Korvalääketieteellisen aineiston luokittelu Bayes -verkoilla

Katja Miettinen

Tampereen Yliopisto

Informaatiotieteiden tiedekunta

Matematiikan, tilastotieteen ja filosofian laitos

1 TIIVISTELMÄ

Tämän työn teoriaosuudessa on esitetty luokittelevien Bayes-verkkojen teoriaa. Koska kysymyksessä on luokitteluongelma, niin teoriaosuudessa esitetään luokittelutehtävään liittyviä näkohtia. Tavoitteena on rakentaa luokittelija siten, että väärinluokituksen mahdollisuus minimoidaan. Muita tärkeitä hyvän luokittelijan piirteitä ovat väärin luokittelemisen ja a priori esiintymistodennäköisyyksien huomioon ottaminen. Koska työn konteksti on Bayes -verkot, niin työssä on johdateltu bayesilaiseen päättelyyn ja esitelty ehdollisen todennäköisyyden käsite sekä Bayesin lause. Lisäksi teoriaosuudessa on vertailtu bayesilaista todennäköisyyden määritelmää frekventistiseen määritelmään. Bayes -verkkoa esittävän graafin oletetaan olevan suunnattu ja syklitön (*DAG*) – siksi teoriosuudessa on esitetty myös graafiteoriaa.

Bayes -verkon oppimisessa on kaksi vaihetta. Ensinnäkin on opittava Bayes -verkon rakenne. Bayes -verkon rakenteen oppimiseen on kaksi lähestymistapaa. Tässä työssä rakenteen oppimiseen on käytetty pistemääräperustaista lähestymistapaa. Siinä haetaan kaikki mahdolliset verkkorakenteet jollain hakualgoritmilla, pisteytetään saadut verkot ja valitaan parhaimman pistemäärän saanut verkko Bayes -verkon rakenteeksi. Tähän liittyen työssä on esiteltynä vuorikiipeilyalgoritmi. Vaihtoehtoinen Bayes -verkon rakenteen oppimiseen käytetty menelmä, nimeltään rajoiteperustainen menetelmä (*constraint based*), on esitelty tässä työssä suppeasti. Toinen Bayes -verkon oppimiseen liittyvä näkökohta on Bayes -verkon parametrien estimointi. Työssä on paneuduttu tarkemmin parametrien estimointiin yleensä frekventistisessä ja bayesilaisessa mielessä. Huomionarvoista on se, että Bayes -verkko voi olla frekventistinen. Tämän työn empiirisessä osuudessa on luokiteltu 815 huimauspotilasta tautiryhmiin käyttäen luokittelijoita: naiivi, *TAN*, GBN_1, GBN_2 ja GBN_3. Naiivi luokittelija perustuu oletukseen, että muuttujat ovat ehdollisesti riippumattomia, kun luokittelumuuttuja on annettu. Verkkorakenne on tällä luokittelijalla puu, jossa ainoa vanhempi on luokittelumuuttuja. *TAN* (*Tree Augmented Naive-Bayes*) -luokittelija sallii toisen vanhemman luokittelumuuttujan lisäksi. *TAN* -luokittelijan rakenteen oppiminen pohjautuu tunnettuun Chown ja Liun vuonna 1968 esittämään menetelmään puutyyppisten Bayes -verkkojen

oppimiseen. Yleisessä Bayes -verkossa (*General Bayes Network, GBN*) luokittelumuuttuja on kuten mikä tahansa solmu, eikä solmujen vanhempien lukumäärää ole rajoitettu.

Tässä työssä käsitellään kolmea yleistä luokittelijaa. Näiden luokittelijoiden rakenteiden oppimiseen on käytetty pistemääräperustaista lähestymistapaa. Käytetyt pistemäärät eroavat näillä luokittelijoilla. Luokittelijalla GBN_1 käytetty pistemäärä on Bayes -pistemäärä. Luokittelijalla GBN_2 käytetty pistemäärä on *MDL* (*Minimum Description Length*) -pistemäärä ja luokittelijalla GBN_3 käytetyn pistemäärän ollessa *AIC* (*Akaike Information Criterion*) -pistemäärä. Koska tässä työssä käsitellään verkkoja, jossa puuttuvia arvoja ei sallita, niin puuttuvat arvot korvattiin muuttujien keskiluvuilla. Puuttuviin arvoihin liittyvää problematiikka on myös näin ollen käsitelty empiriaosuudessa. Kaiken kaikkiaan empiriaosuudessa käsitellään 15 eri luokittelijaa, edellä esitettyjä luokittelijoita selittävien muuttujien lukumäärillä 40, yhdeksän ja viisi. Tämän aineiston potilaiden luokittelu tautiryhmiin Akustikus Neurinoma, Bening positional vertigo, Menièren tauti, Sudden Deaffness, Traumatic Vertigo ja Vesbular Neuritis tehtiin käyttäen open-source ohjelmaa Weka 3 (*Waikato Environment for Knowledge Analysis*).

2 Johdanto

2.1 Esittely

Lopputyön sisältö on jaettu siten, että ensiksi luvussa 2 kerrotaan yleisiä asioita frekventistisestä ja bayesilaisesta päättelystä, sekä esitellään ehdollisen todennäköisyyden käsite kaavoineen ja laajennetaan tätä Bayesin teoreemaan esimerkkeineen. Luvussa 3 esitellään luokitteluun liittyvää teoriaa ja tähdennetään hyvän luokittelijan ominaisuuksia. Luvussa 4 kerrotaan graafiteoriasta, Bayes -verkkojen taustalla olevasta teoriasta sekä esitellään menetelmiä Bayes-verkon rakenteen ja parametrien estimoimiseksi. Verkon parametrien oppimisen alustukseksi on tässä luvussa esitettynä ensin yleisiä parametrien estimointiin liittyviä näkökohtia. Luvussa 5 kerrotaan aineistosta, testiasetelmasta, käytetystä ohjelmasta, käytetyistä luokittelijoista, puuttuvien havaintoihin liittyvästä problematiikasta, sekä lopuksi tuloksista. Luvussa 6 on loppuyhteenveto.

2.2 Frekventistinen ja bayesiläinen todennäköisyys

Useimmat meistä tutustuessaan todennäköisyyslaskentaan tutustuvat aluksi todennäköisyyden frekventistiseen määritelmään. Frekventistinen todennäköisyys määritetään suhteellisen frekvenssin kautta, toisin sanoen jos koe toistetaan n kertaa ja tapahtuma A esiintyy n_A kertaa, niin todennäköisyys $P(A)$ on suhteen n_A/n raja-arvo, kun $n \to \infty$. Frekventististä lähestymistapaa luonnehtivat kaksi näkökulmaa. Ensinnäkin todennäköisyys käsitetään maailman fysikaaliseksi ominaisuudeksi – todennäköisyyden määrittäminen ei riipu todennäköisyyden määrittävästä henkilöstä (objektiivinen todennäköisyys). Toiseksi todennäköisyys voidaan määrittää vain sellaisten kokeiden tuloksille, jotka ovat ainakin periaatteessa toistettavissa. Lisäksi oletetaan, että toistettavissa olevat kokeet ovat riippumattomia toisistaan. Selvästi frekventistinen todennäköisyyden tulkinta toimii totuttuihin esimerkkeihin kolikonheitosta ja nopanheitosta. Frekventistisen koulukunnan edustajalla ei ole juuri sanottavaa ainutlaatuisista, ei-toistettavista tapahtumista. Esimerkiksi kysymykseen mikä on todennäköisyys, että Tappara voittaa jääkiekon SM-kultaa vuonna 2008 ei frekventistiseltä pohjalta voida vakuuttavasti vastata. Vaikka Tapparan menestymisestä

liigassa on tietoa edellisiltä vuosilta, niin tätä tietoa ei voida yksinään käyttää todennäköisyyden, että Tappara voitaa SM-kultaa vuonna 2008 määrittämiseen. Kausi 2007-2008 on ainutkertainen liigakausi: joukkueiden pelaajat vaihtuvat, osa pelaajista kärsii loukkaantumisista jne. Voidaan sanoa, että yksikään aikaisempi kausi ei ole ollut samanlainen kuin kausi 2007-2008. [3]

Bayesiläinen todennäköisyys on erilainen todennäköisyyden tulkinta, jota voidaan soveltaa edellä mainittuun ainutlaatuiseen tilanteeseen ja oikeastaan kaikkiin tilanteisiin, joihin liittyy epävarmuustekijä. Tapahtuman *x* bayesiläinen todennäköisyys on henkilön *uskomuksen aste tapahtumalle saadun uuden datan valossa*. Bayesiläinen lähestymistapa on itse asiassa normatiivinen lähestymistapa, joka määrittää miten datan pitäisi vaikuttaa henkilön ajatteluun ja mahdollisen toiminnan valintaan. Tavoitteena on vähentää tilanteeseen liittyvä epävarmuutta siten, että kaikki mahdollinen informaatio käytetään hyväksi. Bayesiläisen päättelyn tunnuspiirre on, että uskomukset tuntemattomista suureista esitetään suoraan todennäköisyyden termein. Nämä todennäköisyydet käsitetään tavallisesti subjektiivisiksi todennäköisyyksiksi. Formaali menetelmä, joka yhdistää uuden tiedon aikaisemmin saatavilla olevaan tietoon on nimeltään *Bayesin lause* (*Bayes' theorem*) [3, 1, 17]. Bayesilaistä lähestymistapaa kritisoidaan siitä, että se esittää uskomuksen asteet suoraan todennäköisyyden termein. Kysytään, että miksi uskomuksen asteen pitäisi täyttää todennäköisyyden säännöt ja millä asteikolla nämä todennäköisyydet pitäisi mitata.

2.3 Ehdollinen todennäköisyys ja riippumattomuus

Olkoot muuttujat *X* ja *Y* diskreettejä muuttujia. Ehdollinen todennäköisyys ehdolla *Y* määritellään seuraavasti:

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \quad , \; P(Y) > 0 \qquad (1)$$

missä $P(X,Y)$ on muuttujien *X* ja *Y* yhteistodennäköisyysfunktio ja *P(Y)* on muuttujan *Y* reunajakauma. Muuttujat *X* ja *Y* ovat toisistaan riippumattomia, jos *P(X|Y)=P(X)*. Tällöin muuttujan *Y* arvojen havaitseminen ei vaikuta muuttujan *X* arvojen esiintymistodennäköisyyksiin. Yleisesti on voimassa, että jos muuttujat ($X_1,\ldots,X_n$) ovat

riippumattomia täydellisesti, niin niiden yhteistodennäköisyysfunktio on muotoa $P(X_1,\ldots,X_n)=P(X_1)\cdots P(X_n)$.

Yleensä tilanne on se, että muuttujat eivät ole keskenään riippumattomia. Voi kuitenkin olla, että muuttujat ovat ehdollisesti riippumattomia keskenään. Muuttujat X ja Y ovat ehdollisesti riippumattomia ehdolla Z, merkitään $Ind(X;Y|Z)$, jos $P(X|Y, Z)=P(X|Z)$. Tämä tarkoittaa, että kun muuttuja Z on havaittu, niin muuttujan Y havaitseminen ei tuo lisäinformaatiota X ennustamiseen.

2.4 Bayesin teoreema

Aina ei ole helppoa laskea ehdollisia todennäköisyyksiä suoraan käyttäen ehdollisen todennäköisyyden kaavaa. Käytännöllinen kaava, joka yhdistää useita erilaisia ehdollisia todennäköisyyksiä, on Bayesin teoreema. Yksinkertaisin muoto Bayesin teoreemasta on muotoa:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \neg A)P(\neg A)}, \qquad (2)$$

missä A ja B ovat tapahtumia ja ¬A on A:n komplementti.

Tämä Bayesin lauseen yksinkertaisin muoto seuraa suoraan ehdollisen todennäköisyyden (1) määritelmästä:

$$P(A \mid B) = \frac{P(A \cup B)}{P(B)}.$$

Samaan tapaan

$$P(B \mid A) = \frac{P(A \cup B)}{P(A)}$$

ja

$$P(B \mid \neg A) = \frac{P(\neg A \cup B)}{P(\neg A)}.$$

Saadaan

$$P(A \cup B) = P(B \mid A)P(A) \text{ ja } P(\neg A \cup B) = P(B \mid \neg A)P(P(\neg A).$$

Selvästi on niin, että tapahtumat $A \cap B$ ja $\neg A \cap B$ ovat toistensa poissulkevat ja unioni $(A \cap B) \cup (\neg A \cap B) = B$. Tästä seuraa todennäköisyyden additiivisuuden ja ehdollisen todennäköisyyden määritelmän nojalla, että

$$P(B) = P(A \cup B) + P(\neg A \cup B) = P(A)P(B \mid A) + P(\neg A)P(B \mid \neg A) .$$

Nyt ehdollisen todennäkösyyden kaava saadaan muotoon:

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\neg A)P(B \mid \neg A)} . \qquad (3)$$

Bayesin teoreema voidaan kirjoittaa paljon yleisemmässsä muodossa. Jos *i* tapahtumaa $A_1, A_2, ..., A_i$ ovat toistensa poissulkevat ja *B* on toinen tapahtuma, niin tällöin

$$P(A_k \mid B) = \frac{P(B \mid A_k)P(A_k)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \ldots + P(A_i)P(B \mid A_i)}$$
$$= \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} . \qquad (4)$$

Kaavassa (4) esitettiin Bayesin teoreema hyväksi käyttäen tapahtumia A_i . Korvataan nyt tapahtumat hypoteesien joukolla $H_1,...,H_i$. Tästä hypoteesien joukosta olisi löydettävä sopivin hypoteesi käsillä olevaan käytännön tilanteeseen. Olkoon *D* on havaittu otos tarkasteltavan tilanteen datasta. Korvataan tapahtuma *B* kaavassa (4) *D:llä.* Todennäköisyys $P(H_i)$ on hypoteesin H_i prioritodennäköisyys. Prioritodennäköisyydellä tarkoitetaan hypoteesin H_i todennäköisyyttä ennen havaintoja datasta, toisin sanoen tämä todennäköisyys kuvaa ennakkokäsitystä mahdollisuudesta, että H_i on oikea hypoteesi. Todennäköisyydet $P(D \mid H_i)$ tulkitaan todennäköisyyksiksi, että data *D* havaitaan, kun H_i on oikea hypoteesi. Nämä todennäköisyydet ovat havaitun otoksen uskottavuuksia (*likelihood*) eri hypoteesien vallitessa. Bayesin teoreema voidaan tulkita nyt päivitystyökaluksi, joka yhdistää havaitun datan ja hypoteesin H_i prioritodennäköisyyden. Päivitetyt todennäköisyydet, toisin sanoen todennäköisyydet, että eri hypoteesit ovat voimassa vielä datan havaitsemisen jälkeen ovat

posterioritodennäköisyyksiä $P(H_i \mid D)$ *(i=1,...,n)*. Nämä posterioritodennäköisyydet saadaan soveltamalla edellä esitettyä Bayesin kaavaa:

$$P(H_k \mid D) = \frac{P(H_k)P(D \mid H_k)}{\sum_{i=1}^{n} P(H_i)P(D \mid H_i)}. \tag{5}$$

Monesti ollaan kiinnostuneita löytämään se kaikkein todennäköisin hypoteesi $H_i \in H$, kun data *D* on annettu. Jos tällaisia kaikkein todennäköisimpiä hypoteeseja on useita, niin valitaan yksi hypoteesi *maksimaalisesti todennäköisten* hypoteesien joukosta. Mikä tahansa tällainen maksimaalisesti todennäköinen hypoteesi on nimeltään *maximum a posteriori* (*MAP*) hypoteesi. [2]

MAP -hypoteesi määritetään Bayesin teoreemaan avulla laskemalla jokaiselle hypoteesiehdokkaalle posterioritodennäköisyydet. *MAP* -hypoteesi H_i^{MAP} on muotoa:

$$\begin{aligned} H_i^{MAP} &\equiv \arg\max_{H_i \in H} P(H_i \mid D) \\ &= \arg\max_{H_i \in H} \frac{P(D \mid H_i)P(H_i)}{P(D)}, \\ &= \arg\max_{H_i \in H} P(D \mid H_i)P(H_i) \end{aligned} \tag{6}$$

missä viimeinen relaatio seuraa siitä, että *P(D)* on maksimoinnin suhteen vakio. Jos prioritodennäköisyydet $P(H_i)=P(H_j)$, kun $i \neq j$, niin tarkastelun kohteena on enää $P(D|H_i)$, joka on itse asiassa H_i:n uskottavuusfunktio. Toisin sanoen, kun H_i:n prioritodennäköisyydet ovat yhtäsuuret, niin $P(D|H_i)=L(H_i;D)$. Tällöin *MAP* -hypoteesi on *ML* –hypoteesi (*maximum likelihood hypothesis)*. *ML* -hypoteesi on muotoa

$$H_i^{ML} = \arg\max_{h \in H} P(D \mid H_i). \tag{7}$$

2.4.1 Esimerkki lääketieteellisestä diagnosoinnista

Potilaalle tehdään testi, jonka pohjalta tehdään diagnoosi. Hypoteeseja on kaksi: Potilaalla on harvinainen sairaus tai potilaalla ei ole harvinaista sairautta. Testituloksia on

myös kaksi: Negatiivinen testitulos ja positiivinen testitulos. Prioritietämyksenä on, että koko populaatiossa vain 0,8 prosentilla on tämä harvinainen sairaus. Tiedetään myös, että testi antaa oikean positiivisen tuloksen 98 % tapauksista, joilla sairaus oikeasti on. Vastaavasti tiedetään, että testi antaa oikean negatiivisen tuloksen 97 % tapauksista, joilla sairautta ei ole. Potilas saa laboratoriosta positiivisen testituloksen. Pitäisikö potilas nyt diagnosoida harvinaisen sairauden kantajaksi? Tehtävänä on nyt määrittää *MAP* -hypoteesi, toisin sanoen se hypoteesi, joka maksimoi posterioritodennäköisyyden.

Lasketaan todennäköisyydet esitetylle kahdelle hypoteesille, kun evidenssi, joka nyt on testitulos, on havaittu. Merkitään H_1 *="Henkilöllä on sairaus"* ja H_2 *="Henkilöllä ei ole sairautta"*. Todennäköisyydet hypoteeseille ovat $P(H_1)=0,008$ ja $P(H_2)=0,992$. Merkitään *D="Havaitut positiiviset tapaukset"* .

Todennäköisyys ensimmäiselle hypoteesille, kun data on havaittu on muotoa

$$P(H_1 \mid D) = \frac{P(H_1)P(D \mid H_1)}{P(D)} = \frac{0,008 \times 0,98}{0,0376} = 0,207 ,$$

ja todennäköisyys toiselle hypoteesille on muotoa

$$P(H_2 \mid D) = \frac{P(H_2)P(D \mid H_2)}{P(D)} = \frac{0,992 \times 0,03}{0,0376} = 0,793 .$$

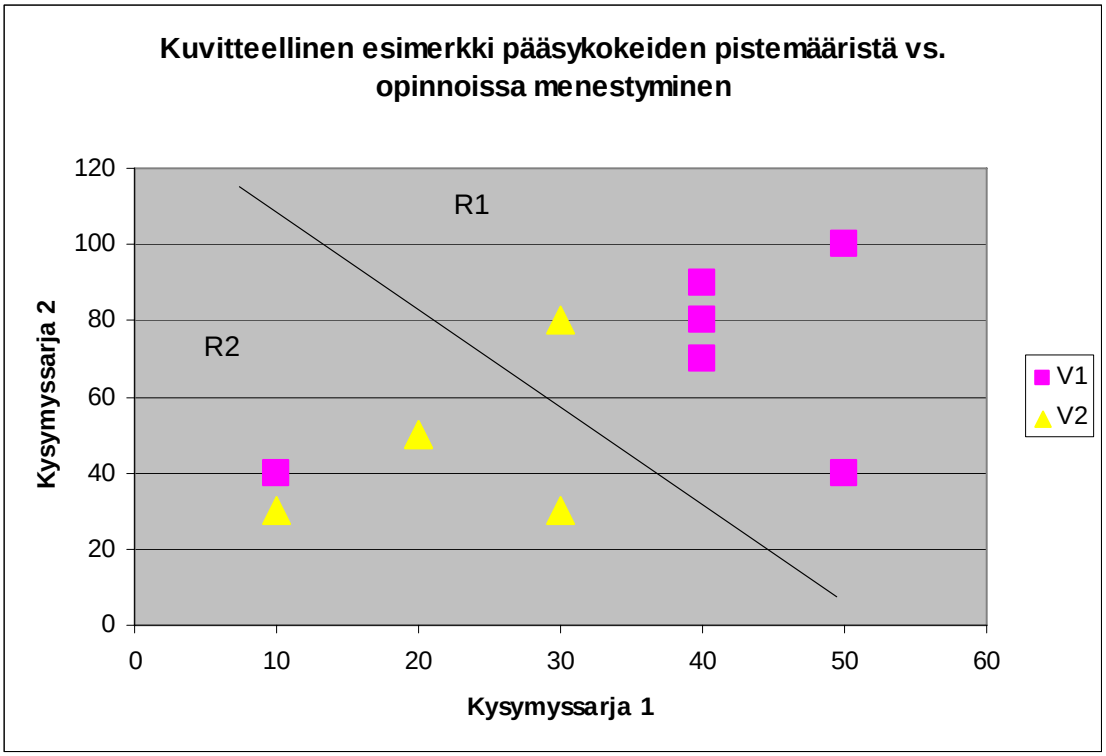
Jälkimmäinen hypoteesi maksimoi posterioritodennäköisyyden. Diagnoosi on nyt, että potilaalla ei ole kyseistä sairautta.

3 Erotteluanalyysi

3.1 Yleistä erotteluanalyysistä

Olkoon m (≥ 2) kappaletta ryhmiä tai luokkia $(v_1, v_2, ..., v_m)$. Erotteluanalyysin (discriminant analysis) tavoitteena on sijoittaa $\boldsymbol{x}$ johonkin näistä m:stä ryhmästä. Havainto $\boldsymbol{x}=(x_1, ..., x_p)$ sisältää yksilön saamat arvot datan muuttujille x_1, $x_2, ..., x_p$. Havainto $\boldsymbol{x}$ on piste p-ulotteisessa avaruudessa. Jos $\boldsymbol{x}$ kuuluu luokkaan v_j $(j=1...m)$, niin sillä on tiheysfunktio $f_j(\boldsymbol{x}) \in R^p$. Erottelusääntö (discriminant rule) d vastaa R^p:n jakoa toistensa poissulkeviin alueisiin $R_1, ..., R_m (\cup R_j = R^p)$. Erottelusääntö määritetään seuraavasti: Sijoita $\boldsymbol{x}$ ryhmään v_j, jos $\boldsymbol{x} \in R_j$ $(j=1...m)$. [6,20]

Edellä esitettiin formaalisti erotteluanalyysin lähtökohdat. Vähemmän formaalisti ilmaistuna erotteluanalyysi vastaa seuraavaan kysymykseen. Kun on annettu yksilö ja tähän kyseiseen yksilöön liittyvät muuttujien arvot, niin mistä populaatiosta tai ryhmästä tämä yksilö on lähtöisin. Kuvitteellinen esimerkki kauppatieteellisen pääsykokeesta. Pääsykokeessa on erilaisia kysymyssarjoja. Pääsykoekysymysten vastausten perusteella halutaan luokitella opiskelijat kahteen luokkaan v_1 ja v_2: Luokassa v_1 ovat opiskelijat, jotka pärjäävät opinnoissaan hyvin. Luokassa v_2 ovat opiskelijat, jotka eivät menesty opinnoissaan. Kuvassa 1 on noita pistemääriä vastaava pisteparvi siten, että jokaisen pisteen kohdalla on merkintä kumpaan luokkaan yksilö kuuluu. Kuvassa on viiva, joka erottaa kaksiulotteisesta avaruudesta kaksi aluetta R_1 ja R_2. Yksilöt, jotka osuvat alueelle R_1, luokitellaan kuuluvaksi luokkaan v_1 ja vastaavasti yksilöt, jotka osuvat alueelle R_2 luokitellaan luokkaan v_2. Huomataan, että jotkut yksilöt, jotka kuuluisivat oikeasti ryhmään v_1, ovat alueella R_2 ja vastaavasti muutama ryhmään v_2 kuuluva sijoittuu alueelle R_1.



Kuva 1 Esimerkki luokittelutehtävästä opiskelijoiden luokittelu opinnoissa menestyviin ja ei-menestyviin

Erotteluanalyysin tavoitteena on luoda sellainen sääntö (alueet R_1 ja R_2), joka minimoi väärinluokituksen mahdollisuutta. Kahden luokan tapauksessa väärinluokitusvaihtoehtoja on kaksi. Toinen on, että yksilö kuuluu luokkaan v_1 ja luokitellaan kuuluvaksi luokkaan v_2, kun toinen väärinluokitusvaihtoehto on, että yksilö kuuluu luokkaan v_2 ja luokitellaan kuuluvaksi luokkaan v_1. [20]

Edellä mainittu väärinluokituksen mahdollisuuden minimointi on erottelusäännön tärkein piirre. Lisäksi on olemassa muita piirteitä, joita optimaalisen luokittelusäännön pitäisi sisältää. Esimerkiksi toisella populaatiolla voi olla suurempi esiintymistodennäköisyys kuin toisella syystä, että toinen populaatio on suhteellisesti paljon suurempi kuin toinen. Optimaalisen luokittelusäännön pitäisi siis ottaa huomioon esiintymistodennäköisyydet a priori [20].

Toinen lisäpiirre, jonka luokittelusääntöön voi kiinnittää, on käsite *tappio*. Voi olla, että oikeasti luokkaan v_1 kuuluvan luokittelu luokkaan v_2 on pahempi virhe kuin, että luokkaan v_2 kuuluva luokiteltaisiin väärin luokkaan v_1. Toisin sanoen ensin mainitusta

väärinluokituksesta aiheutuu suurempi tappio kuin jälkimmäisestä. Esimerkiksi jos tehtävänä on päättää, sairastaako potilas flunssaa vai leukemiaa, olisi kohtalokkaampaa diagnosoida leukemiaa sairastava flunssapotilaaksi, kuin diagnosoida flunssaa sairastava leukemiapotilaaksi. Kuvassa 2 on tappiomatriisi, jossa on väärinluokituksesta aiheutuvat tappiot, kun yksilö on luokiteltu luokkaan v_i sen kuuluessa luokkaan v_j. Ensimmäisestä väärinluokitusvirheestä aiheutuva tappio on $C(v_2|v_1)$ ja vastaavasti toisesta väärinluokitusvirheestä aiheutuva tappio on $C(v_1|v_2)$. Oikeinluokituksen tappio on luonnollisesti *0*. [20]

		Luokitellaan	
		v1	v2
Populaatio	*v1*	*0*	*C(2|1)*
	v2	*C(1|2)*	*0*

Kuva 2 Tappiomatriisi

Olkoon nyt $f_1(\mathbf{x})$ ja $f_2(\mathbf{x})$ todennäköisyysfunktiot, joihin liitetään $p \times 1$-ulotteinen satunnaismuuttujavektori populaatioissa v_1 ja v_2. Objekti, johon on kiinnitettynä mittavektori **x** on luokiteltava joko luokkaan v_1 tai v_2. Olkoon Ω otosavaruus, toisin sanoen kaikkien mahdollisten havaintojen **x** kokoelma. Olkoon R_1 se arvojen **x** joukko, jolle objektit luokitellaan kuuluvaksi luokkaan v_1 ja olkoon $R_2=\Omega-R_1$ jäljelle jäävä arvojen joukko **x**, jolle objektit luokitellaan luokkaan v_2. Koska jokainen objekti on luokiteltava toiseen luokista, niin joukot R_1 ja R_2 ovat keskenään toisensa poissulkevia ja keskenään tyhjentävät. Ehdollinen todennäköisyys $P(v_2|v_1)$, että luokitellaan objekti luokkaan v_2, kun se on oikeasti kotoisin luokasta v_1 on

$$P(v_2 \mid v_1) = P(\mathrm{X} \in R_2 \mid v_1) = \int_{R_2} f_1(\mathrm{x}) d\mathrm{x} \,. \tag{8}$$

Vastaavasti ehdollinen todennäköisyys $P(v_1|v_2)$, että luokitellaan objekti luokkaan v_1, kun se oikeasti on kotoisin luokasta v_2 on

$$P(v_1 \mid v_2) = P(\mathrm{X} \in R_1 \mid v_2) = \int_{R_1} f_2(\mathrm{x}) d\mathrm{x} \,. \tag{9}$$

Olkoon $P(v_1)$ luokkaan v_1 kuulumisen prioritodennäköisyys ja $P(v_2)$ luokkaan v_2 kuulumisen prioritodennäköisyyys, kun $P(v_1)+ P(v_2)=1$. Nyt kaikki oikeinluokituksiin ja

väärinluokituksiin liittyvät todennäköisyydet voidaan johtaa prioritodennäköisyyden ja ehdollisen luokittelutodennäköisyyden tulona:

$$P(\textit{luokiteltu} \quad \textit{oikein} \quad \textit{luokkaan} \quad v_1) = P(\mathrm{X} \in R_1 \mid v_1)P(v_1)$$

$$P(\textit{luokiteltu} \quad \textit{väärin} \quad \textit{luokkaan} \quad v_1) = P(\mathrm{X} \in R_1 \mid v_2)P(v_2),$$

$$P(\textit{luokiteltu} \quad \textit{oikein} \quad \textit{luokkaan} \quad v_2) = P(\mathrm{X} \in R_2 \mid v_2)P(v_2),$$

$$P(\textit{luokiteltu} \quad \textit{väärin} \quad \textit{luokkaan} \quad v_2) = P(\mathrm{X} \in R_2 \mid v_1)P(v_1).$$

Aikaisemmin mainittiin, että väärinluokituksen mahdollisuuden minimoinnin lisäksi olisi joissain tapauksissa järkevää tarkastella myös väärinluokituksesta aiheutuvaa tappiota. Kuvan 2 tappiomatriisissa on kahden luokan tapauksessa aiheutuvat tappiot. Näiden tappioiden avulla määritetään *odotettu väärinluokituksesta aiheutuva tappio* (*ECM*, *expected cost of misclassification*):

$$\begin{aligned} ECM \quad &= C(v_2|v_1)P(v_2|v_1)P(v_1)+0\cdot P(v_2|v_2)P(v_2) \\ &+ C(v_1|v_2)P(v_1|v_2)P(v_2)+0\cdot P(v_1|v_1)P(v_2) \\ &= C(v_2|v_1)P(v_2|v_1)P(v_1)+ C(v_1|v_2)P(v_1|v_2)P(v_2). \end{aligned} \tag{10}$$

Järkevän luokittelusäännön pitäisi olla sellainen, että *ECM* on mahdollisimman pieni. Tehtävänä on nyt keskimääräisen tappion minimointi; halutaan jakaa avaruus kahteen alueeseen R_1 ja R_2 siten, että odotettu tappio on mahdollisimman pieni. Alueet R_1 ja R_2, jotka minimoivat odotetun tappion *ECM* määritetään arvoilla **x**, joille on voimassa seuraavat epäyhtälöt.

$$R_1: \quad \frac{f_1(\mathrm{x})}{f_2(\mathrm{x})} \geq \left(\frac{C(v_1 \mid v_2)}{C(v_2 \mid 1)}\right)\left(\frac{P(v_2)}{P(v_1)}\right) \tag{11}$$

$$R_2: \quad \frac{f_1(\mathrm{x})}{f_2(\mathrm{x})} < \left(\frac{C(v_1 \mid v_2)}{C(v_2 \mid v_1)}\right)\left(\frac{P(v_2)}{P(v_1)}\right) \tag{12}$$

Kun priorijakaumat ovat samat, väärinluokituksesta aiheutuvat tappiot ovat samat, tai sekä priorijakaumat että tappiot ovat samat, niin epäyhtälöt yksinkertaistuvat. Käytännössä tapana on soveltaa vaihtoehtoa, jossa sekä priorijakaumat että väärinluokituksesta aiheutuvat tappiot oletetaan samoiksi, toisin sanoen

$$\frac{P(v_2)}{P(v_1)} = \frac{C(v_1 \mid v_2)}{C(v_2 \mid v_1)} = 1 \text{ [20].} \tag{13}$$

Nyt odotetun tappion minimoivat alueet ovat muotoa:

$$R_1: \quad \frac{f_1(\mathrm{x})}{f_2(\mathrm{x})} \geq 1, \tag{14}$$

$$R_2: \quad \frac{f_1(\mathrm{x})}{f_2(\mathrm{x})} < 1. \tag{15}$$

Edellä esitettiin luokittelukriteeriksi odotettua väärinluokituksen tappiota. Kun väärinluokituksesta aiheutuvaa tappiota ei oteta huomioon, voidaan alueet R_1 ja R_2 valita esimerkiksi siten, että minimoidaan *väärinluokittelun mahdollisuuden kokonaistodennäköisyyttä* (*TPM, Total probability of misclassification*).

$$\begin{aligned} &TPM \\ &= P(\textit{väärinluokitellaan luokkaan } v_1 \textit{ tai väärinluokitellaan luokkaan } v_2) \\ &= P(\textit{väärinluokitellaan luokkaan } v_1) + P(\textit{väärinluokitellaan luokkaan } v_2) \\ &= P(v_1)\int_{R_2} f_1(\mathrm{x})d\mathrm{x} + P(v_2)\int_{R_1} f_2(\mathrm{x})d\mathrm{x} \end{aligned} \tag{16}$$

Itse asiassa tappion huomiotta jättäminen on sama asia kuin tappioiden yhtäsuuruus. Nyt kokonaistodennäköisyyttä minimoivat alueet ovat

$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{P(v_2)}{P(v_1)}, \tag{17}$$

$$R_2: \frac{f_1(x)}{f_2(x)} < \frac{P(v_2)}{P(v_1)}. \tag{18}$$

Esimerkki 1 Olkoot datasta estimoidut ryhmiin v_1 ja v_2 liittyvät tiheysfunktiot $f_1(\mathbf{x})$ ja $f_2(\mathbf{x})$. Oletetaan, että tappiot ovat $C(v_1|v_2)$=100 ja $C(v_2|v_1)$=50 ja että luokkiin kuulumisen prioritodennäköisyydet ovat $P(v_2)$=0,2 ja $P(v_1)$=0,8. Määritetään nyt luokittelualueet

$$R_1: \frac{f_1(x)}{f_2(x)} \geq \left(\frac{100}{50}\right)\left(\frac{0{,}2}{0{,}8}\right) = 0{,}5$$

$$R_2: \frac{f_1(x)}{f_2(x)} < \left(\frac{100}{50}\right)\left(\frac{0,2}{0,8}\right) = 0,5\,.$$

Oletetaan, että tiheysfunktioiden arvot uudella arvolla $\mathbf{x}_0$ ovat $f_1(\mathbf{x}_0)=0,3$ ja $f_2(\mathbf{x}_0)=0,5$. Määritetään osamäärä

$$\frac{f_1(\mathrm{x}_0)}{f_2(\mathrm{x}_0)} = \frac{0.3}{0,5} = 0,6\,,$$

ja huomataan, että saatu osamäärä on suurempi kuin 0.5; toisin sanoen

$$\frac{f_1(\mathrm{x}_0)}{f_2(\mathrm{x}_0)} = \frac{0,3}{0,5} > \left(\frac{C(v_1 \mid v_2)}{C(v_2 \mid 1)}\right)\left(\frac{P(v_2)}{P(v_1)}\right) = 0,5\,.$$

Havainto $\mathbf{x}_0 \in R_1$ ja näin ollen se luokitellaan kuuluvaksi luokkaan v_1.

Edellä esitetyistä luokittelukriteereistä hieman poikkeava lähestymistapa on tarkastella posteriorijakaumaa. Jos luokilla $(v_1,\ldots,v_m)$ on prioritodennäköisyydet $(P(v_1),\ldots, P(v_m))$, niin *Bayesin erottelusääntö* sijoittaa uuden havainnon $\mathbf{x}_0$ siihen ryhmään missä $P(v_j|\mathbf{x}_0)=P(v_j)L(v_j|\mathbf{x}_0)$ maksimoituu. Selvää on, että jos luokkien prioritodennäköisyydet ovat samat, niin Bayesin erottelusääntö palautuu *suurimman uskottavuuden säännöksi*. Toisin sanoen maksimoitavaksi jää vain uskottavuusfunktio [6,20]. Tilastotieteellisissä lähteissä, joissa tarkastellaan luokitteluongelmaa yleisellä tasolla, posteriorijakauman tarkastelun yhteydessä mainitaan posteriorijakauman maksimointi. Kun mietitään uuden havainnon ennustamiseen liittyviä näkökohtia, niin posteriorijaukauman maksimointi on yksi ratkaisu tähän ennustamisongelmaan. Toinen on posteriorijakauman odotusarvo, joka on niin binäärisen kuin moniarvoisten muuttujien tapauksissa sama kuin Bayes -ennuste. Mediaani on yleisesti käytetty piste-estimaatti, mutta luokitteluongelmissa sitä harvemmin käytetään.

4 Bayes -verkot

4.1 Johdatus Bayes –verkkoihin

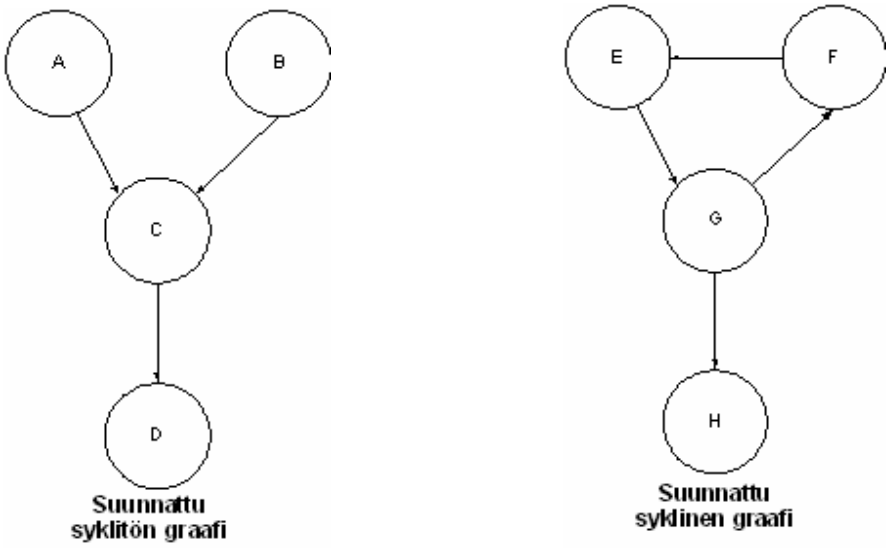
4.1.1 Graafiteoriaa

Olkoon *V* äärellinen *solmujen* (*nodes*, *vertices*) joukko ja olkoon *E kaarien* (*arcs*, *edges*) joukko. Tällöin pari *G*=(*V*,*E)* on *graafi* (*graph*).

Graafin kaaret ovat suunnattuja (*directed*) tai suuntaamattomia (*undirected*). Kaari (*u*,*v*) on suunnattu solmusta *u* solmuun *v* jos parilla (*u*,*v*) on määrätty järjestys, toisin sanoen jos (*u*,*v*)≠(*v*,*u*). Jos (*u*,*v*)=(*v*,*u*), niin kaari (*u*,*v*) on suuntaamaton. Graafia sanotaan suuntaamattomaksi graafiksi (*undirected graph*), jos kaikki sen kaaret ovat suuntaamattomia. Suunnattu graafi (*directed graph*) on graafi, jossa kaikki kaaret ovat suunnattuja. Graafi on *syklitön* (*acyclic*), jos graafissa ei ole yhtään sellaista *polkua* (*path*), jossa aloitussolmu on sama kuin lopetussolmu. Graafista, joka on suunnattu ja syklitön, käytetään yleisesti lyhennettä *DAG (directed acyclic graph)*. [7]

Kuvassa 3 on esimerkit suunnatuista syklittömistä ja syklisistä graafeista. Merkitään edellä olevaa graafia *S*:llä ja jälkimmäistä graafia *P*:llä. Graafin *S* määrittelee solmujen joukko *V={A, B, C, D}* ja kaarien joukko *E={(A, C),(B, C),(C, D)}*. Vastaavasti graafin *P* määrittelee solmujen joukko *V={E, F, G, H}* ja kaarien joukko *E={(E, G),(G, H),(F, E),(G, F)}*. Graafi *P* on syklinen, koska esimerkiksi polku *(E-G-F-E)* on sykli.

Määritellään seuraavaksi solmut *juuri*, *vanhempi*, *lapsi*, *jälkeläinen* ja *lehti*. Juurisolmu on sellainen solmu, johon ei tule yhtään kaarta, toisin sanoen ,jolla ei ole yhtään vanhempaa. Solmulla on lapsi, jos siitä lähtee kaari toiseen solmuun. Lehtisolmulla ei ole yhtään lasta. Kuvan 3 suunnatussa syklittömässä graafissa solmun A lapsi on C ja solmun A jälkeläiset ovat (C,D). Solmun D vanhempi on solmu C.



Kuva 3 Sykliset ja syklittömät graafi

4.1.2 Bayes -verkon määritelmä

Bayes -verkko on graafinen esitys muuttujajoukon $U=(X_i,...,X_n)$ yhteistodennäköisyysjakaumalle. Bayes -verkon katsotaan koostuvan kahdesta osasta:

- Bayes-verkon verkkorakenne on suunnattu syklitön graafi *DAG*. Muuttujajoukko *U* vastaa solmujoukkoa ja kaaret kuvaavat muuttujien välisiä riippuvuuksia.
- Lisäksi jokaiseen muuttujaan/solmuun liitetään paikalliset todennäköisyysjakaumat ehdollistettuna muuttujien vanhempiin.

Formaalimmin ilmaistuna Bayes -verkon määrittää kaksikko $B = \langle G, \Theta \rangle$, missä *G* on suunnattu syklitön graafi ja joukko $\Theta = \{\theta_{x_i|Pa(X_i)}\}$ parametreja, jotka esittävät jokaisen solmun ehdolliset todennäköisyydet vanhempien suhteen, toisin sanoen $\theta_{x_i|Pa(X_i)} = P(x_i \mid Pa(X_i))$. Kuhunkin solmuun liittyvää jakaumaa $P(X_i|Pa(X_i))$ kutsutaan *paikalliseksi todennäköisyysjakaumaksi (local probability distrubution)*. [13, 14]

4.1.3 Bayes -verkon rakenteeseen liittyvät oletukset ja määritelmä

Bayes -verkon ydinidea on, että joukko ehdollisia riippumattomuuksia kuvataan sen rakenteessa. Bayes -verkon avulla saadaan määritettyä yhteistodennäköisyysjakauma

$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$, missä joukko $P(X_i)$ viittaa muuttujan X_i vanhempiin.

Bayes -verkon rakenne *G* on yhteistodennäköisyysjakauman *P I-map (Independency mapping, ks. Määritelmä 3, luku 4.1.6)*. Tämä tarkoittaa, että yhteistodennäköisyysjakauma *P* toteuttaa kaikki verkon *G* ehdolliset riippumattomuudet (*Markov -oletus*). Tämä tarkoittaa, että Bayes -verkon graafinen osuus ei voi ikinä sisältää ehdollisia riippumattomuuksia, jotka eivät ole voimassa yhteistodennäköisyysjaukaumassa. Verkko *G* on Bayes -verkko, jos ja vain jos se on yhteistodennäköisyysjakauman *P minimaalinen I-map* (*minimal I-map*). Verkko *G* on yhteistodennäköisyysjakauman *P* minimaalinen *I-map*, jos yhdenkin kaaren siirtäminen rikkoo *I-map*:in ominaisuuksia. [8]

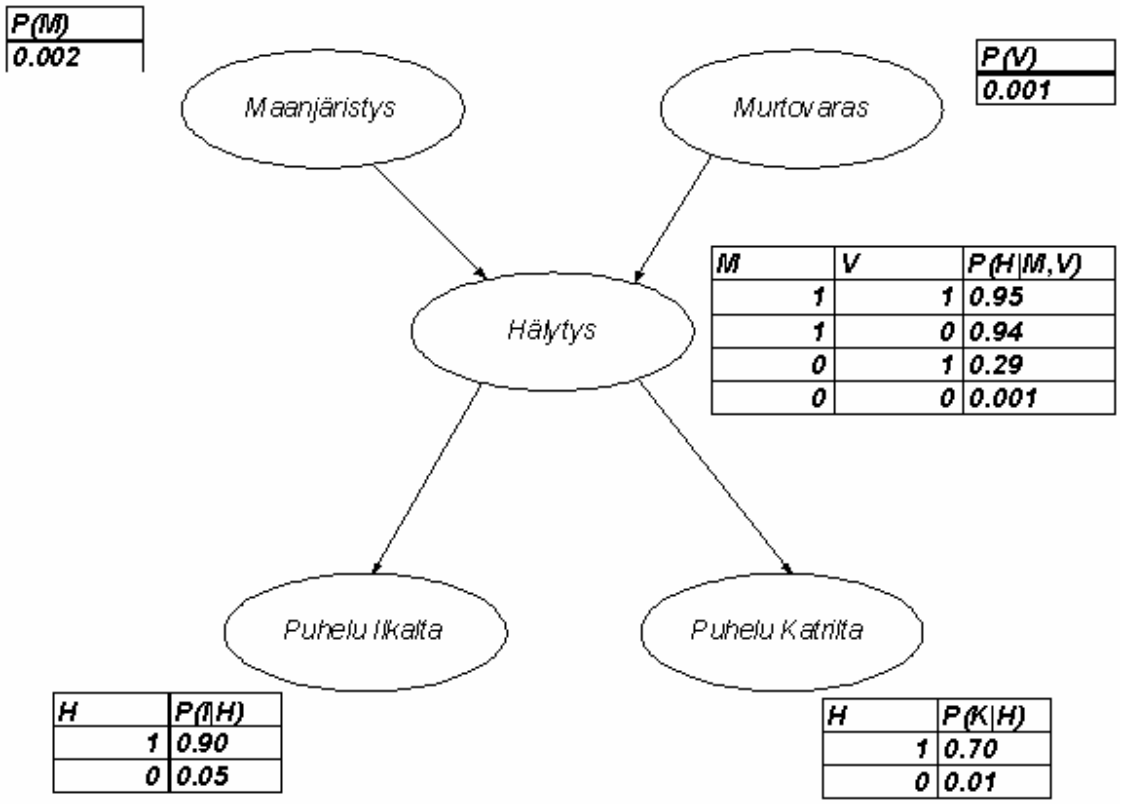
4.1.4 Bayes -verkko ja kausaalisuus

”Samat syyt aiheuttavat samoissa olosuhteissa samoja vaikutuksia”, on eräs mahdollinen kausaalilain formulointi, esittää akateemikko Oiva Ketonen kirjassaan *Se pyörii sittenkin.* Menemättä liian syvälle kausaalisuuden olemukseen, voidaan vain sanoa, että syy-seuraus-suhteen todistaminen ei ole helppo tehtävä. Milloin syyt tai olosuhteet tai vaikutukset ovat riittävän samat, ja miten riittävä samanlaisuus todennetaan, pohtii Ketonen. Oma esimerkkini kausaalisuudesta on seuraava: Rasvaisen ruoan syöminen lihottaa. Vai oliko tuo sittenkään esimerkki kausaalisuudesta?

Luvussa 4.1.2 esittelen, kuinka Bayes -verkossa kaaret kuvaavat solmujen tai muuttujien välisiä riippuvuuksia. Riippuvuuksien tulkinnassa on syytä olla varovainen. Bayes -verkko voi olla kausaaliverkko. Tällöin kaaret solmujen välillä ovat tulkittavissa siten, että jos solmusta *A* lähtee kaari solmuun *B*, niin solmulla *A* on *suora vaikutus* solmuun *B*. Bayes -verkko on kuitenkin tulkittavissa kausaaliseksi, jos ja vain jos se on rakennettu tietyin menetelmin. Kausaalisen Bayes -verkon saa muodostettua käyttäen hyväksi rajoiteperustaisia menetelmiä (*constraint based methods*) [8]. Rajoiteperustaiset menetelmät on lyhyesti esiteltynä luvussa 4.2.2.

Kuten edellä esitettiin, niin solmujen väliset suhteet voidaan tulkita kausaalisiksi. Tässä työssä esitellään tarkemmin Bayes -verkon, jonka solmujen väliset kaaret tulkitaan tilastollisiksi riippuvuuksiksi (*probabilistic dependency*), ideaa. Työni empiirisessä

osuudessa esittelen Bayes -verkkoja, jotka on rakennettu luokittelemaan potilaat oikeisiin otoneurologisiin tautityyppeihin ja tässä yhteydessä solmujen väliset kaaret tulkitaan tilastollisiksi riippuvuuksiksi.



Kuva 4 Esimerkki Bayes -verkosta

Kuvitellaan kuvan 4 verkko kausaaliseksi verkoksi. Verkon takana on seuraavanlainen tarina: Pekka asuu alueella, jossa maanjäristysten esiintymistodennäköisyys on *P(M)=0.002* ja murtojen todennäköisyys on *P(V)=0.001*. Pekka on lähdössä viikonloppureissulle ja on sopinut kahden eri naapurin (*Ilkka, Katri*) kanssa, että he soittavat, jos Pekan murtohälytin laukeaa päälle. Murtohälytin on niin herkkä, että se reagoi myös maanjäristyksiin. Kuvan 2 graafissa ovat kaaret muuttujista *maanjäristys* ja *murtovaras* muuttujaan *hälytys*, koska niillä katsotaan olevan *suora vaikutus* siihen, laukeaako hälytin. Vastaavasti muuttujalla *hälytys* on suora vaikutus muuttujiin *puhelu Ilkalta* ja *puhelu Katrilta*. Sitä vastoin muuttujat *maanjäristys* ja *murtovaras* eivät vaikuta suoraan puheluihin Ilkalta ja Katrilta, vaan muuttujan *hälytys H* kautta. Siis *P(K|H, M, V)=P(K|H)*. Muuttujat *maanjäristys* ja *murtovaras* ovat riippumattomia merkitään $M \perp V$.

Kyseiset muuttujat eivät kuitenkaan ole riippumattomia, kun hälytin on lauennut toisin sanoen $P(M|V,H)\neq P(M|H)$. Muuttujat *puhelu Ilkalta* ja *puhelu Katrilta* ovat riippumattomia, kun muuttuja *hälytys* on havaittu, siis $P(I|K, H)=P(I|H)$. Yhteistodennäköisyysjaukama on nyt muotoa

$$P(M, V, H, I, K) = P(M)P(V)P(H|M,V)P(I|H)P(K|H).$$

Jokaiseen solmuun/muuttujaan liitetään paikallinen todennäköisyysjakauma, kun vanhemmat on annettu. Kuvassa 4 on solmujen vieressä ehdolliset todennäköisyystaulukot (*CPT*). Nyt esimerkiksi,

$$P(\neg M, \neg V, H, I, K) = P(\neg M)P(\neg V)P(H \mid \neg M, \neg V)P(I|H)P(K|H)$$

$$=0,998\times0,999\times0,001\times0,90\times0,70=0,0062.$$

Kuvan 4 Bayes -verkon muuttujat ovat binäärimuuttujia. Yleisesti, jos määritetään n binääriarvoisen muuttujan yhteistodennäköisyysjaukauma, sen laskemiseen tarvitsee määrittää $2^n - 1$ kappaletta yhteistodennäköisyyksiä. Esimerkin Bayes -verkossa muuttujia on *5* kappaletta, siis muuttujien yhteistodennäköisyysjakauman laskemiseksi tarvitsisi teoriassa laskea $2^5-1 = 31$ kappaletta yhteistodennäköisyyksiä. Esimerkkitapauksessa tarvitsi määrittää *10* todennäköisyyttä. Artikkelissa Bayesian Network without Tears [9, s.53] esimerkin Bayes -verkossa on sijoitettu kymmenen solmua toisiinsa nähden tietyllä tavalla. Yhteistodennäköisyysjakauman määrittämiseksi tarvitsi määrittää 21 todennäköisyyttä. Tämä luku on huomattavasti pienempi kuin $2^{10}-1 = 1023$! [9]

Bayes -verkon avulla pystytään esittämään yhteistodennäköisyysjakauma kompaktisti. Yhteistodennäköisyysjakauman laskemisen lisäksi Bayes -verkosta saadaan määritettyä myös reunajakaumat. Jos halutaan esimerkiksi määrittää yllä olevan esimerkin tapauksessa todennäköisyys $P(K=0)$, niin tämä tapahtuu marginalisoinnilla. Toisin sanoen reunatodennäköisyys on nyt muotoa

$$\begin{aligned} P(K=0) &= \sum_{M\in\{0,1\}}\sum_{V\in\{0,1\}}\sum_{I\in\{0,1\}}\sum_{H\in\{0,1\}} P(K=0,M,V,I,H) \\ &= \sum_{M\in\{0,1\}}\sum_{V\in\{0,1\}}\sum_{I\in\{0,1\}}\sum_{H\in\{0,1\}} P(K=0\mid H)P(M)P(V)P(I\mid H)P(H\mid M,V) \quad . \quad (19) \\ &= \sum_{M\in\{0,1\}} P(M)\times\sum_{V\in\{0,1\}} P(V)\times\sum_{I\in\{0,1\}} P(I)\times\sum_{H\in\{0,1\}} P(H\mid M,V)P(K=0\mid H) \end{aligned}$$

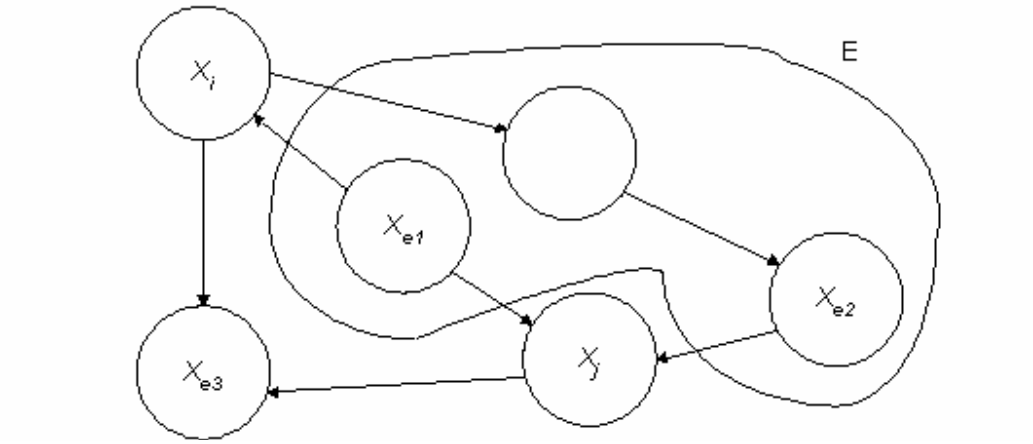
4.1.5 D-separaatio

Edellisen luvun esimerkissä kerroin, mitkä muuttujat ovat ehdollisesti riippumattomia ja mitkä riippuvia. Selvitän seuraavaksi, kuinka muuttujien väliset ehdolliset riippumattomuudet saadaan selvitettyä Bayes -verkosta. Usein käytetty menetelmä ehdollisten riippumattomuuksien löytämiseen Bayes -verkon rakenteesta on Pearlin vuonna 1988 esittämä *d-separaatio* (*d-separation*).

Määritelmä 1 *(d-separaatio)* Olkoon polku vuorottelevien solmujen ja kaarien sekvenssi ja olkoon Z solmujen joukko suunnatussa syklittömässä graafissa G. Tällöin joukko Z d-separoi polun p, jos ainakin toinen ehdoista on voimassa:

1. polku p sisältää ketjun $i \rightarrow j \rightarrow k$ tai divergoituvan haaran $i \leftarrow j \rightarrow k$, että $j \in Z$.
2. polku p sisältää konvergoituvan haaran $i \rightarrow j \leftarrow k$ siten, että $j \notin Z \wedge NonDesc(j) \notin Z$, missä merkintä *NonDesc(j)* viittaa solmun j ei-jälkeläisiin.

Jos X, Y ja Z ovat toistensa poissulkevat verkon G solmujen osajoukot, niin sanotaan, että Z *d-separoi* joukot X ja Y, merk. $(X \perp Y|Z)_G$, jos ja vain jos Z *d-separoi* jokaisen polun, joka lähtee joukon X solmusta ja päättyy joukon Y solmuun. Toisin sanoen joukot X ja Y ovat toisistaan riippumattomia ehdolla Z, jos joukko Z d-separoi joukot X ja Y. [11]



Kuva 5 Esimerkkinä Bayes -verkko, jolla demonstroidaan d-separaatiota.

Demonstroidaan kuvan 5 avulla d-separaatiota. Joukko E sisältää niin sanotut evidenssisolmut. Evidenssisolmu on solmu, johon liittyvän muuttujan arvo on havaittu. Muuttuja X_i on riippumaton muuttujasta X_j kun solmujen X_{e1} ja X_{e2} arvot on annettu. Näin

on, koska kaikki kolme polkua muuttujien X_i ja X_j välillä on suljettu (*blokattu*, engl. *blocked*). Polut on suljettu seuraavasti:

- X_{e1} on evidenssisolmu ja molemmat kaaret suuntautuvat siitä poispäin.
- X_{e2} on evidenssisolmu ja toinen kaari on siihen itseensä päin ja toinen kaari on siitä poispäin.
- X_{e3} ,eikä yksikään sen jälkeläisistä, ole evidenssisolmu ja molemmat kaaret ovat siihen itseensä päin.

Kuvan 5 graafissa on yksi solmu jätetty ilman nimeä. Nimettömän evidenssisolmun merkitys on se, että se on osa polkua solmusta X_i solmuun X_j.

4.1.6 I-map

Määritelmä 2 *(Markov -oletus)* Jokainen muuttuja X_i on riippumaton ei-jälkeläisistään, kun sen vanhemmat ($X \perp_G Y \mid Z$) $Pa(X_i)$ on annettu, toisin sanoen $P(X_i|Pa(X_i), NonDesc(X_i))=P(X_i| Pa(X_i))$.

Määritelmä 3 (*I-map*) *DAG G* on todennäköisyysjakauman *P I-map*, jos d-separaation avulla graafissa *G* esitetyt ehdolliset riippumattomuudet pätevät myös todennäköisyysjakaumassa *P*, ts. jos $(X \perp_G Y \mid Z) \Rightarrow (X \perp Y_P \mid Z)$, missä merkinnällä tarkoitetaan *Y*:n ja *X*:n d-separaatiota ehdolla *Z* ja merkinnällä $(X \perp Y_P \mid Z)$ viitataan ehdolliseen todennäköisyyteen todennäköisyysjakaumassa *P*. [8]

Teoreema 1 Jos *G* on jakauman *P I-map*, niin tällöin graafin *G* yhteistodennäköisyysjakuma on muotoa

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) . \qquad (20)$$

Todistus (Teoreema1) Yleisesti muuttujajoukon yhteistodennäköisyysjaukauma määritetään

$$P(X_1,...,X_n) = P(X_n \mid X_1,...,X_{n-1})P(X_{n-1} \mid X_1,...,X_{n-2}) \cdots P(X_2 \mid X_1)P(X_1) .$$

Tämän saa helposti todistettua käyttäen ketjusääntöä:

$$
\begin{aligned}
&P(X_n \mid X_1,...,X_{n-1}) \cdot P(X_{n-1} \mid X_1,...,X_{n-2}) \cdot ... \cdot P(X_2 \mid X_1) P(X_1) \\
&= \frac{P(X_1,...,X_n)}{P(X_1,...,X_{n-1})} \cdot \frac{P(X_1,...,X_{n-1})}{P(X_1,...,X_{n-2})} \cdot ... \cdot \frac{P(X_1,X_2)}{P(X_1)} \cdot P(X_1) \\
&= P(X_1,...,X_n)
\end{aligned} \tag{21}
$$

Tehdään oletus, että muuttujien joukko $(X_1,...,X_n) \subseteq \mathbf{X}$ on tietyssä järjestyksessä, ja että tämä järjestys pätee myös graafissa *G*; jos muuttuja X_i on muuttujan X_j vanhempi graafissa *G* niin tällöin $i<j$, siis $Pa(X_i) \subseteq (X_1,...,X_{i-1})$ ja $(X_1,...,X_{i-1}) - Pa(X_i) \subseteq NonDesc(X_i)$. Ketjusäännön mukaan $P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid X_1,...,X_{i-1})$. Koska *G* on jakauman *P I-map*, niin $P(X_i \mid NonDesc(X_i), Pa(X_i)) = P(X_i \mid Pa(X_i))$, ja täten $P(X_i \mid (X_1,...,X_{i-1}) - Pa(X_i), Pa(X_i)) = P(X_i \mid Pa(X_i))$, mistä voidaan päätellä että $P(X_i \mid X_i,...,X_{i-1}) = P(X_i \mid Pa(X_i))$.

4.1.7 Bayes -verkon riippumattomuusoletusten tausta

Kuten on jo tullut mainittua, Bayes -verkko koodaa joukon riippumattomuuksia rakenteessaan. Riippumattomuusoletukset pohjautuvat seuraaviin oletuksiin [8, 35]:

- **Kausaalinen riittävyys-ehto (Causal Sufficiency Assumption):** Ei ole olemassa yleisiä piilo-muuttujia (engl. *Hidden, latent*), jotka ovat havaittujen muuttujien vanhempia.
- **Markov -oletus**: Kun on annettu Bayes -verkon malli, kaikki muuttujat ovat riippumattomia ei-jälkeläisistään, ehdolla niiden vanhemmat.
- **Uskollisuus-oletus (Faithfulness Assumption)**: Bayes -verkon graafi *G* ja sen todenäköisyysjakauma *P* ovat toisilleen uskollisia siten, että todennäköisyysjakaumassa *P* on voimassa vain ne riippumattomuudet (eikä yhtään sen enempää tai vähempää), jotka on esitetty verkon rakenteessa *G*.

Määritelmä 4 (*Markov -peitto*) Kaikille muuttujille $X \in U$ Markov -peitto (engl. *Markov blanket*) $BL(X) \subseteq U$ on mikä tahansa muuttujien joukko siten, että jokaiselle muuttujalle $Y \in U - BL(X) - \{X\}, X \perp Y \mid BL(X)$.

Bayes -verkkokontekstissa muuttujan *X* Markov -peiton löytää helposti graafista. Se sisältää muuttujan *X* vanhemmat, muuttujan *X* lapset ja muuttujan *X* lapsien muut vanhemmat. Nyt ehdolliset riippumattomuudet ovat ilmaistavissa seuraavasti: Muuttuja X on ehdollisesti riippumaton verkon muista muuttujista, kun muuttujan X Markov-peitto on annettu [34].

4.2 Bayes -verkon rakenteen oppiminen datasta

4.2.1 Pistemääräfunktioihin perustuvat menetelmät

Yleisesti kun puhutaan Bayes -verkon rakenteen oppimisesta, halutaan löytää ratkaisu seuraavaan ongelmaan. Kun on annettu opetusjoukko $D=(\boldsymbol{u_1},\ldots,\boldsymbol{u_n})$ pitää löytää Bayes -verkko, joka parhaiten sopii opetusjoukkoon *D*. Yksi yleisimmistä tavoista ratkaista tämä ongelma on käyttää pistemääräfunktiota parhaimman verkkorakenteen löytämiseen. Pistemäärä-perustainen (*score-based*) menetelmä on varsin käyttökelpoinen menetelmä, varsinkin yhteistodennäköisyysjakauman estimointiin. Menetelmän idea pähkinänkuoressa on, että jollain hakualgoritmilla haetaan ”kaikki” mahdolliset Bayes -verkot, pisteytetään saadut verkot käyttäen jotain tunnettua metriikkaa ja valitaan se rakenne Bayes -verkon rakenteeksi, joka antaa parhaimman pistemääräfunktion arvon. Parhaimman pistemääräfunktion arvon antaa se rakenne, joka *parhaiten kuvaa annettua dataa* [8,10]

Parhaiten dataa kuvaa se verkkorakenne, joka maksimoi verkkorakenteen *G* posterioritodennäköisyyden $Score(G, D)=P(G|D)$, missä *D* on opetusjoukko. Käytetään Bayesin sääntöä ja saadaan

$$Score(G,D) = P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)}. \qquad (22)$$

Koska nimittäjä ei riipu rakenteesta *G*, niin maksimoitavaksi jää osoittaja. Todennäköisyys $P(G)$ voidaan jättää huomiotta, toisin sanoen oletetaan, että kaikki verkkorakenteet ovat yhtä todennäköisiä tai sitten asetetaan verkkorakenteen priorijakaumaksi jokin muu kuin tasajakauma.[8]

Kaksi usein käytettyä pistemääräfunktiota ovat *Bayesin pistemäärää (Bayesian scoring)* ja *MDL -mitta (minimun description length)*. Nämä pistemääräfunktiot ovat asymptoottisesti ekvivalentteja ja asymptoottisesti oikeita, toisin sanoen otoskoon kasvaessa datasta opitulla Bayes -verkolla koodattu yhteistodennäköisyysjakauma lähenee sitä oikeaa jakaumaa, mistä havainnot on poimittu, todennäköisyydellä 1. [13,14]

Jorma Rissasen vuonna 1978 esittämän *MDL* -periaatteen mukaan tavoitteena on löytää malli, jonka avulla havaintoaineisto voidaan kuvata mahdollisimman lyhyesti. *MDL* -periaatteeseen perustuva kaksiosainen kriteeri ottaa huomioon mallin itsessään ja mallin antaman kuvauksen datasta. Malli on tämän gradun kontekstissa Bayes -verkko, joka kuvaa datan yhteistodennäköisyysjakauman P_B. Olkoon $B = \langle G, \Theta \rangle$ Bayes -verkko ja olkoon $D=(\mathbf{u}_1,\ldots,\mathbf{u}_N)$ opetusjoukko, missä $\mathbf{u}_i$:lla merkitään kaikkien niiden muuttujien arvoja, jotka kuuluvat muuttujajoukkoon U. Bayes -verkon B *MDL* -mitta ehdolla opetusjoukko D on muotoa

$$MDL(B \mid D) = \frac{\log N}{2} \mid B \mid -LL(B \mid D), \tag{23}$$

missä $|B|$ on verkon parametrien lukumäärä. Ensimmäinen termi $\frac{\log N}{2} \mid B \mid$ kokonaisuudessaan kertoo verkon B kuvauksen pituuden siten, että lasketaan kuinka monta bittiä tarvitaan verkon B koodamiseen, kun jokaista parametria $\vartheta \in \Theta$ kohti tarvitaan $\frac{\log N}{2}$ bittiä. Toinen termi on verkon B miinusmerkkinen logaritmoitu uskottavuusfunktio, kun D on annettu:

$$LL(B \mid D) = \sum_{i=1}^{N} \log(P_B(\mathbf{u}_i)), \tag{24}$$

joka mittaa kuinka kuinka monta bittiä tarvitaan kuvaamaan opetusjoukko D pohjautuen yhteistodennäköisyysjakaumaan P_B. Tilastollinen tulkinta logaritmoidulle uskottavuudelle on, että mitä suurempi logaritmoitu uskottavuus on, sitä paremmin B mallintaa opetusjoukon D todennäköisyysjakaumaa. *MDL* -mitan ensimmäinen termi ohjaa verkon kompleksisuutta, toisin sanoen tämä termi rankaisee verkkoja, joissa on paljon parametreja. [10]

Kun *MDL* -mitta on informaatioteoreettinen kriteeri, niin bayesiläinen lähestymistapa verkkorakenteen valintaan on käyttää Bayesin Dirichlet (*BD*) pistemäärää. Maksimoitavana on nyt

$$P(G \mid D) = P(G)P(D \mid G) \propto P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_{ij}}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}, \qquad (25)$$

missä $\Gamma()$ on gammafunktio, α_{ijk} on *hyperparametri*, α_{ij} on *ekvivalentti otoskoko*, q_i on muuttujan x_i vanhempien eri tilojen yhdistelmien lukumäärä ja r_i on muuttujan x_i tilojen lukumäärä. Oletuksena on, että paikalliset todennäköisyydet noudattavat multinomijakaumaa (Dirichlet -jakaumasta tarkemmin luvussa 4.3).

Pistemääräperustaiset menetelmät siis yrittävät optimoida valitun pistemäärän. Tuloksena on verkkorakenne, joka maksimoi tämän pistemäärän. Kaikkien mahdollisten verkkorakenteiden avaruus on valitettavasti valtava; avaruus on kombinatorinen, pitäen sisällään supereksponentiaalisen lukumäärän erilaisia rakenteita. Yleisesti ongelma pistemäärän maksimoivan rakenteen löytämiseen on *NP -kova* (engl. *NP-hard*). Ratkaisu tähän ongelmaan on käyttää *heuristisia* hakualgoritmeja. Paljon käytettyjä heuristisia hakualgoritmeja ovat mm. *vuorikiipeily (hill climbing)*, *simuloitu jäähdytys (simulated annealing)*, *best first search* ja *geneettiset algoritmit*. [8,15]

4.2.1.1 Hill Climbing

Bayes -verkon rakenteen oppimista varten pitää olla päätettynä seuraavat asiat (kun käytetään pistemääräperustaista lähestymistapaa):

- Pistemäärä, jolla arvioidaan verkon hyvyyttä.
- Hakuavaruus ja siihen liittyvät sallitut operaatiot, jotka tuottavat annetusta rakenteesta toisen.
- Hakualgoritmi, joka optimoi haun.

Vuorikiipeilyalgoritmissa määritetyn hakuavaruuden tilat ovat mahdollisia verkkorakenteita ja operaatioilla määritetään verkkorakenteiden läheisyydet. Operaatiot hakuavaruudessa ovat kaaren lisäys, kaaren poisto ja kaaren kääntäminen siten että yksi operaatio kerrallaan on sallittu. Haku voidaan aloittaa täysin tyhjällä verkolla tai sitten ei-

tyhjällä, jonka asiantuntija on tuottanut alkuverkoksi. Käyttäen operaatioita löydetään suurimman pistemäärän tuottava verkko. Vaikka vuorikiipeilyalgoritmi on paljon käytetty heuristinen hakualgoritmi, sen käyttö ei ole täysin ongelmatonta, nimittäin aina ei ole takeita siitä, että globaali maksimi saavutetaan. Tämän ongelman välttämiseksi voidaan käyttää esimerkiksi menetelmiä *TABU* -haku (engl. *Tabu-search)* ja *simulated annealing*. Kolmas ehdotettu keino on käyttää satunnaisesti tuotettuja aloitusverkkoja (engl. *multiple random restarts*). [8,10,15]

4.2.2 Rajoiteperusteiset menetelmät

Pistemääräperusteinen lähestymistapa verkon rakenteen oppimiseen käyttää kulmakivenään sitä, että Bayes -verkko esittää muuttujien yhteistodennäköisyysjakauman. Rajoiteperusteinen lähestymistapa puolestaan käyttää hyväkseen tietoa siitä, että verkkorakenne pitää sisällään joukon ehdollisia riippumattomuussuhteita käsitteen d-separaatio mukaan (Pearl 1988). Rajoiteperusteinen menetelmä käyttääkin riippumattomuustestejä ehdollisten riippumattomuuksien löytämiseen attribuuttien välillä ja edelleen näitä suhteita verkon rakentamiseen. Kaksi paljon käytettyä riippumattomuustestiä tässä yhteydessä ovat X^2*-testi* ja *keskinäinen informaatio-testi (mutual information test)*. Englanninkielisessä kirjallisuudessa rajoiteperustaiset menetelmät tunnetaan nimeltä *CI-based algorithms* tai *constraint-based algorithms* [8,36].

4.3 Bayes -verkon parametrien oppiminen

4.3.1 Johdanto parametrien estimointiin

Bayes -verkkokontekstissa lähestyminen parametrien estimointiin voi olla joko frekventistinen tai bayesiläinen. Käsittelen aluksi parametrien estimointia yksisolmuisen verkon kautta, jonka jälkeen esittelen, miten estimointi tapahtuu monisolmuisten verkkojen yhteydessä. Suurimman uskottavuuden menetelmän esittelyssä pohjatietona on käytetty kurssimonisteita, sekä Pentti Huuhtasen teosta Matemaattinen tilastotiede [16].

Suurimman uskottavuuden menetelmässä lähtökohtana on valita parametrin estimaatiksi se parametriavaruuden piste, jossa saadulla kiinteällä otosavaruuden pisteen arvolla

otoksen yhteisjakauman pistetodennäköisyys saavuttaa suurimman arvonsa. Näin saatua parametrin arvoa voidaan pitää uskottavimpana ehdokkaana parametrille.

Olkoon $(X_1,...,X_n)$ satunnaisvektori, jonka yhteisjakauman tiheysfunktio tai pistetodennäköisyysfunktio on $f(x_1, ..., x_n; \theta)$, missä $\theta \in \Theta$ on tuntematon parametri. Tässä θ voi olla joko skalaari tai vektori.

Määritelmä 4 (Uskottavuusfunktio) Havaitun otoksen $(x_1, ..., x_n)$ uskottavuusfunktio määritellään parametrin $\theta \in \Theta$ funktiona $L(\theta; x_1, ..., x_n)=f(\theta; x_1, ..., x_n)$.

Diskreeteisssä tapauksissa $L(\theta; x_1, ..., x_n)$ kuvaa tapahtuman $\{X_1= x_1, ..., X_n= x_n\}$ todennäköisyyttä. Käytännössä $(X_1, ..., X_n)$ on yleensä otos jakaumasta, jonka tiheysfunktion muoto on tunnettu. Uskottavuusfunktio on nyt muotoa

$$L(\theta; x_1,...,x_n) = \prod_{i=1}^{n} P_\theta(X_i = x_i), \tag{26}$$

kun X on diskreetti.

Määritelmä 5 (*ML* -estimaatti (Suurimman uskottavuuden estimaatti)) Parametrin θ *ML* -estimaatti on sellainen parametriavaruuden Θ piste $\hat{\theta} = \hat{\theta}(x_1,...,x_n)$, joka toteuttaa ehdon $L(\hat{\theta}; x_1,...,x_n) = \sup_{\theta \in \Theta} L(\theta; x_1,...,x_n)$. *ML* -estimaattori on vastaavasti satunnaismuuttujista $(X_1, ..., X_n)$ näin riippuva otossuure $\hat{\theta}(X_1,...,X_n)$.

Koska uskottavuusfunktio on aina positiivinen, voidaan uskottavuusfunktion sijasta tarkastella logaritmoitua uskottavuusfunktiota; uskottavuusfunktio saavuttaa maksiminsa, kun sen muunnos $ln\ L(\theta; x_1, ..., x_n)$ saavuttaa maksiminsa. Tavallisesti *ML* - estimaatti löydetään siten, että ratkaistaan niin sanotut *uskottavuusyhtälöt*. Olkoon $\theta = (\theta_1,...,\theta_k)$ parametrivektori. Tällöin välttämätön ehto ääriarvokohdalle pisteessä $\hat{\theta} = (\hat{\theta}_1,...,\hat{\theta}_k)$ on, että

$$\frac{\partial}{\partial \theta_1} \ln L(\hat{\theta}_1,..,\hat{\theta}_k) = 0$$

:

$$\frac{\partial}{\partial \theta_k} \ln L(\hat{\theta}_1,..,\hat{\theta}_k) = 0 .$$

Ratkaistaan nämä uskottavuusyhtälöt ja tuloksena on haluttu maksimikohta.

Esimerkki 3 (Binäärinen muuttuja, ML -estimaatti) Olkoon Bayes -verkko yksisolmuinen verkko, jonka solmu voi saada kaksi eri arvoa, toisin sanoen $\Omega_X=(x_1, x_2)$. Tavoitteena on estimoida parametri $\theta=P(X= x_1)$. Oletetaan, että data-alkiot $D=(D_1, ..., D_m)$ ovat riippumattomia, kun θ on annettu;

$$P(D_1,...,D_m \mid \theta) = \prod_{i=1}^{m} P(D_i \mid \theta) .$$

Oletetaan myös, että tapahtumat ovat identtisesti jakautuneita, toisin sanoen $P(D_i=x_1)= \theta$ ja $P(D_i=x_2)=1- \theta$, $i=1,...,m$. Uskottavuusfunktio on nyt muotoa

$$L(\theta \mid D) = P(D \mid \theta) = P(D_1,...,D_m \mid \theta) = \prod_{i=1}^{m} P(D_i \mid \theta) = \theta^{m_1}(1-\theta)^{m_2} ,$$

missä m_1 on arvon x_1 esiintymisten lukumäärä ja vastaavasti m_2 on arvon x_2 esiintymisten lukumäärä. Estimaatti todennäköisyydelle, että seuraava data-alkio on x_1, on sama kuin parametrin θ *ML* -estimaatti.

Logaritmoitu uskottavuusfunktio on muotoa

$$\ln L(\theta \mid D) = m_1 \ln \theta + m_2 \ln(1-\theta) .$$

Derivoidaan saatu funktio parametrin θ suhteen, saadaan

$$\frac{\partial}{\partial \theta}\left[m_1 \ln \theta + m_2 \ln(1-\theta)\right] = \frac{m_1}{\theta} - \frac{m_2}{(1-\theta)} .$$

Asetetaan derivaatta nollaksi ja ratkaistaan θ:

$$\frac{m_1}{\theta} - \frac{m_2}{(1-\theta)} = 0 \Leftrightarrow \frac{(1-\theta)m_1 - \theta m_2}{(1-\theta)\theta} \Leftrightarrow (1-\theta)m_1 - \theta m_2 = 0 \Leftrightarrow \theta = \frac{m_1}{m_1 + m_2} .$$

Nyt siis

$$P(D_{m+1} = x_1 \mid D) = \frac{m_1}{m_1 + m_2} = \frac{m_1}{m}. \tag{27}$$

Esimerkki 4 (Multinomijakauma, *ML* -estimaatti) Olkoon Bayes -verkko yksisolmuinen verkko, jonka solmu/muuttuja on moniarvoinen, toisin sanoen $\Omega_X=(x_1, \ldots, x_r)$, ja olkoon $\theta_i=P(X=x_i)$, sekä $\theta=(\theta_1, \ldots, \theta_r)$. Oletuksena luonnollisesti on, että $\Sigma\, \theta_i=1$. Oletetaan, että datajoukossa D, muuttuja X saa arvon x_i m_i kertaa. Nyt uskottavuusfunktio ja logaritmoitu uskottavuusfunktio ovat muotoa

$$L(\theta \mid D) = \prod_{i=1}^{r} \theta_i^{m_i} \tag{28}$$

ja

$$\ln L(\theta \mid D) = \sum_{i=1}^{r} m_i \ln \theta_i . \tag{29}$$

Jälkimmäinen on nyt maksimoitava ehdolla, että $\Sigma\, \theta_i=1$. Määritetään ensiksi *Lagrangen funktio*

$$F(\theta_i, \lambda) = \sum_{i=1}^{r} m_i \ln \theta_i + \lambda(\sum_{i=1}^{r} \theta_i - 1), \tag{30}$$

joka sitten derivoidaan θ_i ($i=1, \ldots, r$) suhteen. Derivaataksi saadaan

$$\frac{\partial}{\partial \theta_i} F(\theta_i, \lambda) = \frac{\partial}{\partial \theta_i}\left[\sum_{i=1}^{r} m_i \ln \theta_i + \lambda(\sum_{i=1}^{r} \theta_i - 1)\right] = \frac{m_i}{\theta_i} + \lambda\,,\ i=1, \ldots, r.$$

Asetetaan tulos nollaksi ja ratkaistaan θ_i, saadaan

$$\theta_i = -\frac{m_i}{\lambda}.$$

Otetaan seuraavaksi Lagrangen funktiosta derivaatta λ:n suhteen ja asetetaan tulos nollaksi, saadaan

$$\sum_{i=1}^{r} \theta_i = 1.$$

Korvataan nyt θ_i: t äsken saadulla tuloksella ja saadaan yhtälö

$$-\frac{1}{\lambda}\sum_{i=1}^{r} m_i = 1,$$

josta saadaan, että

$$-\frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{r} m_i}$$

ja

$$\lambda = -\sum_{i=1}^{r} m_i = -m,$$

mistä seuraa, että

$$\theta_i = \frac{m_i}{m}. \tag{31}$$

Nyt halutaan ennustaa, että seuraavan alkion arvo on x_i, toisin sanoen $P(D_{m+1}=x_i|D)$. Frekventistisessä lähestymistavassahan tämä on *ML* -estimaatti, eli tässä tapauksessa

$$P(D_{m+1} = x_i \mid D) = \frac{m_i}{m}. \tag{32}$$

Bayesiläisessä lähestymistavassa parametri θ käsitetään satunnaismuuttujaksi. Tälle parametrille asetetaan priorijakauma $P(\theta)$, yhdistetään tähän prioritietoon datasta D saatu informaatio, ja saadaan parametrin θ posteriorijakauma $P(\theta|D)$. Bayes -sääntö on muotoa

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}, \tag{33}$$

joka voidaan sanallisesti kirjoittaa muotoon $posteriori \propto uskottavuus \times priori$. Todennäköisyys $P(D)$ voidaan siis jättää posterioritodennäköisyyden määrittämisestä pois, koska se on parametrista θ riippumaton vakio. Yleisesti kun priori-informaatiota on käytettävissä, priojakauman $P(\theta)$ valinnassa voidaan käyttää kahta eri menetelmää: Priorijakauma muodostetaan subjektiivisen ennakkokäsityksen perusteella esimerkiksi

asiantuntijan avustuksella tai sitten valitaan priorijakaumaksi *konjugaattinen priorijakauma*.

Kirjoitetaan Bayes -sääntö toisessa muodossa:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} = \frac{P(D \mid \theta)P(\theta)}{\sum_{\theta} P(D \mid \theta)P(\theta)}, \tag{34}$$

diskreetissä tapauksessa ja jatkuvassa tapauksessa

$$f(\theta \mid D) = \frac{f(D \mid \theta)f(\theta)}{\int f(D \mid \theta)f(\theta)d\theta}. \tag{35}$$

Tarkastellaan Bayes -säännön määritelmää jatkuvassa tapauksessa. Kaavan avulla saadaan kätevästi, ainakin periaatteessa, uuden datan perusteella muokattua tiheysfunktioita. Käytännössä kaavan soveltaminen ei kuitenkaan ole ongelmatonta. Jos uskottavuusfunktio $f(D|\theta)$ ja priorijakauma $f(\theta)$ eivät ole suhteellisen yksinkertaisia matemaattisia funktioita, voi kaavan (35) nimittäjän integrointi osoittautua hankalaksi tehtäväksi. Bayesiläiset tilastotieteilijät ovat kehittäneet käsitteen konjugaattiset priorijakaumat, jonka avulla laskenta saadaan suoritettua juohevammin. Konjugaattisten perheiden avulla posteriorijakauma voidaan esittää suljetussa muodossa. Tätä kautta ratkaisu ennustamiseen on myös löydettävissä suljetussa muodossa. [1]

Uskottavuusfunktio määrätään pohjalla olevien oletusten perusteella (otos noudattaa esimerkiksi Bernoulli -jakaumaa) yksikäsitteisesti. Konjugaattinen priorijakauma riippuu yksin ainoastaan uskottavuusfunktion muodosta. Kun priorijakaumaksi valitaan konjugaattinen priorijakauma, on myös posteriorijaukauma samasta jakaumaperheestä. Konjugaattinen priorijakauma määritetään systemaattisesti uskottavuusfunktion avulla siten, että uskottavuusfunktiossa otoksesta riippuvat termit korvataan priorijakauman parametreillä.

Esimerkki 5 (Konjugaattinen priorijakauma binomiaaliselle uskottavuudelle) Esimerkissä 3 Bernoulli -jakauman uskottavuusfuntioksi saatiin $L(\theta \mid D) = \theta^{m_1}(1-\theta)^{m_2}$. Korvataan nyt otoksesta riippuvat termit priorijakauman parametreilla: $m_1 \rightarrow \alpha_1$ ja $m_2 \rightarrow \alpha_1$. Uskottavuusfunktio on nyt muotoa $L(\theta \mid D) = \theta^{\alpha_1}(1-\theta)^{\alpha_2}$, joka muistuttaa, vakiota

vailla olevaa *Beta-jakauman* tiheysfunktiota. Nyt priorijakauma on $P(\theta) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$, jonka normalisoiva vakio on

$$\frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} = \frac{1}{\mathrm{B}(\alpha_1, \alpha_2)}.$$

Mille tahansa kokonaisluvulle γ gammafunktio määritetään $\Gamma(\gamma)=(\gamma-1)!$. Parametrin θ priorijakauma on beta-jakauma; $\theta \sim Beta(\alpha_1, \alpha_2)$. Posteriorijakauma on muotoa

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{\int P(D \mid \theta)P(\theta)d\theta} = \frac{\frac{1}{\mathrm{B}(\alpha_1, \alpha_2)}\theta^{m_1+\alpha_1-1}(1-\theta)^{m_2+\alpha_2-1}}{\int_{-1}^{1}\frac{1}{\mathrm{B}(\alpha_1, \alpha_2)}\theta^{m_1+\alpha_1-1}(1-\theta)^{m_2+\alpha_2-1}d\theta}. \qquad (36)$$

Täydennetään nimittäjää sopivasti ja saadaan

$$\frac{\mathrm{B}(m_1+\alpha_1, m_2+\alpha_2)}{\mathrm{B}(\alpha_1, \alpha_2)}\int_{-1}^{1}\frac{1}{\mathrm{B}(m_1+\alpha_1, m_2+\alpha_2)}\theta^{m_1+\alpha_1-1}(1-\theta)^{m_2+\alpha_2-1}d\theta,$$

$$= \frac{\mathrm{B}(m_1+\alpha_1, m_2+\alpha_2)}{\mathrm{B}(\alpha_1, \alpha_2)} \qquad (37)$$

mistä seuraa, että posteriorijakauma on

$$P(\theta \mid D) = \frac{1}{\mathrm{B}(m_1+\alpha_1, m_2+\alpha_2)}\theta^{m_1+\alpha_1-1}(1-\theta)^{m_2-\alpha_2-1}. \qquad (38)$$

Myös posteriorijakauma on täten beta -jakauma; $\theta|D \sim Beta(\alpha_1+m_1, \alpha_2+m_2)$.

Esimerkki 6 (Konjugaattinen priorijakauma Multinominaaliselle uskottavuudelle) Esimerkissä 3 moniarvoisen muuttujan uskottavuusfunktioksi saatiin $L(\theta \mid D) = \prod_{i=1}^{r}\theta_i^{m_i}$.

Multinomiaaliselle uskottavuudelle konjugaattinen perhe on *Dirichlet-jakaumat*. Yleisesti Dirichlet -jakauman tiheysfunktio on

$$P(\theta) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{r}\Gamma(\alpha_i)}\prod_{i=1}^{r}\theta_i^{\alpha_1-1}, \qquad (39)$$

missä $\alpha=\alpha_1+\ldots+\alpha_r$. Posteriorijakauma on nyt

$$P(\theta \mid D) \propto L(\theta \mid D)P(\theta) \propto \prod_{i=1}^{r} \theta_i^{m_i} \prod_{i=1}^{r} \theta_i^{\alpha_i - 1}$$

$$= \prod_{i=1}^{r} \theta_i^{(m_i+\alpha_i)-1} \propto \frac{\Gamma(\alpha + m)}{\prod_{i=1}^{r} \Gamma(\alpha_i + m_i)} \prod_{i=1}^{r} \theta_i^{(m_i+\alpha_i)-1}, \qquad (40)$$

missä $m=m_1+\ldots+m_r$. Nyt $\theta \sim Dir(\alpha_1, \ldots, \alpha_r)$ ja $\theta|D \sim Dir(m_1+\alpha_1, \ldots, m_r+\alpha_r)$.

Edellä esitettiin konjugaattisen priorijakauman käsite. Luonnollinen kysymys mielestäni on kysymys priorijakauman parametreistä; miten määrätä konjugaattisen priorijakauman parametrit? Käytetään esimerkkinä saatua konjugaattista priorijakaumaa binomiaaliselle uskottavuudelle, siis beta-jakaumaa $Beta(\alpha_1, \alpha_2)$. Jos on olemassa näkemys beta-jakauman odotusarvosta ja varianssista, niin parametrit voidaan ratkaista näiden avulla. Esimerkkinä tilanne, jossa tilastotieteen professori pähkäilee, että selvittääkö Maija hänen tulevan kurssinsa. Edellisvuosien perusteella hän arvioi, että keskimäärin hänen kurssinsa on selvittänyt 70 % opiskelijoista varianssin ollessa 0,02. Beta-jakauman odotusarvo ja varianssi ovat muotoa

$$E(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_2} \overset{sijoitus}{=} 0{,}7 \qquad \text{ja} \qquad Var(\theta) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)} \overset{sijoitus}{=} 0{,}02\,.$$

Ratkaistaan parametrit kaavoista ja parametreille saadaan arvot $\alpha_1=7$ ja $\alpha_2=3$.

Esitelty menetelmä on teoriassa käyttökelpoinen menetelmä priorijakauman parametrein määrämiseen. Käytännössä tätä menetelmää voi olla hankala soveltaa, nimittäin jakauman varianssin asettaminen intuition pohjalta on vaikeaa. Samantyyppinen menetelmä kuin mitä edellä esitettiin, on käyttää hyväksi jakauman fraktiili-paria tai keskiarvoa ja yhtä fraktiilia. Päätöksentekijällä voi olla esimerkiksi näkemys, että parametrin priorijakauma noudataa beta-jakaumaa odotusarvonaan 0,25 ja 25 % fraktiilinaan 0,2. Näiden suureiden avulla ratkaistaan kuten edellä priorijakauman parametrit. [1]

Toisentyyppinen tapa ratkaista asia, on käyttää niin sanottua kuvitteellista otostietoa hyväksi. Olkoon θ niiden opiskelijoiden osuus, joka läpäisee erään tilastotieteen kurssia ja oletetaan, että prioritieto voidaan esittää beta-jakauman avulla. Parametrin θ odotettu

arvo on 0,2. Lisäksi väitetään, että jos havainnoidaan 100 opiskelija otos, niin vain 10 opiskelijaa läpäisisi kurssin ja siten parametrin θ odotusarvo laskisi arvoon 0.18. Toisin sanoen, nykyinen odotusarvo on

$$\frac{\alpha_1}{\alpha_1 + \alpha_2} \overset{sijoitus}{=} 0{,}2 \quad ,$$

ja posteriorijakauman odotusarvo on

$$E(\theta \mid D) = \frac{\alpha_1 + 10}{\alpha_1 + \alpha_2 + 100} \overset{sijoitus}{=} 0{,}18$$

Kun näistä kahdesta odotusarvosta ratkaistaan prioriparametrit, saadaan $\alpha_1=80$ ja $\alpha_2=400$.

Kun halutaan määrittää todennäköisyys $P(D_{m+1}=x_i|D)$, missä data pitää sisällään riippumattomia identtisesti jakautuneita muuttujia, on vaihtoehtoja useampi kuin yksi riippuen siitä, halutaanko löytää yksittäinen arvo parametreille, vai halutaanko käyttää hyväksi koko posteriorijakauman $P(\theta|D)$ antama informaatio. Jälkimmäisen lähestymistavan mukaan bayesiläinen ennustaminen pohjautuu mallien ennusteiden keskiarvottamiseen, painotettuna mallin parametrien posterioritodennäköisyyksillä:

$$P(D_{m+1} \mid D) = \int P(D_{m+1}, \theta \mid D) d\theta = \int P(D_{m+1} \mid \theta) P(\theta \mid D) d\theta \,. \qquad (41)$$

Palataan esimerkkiin diskreetistä kaksiarvoisesta muuttujasta. Määritetään nyt Bayes -ennuste tulevalle havainnolle D_{m+1}. Määritetään ensin $P(D_{m+1}|\theta,D)$:

$$P(D_{m+1} \mid \theta, D) = \theta^{D_{m+1}} (1-\theta)^{1-D_{m+1}} \,. \qquad (42)$$

Posterioritodennäköisyydeksihän saatiin

$$P(\theta \mid D) = \frac{1}{\mathrm{B}(m_1 + \alpha_1, m_2 + \alpha_2)} \theta^{m_1 + \alpha_1 - 1} (1-\theta)^{m_2 - \alpha_2 - 1} \,.$$

Nyt, koska

$$P(D_{m+1},\theta|D)=P(D_{m+1}|\theta, D)P(\theta|D),$$

niin

$$P(D_{m+1}, \theta \mid D) = \frac{\theta^{D_{m+1}+A}(1-\theta)^{B+1-D_{m+1}}}{\mathrm{B}(A+1, B+1)}, \tag{43}$$

missä $A=m_1+\alpha_1-1$ ja $B= m_2+\alpha_2-1$ ja

$$\int P(D_{m+1}, \theta \mid D) = \frac{1}{\mathrm{B}(A+1, B+1)} \int \theta^{D_{m+1}+A}(1-\theta)^{B+1-D_{m+1}} d\theta = \frac{\mathrm{B}(D_{m+1}+A+1, B+2-D_{m+1})}{B(A+1, B+1)}$$

.

Nyt ennusteet ovat

$$P(D_{m+1}=1 \mid D) = \frac{\mathrm{B}(A+2, B+1)}{\mathrm{B}(A+1, B+1)} = \frac{A+1}{A+B+1} = \frac{m_1+\alpha_1}{m+\alpha_1+\alpha_2} = \frac{m_1+\alpha_1}{m+\alpha} \tag{44}$$

$$P(D_{m+1}=0 \mid D) = \frac{\mathrm{B}(A+1, B+2)}{\mathrm{B}(A+1, B+1)} = \frac{B+1}{A+B+2} = \frac{m_2+\alpha_2}{m+\alpha}. \tag{45}$$

Vastaavalla tavalla saadaan määritettyä Bayes-ennuste havainnolle D_{m+1} moniarvoisen diskreetin muuttujan tapauksessa. Posterioritodennäköisyydeksi saatiin

$$P(\theta \mid D) = \frac{\Gamma(\alpha+m)}{\prod_{i=1}^{r}\Gamma(\alpha_i+m_i)} \prod_{i=1}^{r} \theta_i^{(m_i+\alpha_i)-1}$$

Ja, koska

$$P(D_{m+1} \mid \theta, D) = \theta_1^{I(D_{n01}=1)} \cdot \theta_2^{I(D_{n01}=2)} \cdot \ldots \cdot \theta_r^{I(D_{n01}=r)} = \prod_{i=1}^{r} \theta_i^{I(D_{m+1}=i)},$$

missä $I(.)$ on indikaattorifunktio, niin

$$P(D_{m+1}, \theta \mid D) = \frac{\Gamma(\alpha+m)}{\prod_{i=1}^{r}\Gamma(\alpha_i+m_i)} \prod_{i=1}^{r} \theta_i^{\alpha_i+m_i-1+I(D_{m+1}=i)}, \tag{46}$$

mistä seuraa, kun merkitään $m_i^{'} = m_i - 1 + I(D_{m+1}=i)$, että

$$P(D_{m+1} \mid D) = \frac{\Gamma(\alpha + m)}{\prod_{i=1}^{r} \Gamma(\alpha_i + m_i)} \cdot \frac{\prod_{i=1}^{r} \Gamma(\alpha_i + m_i')}{\Gamma(\alpha + m + 1)} \quad . \tag{47}$$
$$= \frac{\Gamma(\alpha_i + m_i')}{\Gamma(\alpha_i + m_i)} \times \frac{1}{\alpha + m} = \frac{\alpha_i + m_i}{\alpha + m}, D_{m+1} = i$$

Kun halutaan käyttää piste-estimaattia, on tässäkin valittavana teoriassa useampi kuin yksi lähestymistapa; Posteriorijakauman odotusarvo, moodi tai mediaani. Posteriorijakauman mediaanin käyttöä Bayes -estimaattina ei kylläkään diskreettien Bayes -verkkojen tapauksessa suosita, koska mediaanin määrittäminen ei ole suoraviivaista.

Bayes -estimaattorin takana on valittu *tappiofunktio L(T; θ)*, joka ilmaisee estimaattorin $T(X_1, \ldots, X_n)$ arvon poikkeaman parametrin θ arvosta aiheutuvan tappion. *Riskifunktio R(T; θ)* on tappiofunktion odotusarvo satunnaisotoksen yhteisjakauman suhteen:

$$R(T;\theta) = E_\theta\left[L(T;\theta)\right] = E\left[L(T(x);\theta) \mid \theta\right] = \int L(T(x);\theta)P(x \mid \theta)dx\,. \tag{48}$$

Riskifunktio ilmaisee sen keskimääräisen tappion, joka aiheutuu estimaattorin $T(X_1, \ldots, X_n)$ valinnasta. Koska todellista parametrin θ arvoa ei luonnollisestikaan tunneta, valitaan se estimaattori, jolla on pienin riskifunktion arvo kaikilla θ:n arvoilla. Kun painotetaan riskifunktiota priojakauman sisältämällä informaatiolla, toisin sanoen kun otetaan riskifunktiosta odotusarvo priorijakauman $P(\theta)$ suhteen, saadaan *Bayes -riski*

$$R(P_\theta) = E_{P_\theta}\left[R(T;\theta)\right] = \int R(T;\theta)P(\theta)d\theta\,, \tag{49}$$

jota minimoimalla minimoidaan estimaattorin $T(X_1, \ldots, X_n)$ arvon poikkeaman parametrin θ arvosta aiheutuvan tappio. Estimaattoria, joka minimoi Bayes-riskin, kutsutaan Bayes -estimaattoriksi. Kun määritetään Bayes -estimaattoria, voidaan Bayes -riskin minimoinnin sijasta minimoida posteriorijakauman suhteen laskettua tappiofunktion odotusarvoa.

Olkoon,

$$
\begin{aligned}
&R(P_\theta) = E_{P_\theta}\left[R(T;\theta)\right] = E_{P_\varpi}\left[E_\theta\left[L(T;\theta)\right]\right] = \int E_\theta\left[L(T;\theta)\right]P(\theta)d\theta \\
&= \int\left[\int L(T;\theta)f(x;\theta)dx\right]P(\theta)d\theta = \int\left[\int L(T;\theta)f(x;\theta)P(\theta)d\theta\right]dx \\
&= \int\left[\int L(T;\theta)P(\theta \mid x)P(x)d\theta\right]dx \\
&= \int P(x)\left[\int L(T;\theta)P(\theta \mid x)d\theta\right]dx,
\end{aligned}
\tag{50}
$$

mistä näkee, että Bayes -riski minimoituu, kun $E_{\theta|x}[L(T;\theta)]$ minimoituu.

Piste-estimoinnissa tappiofunktio heijastaa valitun estimaatin arvon ja todellisen estimaatin arvon yhteensopivuutta. Merkitään nyt valittua arvoa notaatiolla *a* ja todellista parametrin arvoa notaatiolla θ. Jos *a* on lähellä todellista parametrin arvoa θ, niin aiheutuva tappio ja tappiofunktion arvo $L(a, \theta)$ ovat pieniä. Kolme yleisesti käytettyä tappiofunktiota ovat *neliöllinen tappiofunktio*, *itseisarvotappiofunktio* ja *0-1 tappiofunktio*.

Neliöllinen tappiofunktio, joka on muotoa $L(\theta, a)=(\theta-a)^2$, rankaisee suurista poikkeamista ankarammin, kun taas itseisarvotappiofunktio $L(\theta, a)=|\theta-a|$ ja 0-1 tappiofunktio

$$
L(\theta, a) = \begin{cases} 0, & |a-\theta| \le b \\ 1, & |a-\theta| > b \end{cases}
\tag{51}
$$

ovat vähemmän herkkiä suurille poikkeamille. Neliölliseen tappiofunktioon liittyvä Bayes-estimaattori on posteriorijakauman odotusarvo

$$\hat{\theta}_B = \int \theta P(\theta \mid D)d\theta .$$

ja itseisarvotappifunktioon liittyvä Bayes-estimaattori on posteriorijakauman $P(\theta|D)$ mediaani. 0-1 tappiofunktioon liitetään yleensä niin sanottu *MAP -estimaattori*, toisin sanoen valitaan se estimaatin arvo, joka maksimoi posteriorijakauman: $\hat{\theta}_B = \arg\max_\theta P(\theta \mid D)$. Estimaatin arvo, joka maksimoi posteriorijakauman, on tunnetusti moodi. Posteriorijakauman maksimoinnin sijaan voidaan tarkastella logaritmoidun posteriorijakauman maksimointia:

$$\hat{\theta}_B = \arg\max_\theta P(D \mid \theta)P(\theta) = \arg\max_\theta \ln P(D \mid \theta) + \ln P(\theta) .$$

Ensimmäinen termihän on logaritmoitu uskottavuusfunktio. Maksimointitehtävä eroaa siis suurimman uskottavuuden maksimoinnista termin *ln P(θ)* osalta; *MAP* -estimaattoria kutsutaankin *rangaistuksi suurimman uskottavuuden estimaattoriksi* (*penalized maximum likelihood estimator*) [18]. Selvää on, että kun parametrin θ priorijakauma on tasajakauma, niin *ML*-estimaattori ja *MAP* -estimattori ovat samoja. Kun taas otoskoko on suuri, niin *ML* -estimaattori ja *MAP* -estimaattori lähenevät toisiaan; tällöin parametrin θ priorijakauman merkitys heikkenee.

Viitaten edellisiin esimerkkeihin, joissa posteriorijaukamiksi saatiin *Beta*($\alpha_1+m_1, \alpha_2+m_2$)-jakauma ja *Dir*($\alpha_1+m_1, \ldots, \alpha_i+m_i$) -jakauma, määritän seuraavaksi näitä jakaumia vastaavat parametrin θ piste-estimaatit. *Beta*($\alpha_1+m_1, \alpha_2+m_2$) -jakauman odotusarvo ja moodi ovat muotoa

$$\begin{aligned} &P(D_{m+1} = x_1 \mid D) = E(\theta \mid D) = \int \theta f(\theta \mid D) d\theta \\ &= c\int_0^1 \theta\theta^{(m_1+\alpha_1)-1}(1-\theta)^{(m_2+\alpha_2)-1} d\theta = \frac{m_1+\alpha_1}{m_1+\alpha_1+m_2+\alpha_2} = \frac{m_1+\alpha_1}{m+\alpha} \end{aligned} \tag{52}$$

ja

$$\begin{aligned} &P(D_{m+1} = x_1 \mid D) = \arg\max_{\theta} P(\theta \mid D) \\ &= \frac{\alpha_1+m_1-1}{\alpha_1+m_1+\alpha_2+m_2-2} = \frac{\alpha_1+m_1-1}{m+\alpha-2}. \end{aligned} \tag{53}$$

Jakauman *Dir*($\alpha_1+m_1, \ldots, \alpha_r+m_r$) i. komponentin odotusarvo on

$$E(\theta_i \mid D) = \int \theta_i f(\theta \mid D) d\theta = \frac{\alpha_i+m_i}{\alpha+m}, \tag{54}$$

missä $\alpha = \alpha_1+\ldots+\alpha_r$ ja $m = m_1+\ldots+m_r$. Dirichlet-jakauman moodin i. komponentti on

$$\hat{\theta}_{i,MAP} = P(D_{m+1} = x_i \mid D) = \arg\max_{\theta} P(\theta_i \mid D) = \frac{\alpha_i+m_i-1}{m+\alpha-r}. \tag{55}$$

Huomataan, että sekä kaksiarvoisen että moniarvoisen muuttujan tapauksessa parametrin θ posteriorijakauman odotusarvo on sama kuin mitä saatiin laskemalla Bayes-ennuste. Frekventistinen ratkaisu todennäköisyyteen $P(D_{m+1} = x_1 \mid D)$ on $\hat{\theta}_{ML} = \frac{m_1}{m}$.

Posteriorijakakauman odotusarvo ja *ML* -estimaattori eroavat toisistaan termien α_1 ja $\alpha_1+\alpha_2$ osalta. Kun $\alpha_1=\alpha_2=0$, niin odotusarvo ja *ML* -estimaattori ovat samat. Jos taas $\alpha_1=\alpha_2=1$ (priorijakaumana tasajakauma), niin *MAP* -estimaattori ja *ML* -estimaattori ovat samat.

4.3.2 Verkon parametrien estimointi

Olkoon G verkko, jossa on n-kappaletta diskreettejä muuttujia $(X_1,\ldots, X_n)$. Merkitään termillä r_i muuttujan X_i $(i=1,\ldots, n)$ eri arvojen lukumäärä; $r_i = \left|\Omega_{X_i}\right|$. Vastaavalla tavalla $q_i = \left|\Omega_{Pa(X_i)}\right|$, mikä siis tarkoittaa sitä, että q_i kertoo muuttujan X_i vanhempien eri tilojen lukumäärän. Esimerkiksi jos muuttujan X_1 toinen vanhempi voi saada kaksi arvoa ja toinen vanhempi kolme arvoa, niin tällöin $q_1 = \left|\Omega_{Pa(X_1)}\right| = 2\times 3 = 6$. Estimoitavat parametrit ovat nyt

$$\begin{aligned}&\theta_{ijk} = P(X_i = k \mid Pa(X_i) = j), i = 1,\ldots,n; k = 1,\ldots,r_i;\\ &j = 1,\ldots,q_i (\sum_k \theta_{ijk} = 1 \forall i, j)\end{aligned} \quad . \tag{56}$$

Nyt θ on parametrivektori: $\theta=\{\ \theta_{ijk}|i=1,\ldots,\ n;\ j=1,\ldots,\ q_i;\ k=1,\ldots,\ r_i\}$ ja θ_{ij} on parametrivektori todennäköisyyksille $P(X_i|Pa(X_i)=j)$, ts. $\theta_{ij}=\{\ \theta_{ijk}|k=1,\ldots,\ r_i\}$. Parametrivektori θ_i on vektori todennäköisyyksille $P(X_i|Pa(X_i))$; $\theta_i=\{\theta_{ijk}|j=1,\ldots,\ q_i;\ k=1,\ldots,\ r_i\ \}$. Olkoon D_l havainto; vektori, jossa on jokaista muuttujaa kohden arvo. Määritetään karakteristinen funktio

$$\chi(i, j, k : D_l) = \begin{cases} 1 & , X_i = k, Pa(X_i) = j \\ 0 & \quad muuten \end{cases} . \tag{57}$$

Merkitään $m_{ijk} = \sum_l \chi(i, j, k : D_l)$, mikä on niiden tapausten lukumäärä, joissa $X_i=k$ ja $Pa(X_i)=j$. Nyt logaritminen uskottavuus on

$$\ln L(\theta \mid D_l) = \ln \prod_l P(D_l \mid \theta) = \sum_l \ln P(D_l \mid \theta)$$
$$= \sum_l \sum_{i,j,k} \chi(i,j,k : D_l) \ln \theta_{ijk}$$
$$= \sum_{i,j,k} \sum_l \chi(i,j,k : D_l) \ln \theta_{ijk} \quad . \tag{58}$$
$$= \sum_{i,j,k} m_{ijk} \ln \theta_{ijk}$$
$$= \sum_{i,j} \sum_k m_{ijk} \ln \theta_{ijk}$$

Parametrin θ_{ijk} *ML* -estimaatti löydetään nyt maksimoimalla saatu logaritminen uskottavuus:

$$\arg\max_\theta l(\theta \mid D) = \arg\max_{\theta_{ijk}} \sum_{i,j} \sum_k m_{ijk} \ln \theta_{ijk} \ . \tag{59}$$

Nyt on huomioitava, että $\theta_{ijk}=(X_i=k|Pa(X_i)=j)$ ja $\theta_{i'j'k'}=(X_{i'}=k'|Pa(X_{i'})=j')$ eivät ole sama asia, jos $i \neq i'$ tai $j \neq j'$. Näin ollen jokainen summan $\sum_{i,j}$ termi voidaan maksimoida erikseen, toisin sanoen

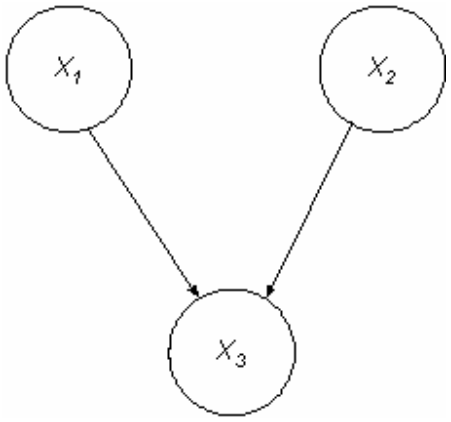
$$\arg\max_\theta l(\theta \mid D) = \arg\max_{\theta_{ijk}} \sum_k m_{ijk} \ln \theta_{ijk} \ . \tag{60}$$

ML -estimaatti on nyt

$$\hat{\theta}_{ijk} = \frac{m_{ijk}}{\sum_k m_{ijk}} . \tag{61}$$

Esimerkki 7 (Yksinkertainen Bayes -verkko, *ML* -estimaatti) Kuvassa 6 on yksinkertainen Bayes-verkko, jossa muuttujalla X_3 on kaksi vanhempaa ja muuttujilla X_1 ja X_2 on lapsenaan X_3. Havainnot on esitetty verkon vieressä. Havaintoja on kymmenen ja muuttujat ovat binäärisiä. Nyt $r_1= r_2= r_3=2$ ja $q_3=2^3=8$. *ML* -estimaatti parametrille $\theta_{312}=P(X_3=2|X_1=1,X_2=1)$ on

$$\hat{\theta}_{312} = \frac{m_{312}}{\sum_{k=1}^{2} m_{31k}} = \frac{1}{2} .$$



X1	X2	X3
1	1	1
1	2	2
1	1	2
1	2	2
2	1	1
2	1	2
2	2	1
2	2	1
1	2	2
2	1	1

Kuva 6 Kolmisolmuinen Bayes -verkko ja muuttujia vastaavat havainnot

Olkoon θ satunnaismuuttujien vektori johon liittyy prioritieto $P(\theta)$. Posteriorijakauma on muotoa

$$P(\theta \mid D) \propto P(\theta)L(D \mid \theta) = P(\theta)\prod_{i,j}\prod_{k}\theta_{ijk}^{\ m_{ijk}} \ . \tag{62}$$

Kun oletetaan *parametrien paikallinen ja globaali riippumattomuus* [8,17,19], niin

$$P(\theta) = \prod_{i,j} P(\theta_{ij}) \tag{63}$$

Nyt posteriorijakauma on muotoa

$$\begin{aligned} P(\theta \mid D) &\propto P(\theta)L(D \mid \theta) = P(\theta)\prod_{i,j}\prod_{k}\theta_{ijk}^{\ m_{ijk}} \\ &= \left[\prod_{i,j} P(\theta_{ij})\right]\prod_{i,j}\prod_{k}\theta_{ijk}^{\ m_{ijk}} = \prod_{i,j} P(\theta_{ij})\prod_{k}\theta_{ijk}^{\ m_{ijk}} \end{aligned}, \tag{64}$$

joka voidaan kirjoittaa muodossa

$$P(\theta \mid D) = \prod_{i,j} P(\theta_{ij} \mid D)\,, \tag{65}$$

missä

$$P(\theta_{ij} \mid D) \propto P(\theta_{ij})\prod_{k}\theta_{ijk}^{\ m_{ijk}} \ .$$

$P(\theta_{ij})$ kuuluu Dirichlet -perheeseen:

$$P(\theta_{ij}) = \Gamma(\alpha_{ij}) \prod_{k=1}^{r_i} \frac{\theta_{ijk}^{\alpha_{ijk}-1}}{\Gamma(\alpha_{ijk})}, \tag{66}$$

missä $\alpha_{ij} = \alpha_{ij1} + \alpha_{ij2} + ... + \alpha_{ijr_i}$. Nyt priorijakauma $P(\theta_{ij}) \sim Dir(\alpha_{ij1}, \alpha_{ij2}, ..., \alpha_{ijr_i})$. Hyperparametrit α_{ijk} välittävät siis havainnoijan prioriuskomuksen; hyperparametrit voidaan käsittää kuvitteellisten tapausten (*imaginary counts*) frekvensseiksi. Ekvivalentti otoskoko (*equivalent sample size*) α_{ij} kertoo kuinka varma havainnoija on prioritietämyksestään. Toisin sanoen, mitä suurempi α_{ij} on, sitä vakuuttuneempi havainnoija on omasta priorinäkemyksestään. Posteriorijakauma noudattaa nyt myös Dirichlet -jakaumaa:

$$P(\theta_{ij} \mid D) \sim Dir(\alpha_{ij1} + m_{ij1}, \alpha_{ij2} + m_{ij2}, ..., \alpha_{ijr_i} + m_{ijr_i}),$$

jonka tiheysfunktio on muotoa

$$P(\theta_{ij} \mid D) \propto \prod_{k}^{r_i} \theta^{m_{ijk} + \alpha_{ijk} - 1} . \tag{67}$$

Tämän posteriorijakauman k. komponentin odotusarvo on

$$\tilde{\theta}_{ijk} = E(\theta_{ijk} \mid D) = \frac{\alpha_{ijk} + m_{ijk}}{\alpha_{ij} + m_{ij}}, \tag{68}$$

missä $m_{ij} = m_{ij1} + m_{ij2} + ... + m_{ijr_i}$. Aikaisemmin saatiin vastaavasti *ML* -estimaatiksi

$$\hat{\theta}_{ijk} = \frac{m_{ijk}}{\sum_{k} m_{ijk}} = \frac{m_{ijk}}{m_{ij}} .$$

ML -estimaatti ja Bayes -estimaatti eroavat näin ollen toisistaan prioritiedon verran. Posteriorijakauman moodin avulla saatava *MAP* -estimaatti on nyt

$$\hat{\theta}_{ijk}^{\ MAP} = \arg\max_{\theta_{ijk}} P(\theta_{ijk} \mid D) = \frac{\alpha_{ijk} + m_{ijk} - 1}{m_{ij} + \alpha_{ij} - r_i} . \tag{69}$$

Kun muistellaan mitä *MAP -estimaatit* ovat yksisolmuisen kaksi- ja moniarvoisen muuttujan tapauksessa, niin huomataan viimeistään tässä vaiheessa, että nimittäjissä luvut *2*, *r* ja r_i ovat estimoitavien parametrien lukumäärä.

4.3.2.1 Laplace -estimaatti ja m -estimaatti luokittelutehtävässä

Bayes -verkkojen parametrien estimoinnissa, johtuen varsinkin puutteellisista ohjelmista, käytetään parametrien estimoinnissa varsin usein frekventistä lähestymistapaa. Nollatodennäköisyyksien välttämiseksi lisätään *ML* -estimaattiin niin sanottua täristystä Laplace -tai *M* -estimaatin muodossa. Luvussa 4.4.1 esitellään Naiivi Bayes -luokittelija. Havainto luokitellaan kuuluvaksi luokkaan C_k yhteistodennäköisyysjakauman $P(C_k)\prod_{i=1}^{n} P(X_i \mid C_k)$ perusteella. Jos opetusjoukossa ei ole esiintynyt tietyn luokan edustajia, niin tällöin *ML* -estimaatti on huono estimaatti, nimittäin tällöin ehdollinen todennäköisyys $P(X_i|C_k)$ olisi nolla. Esimerkkinä olkoon tilanne, jossa potilas luokitellaan kuuluvaksi tautiryhmään A tai B (luokittelumuuttuja C on binäärinen) muuttujien x_1, x_2 ja x_3 arvojen perusteella. Diskreettit muuttujat x_1, x_2 ja x_3 ovat kolmiarvoisia. Käytetty luokittelija on Naiivi-luokittelija. Taulukossa 1 on opetusjoukon perusteella lasketut todennäköisyydet. Olkoon luokittelukohteena oleva havainto seuraavanlainen: (2, 3, 1). Havainto määritetään siiheen luokkaan, jonka arvo maksimoi posteriorijakauman, toisin sanoen

$$c_{map} = \arg\max_{C_k \in C} P(c_k)\prod_i P(x_i \mid c_k) \text{ [5],} \tag{70}$$

niin saadaan todennäköisyydet

$$\mathrm{P(A)\,P(x_{12} \mid A)\,P(x_{23} \mid A)\,P(x_{31} \mid A)} = \frac{9}{14}\times\frac{4}{9}\times\frac{3}{9}\times\frac{3}{9} = 0{,}03175$$

$$\mathrm{P(B)\,P(x_{12} \mid B)\,P(x_{23} \mid B)\,P(x_{31} \mid B)} = \frac{5}{14}\times 0\times\frac{1}{15}\times\frac{3}{5} = 0\,.$$

Opetusjoukon huono edustavuus toisin sanoen veti toisen posterioritodennäköisyyden nollaksi. Yksi ratkaisu on käyttää *m -estimaattia*, joka on esitetty kaavassa 71. Termi $|x_{ijk}|$ on niiden havaintojen lukumäärä, jotka saavat muuttujalle x_i arvon j, ja jotka saavat luokittelumuuttujan c arvon k. Termi n_k on luonnollisesti luokkaan c_k kuuluvien lukumäärä. Termi p on prioritodennäköisyys $P(x_{ijk} \mid c_k)$. Termi m on vakio, niin sanottu kuvitteellinen otoskoko, joka on havainnoijan painoarvo prioritodennäköisyydelle. [5]

$$P(x_{ijk} \mid c_k) = \frac{|\, x_{ijk} \mid + mp}{n_k + m}. \tag{71}$$

Laplace-estimaatti, m -estimaatin vähemmän sofistikoitunut versio, on muotoa

$$P(x_{ijk} \mid c_k) = \frac{|\, x_{ijk} \mid + 1}{n_k + k}, \tag{72}$$

jossa nimittäjässä oleva k viittaa muuttujan x_i arvojen lukumäärään. Käyttämällä jompaakumpaa estimaattia, vältytään posterioritodennäköisyyden menemiseltä nollaksi opetusjoukon edustavuuden takia.

Taulukko 1 Esimerkin potilaiden luokittelu tautiryhmiin: priorijakaumat

P(A)=9/14	P(B)=5/14	$P(x_{23}\|A)=3/9$	$P(x_{21}\|B)=2/5$
$P(x_{11}\|A)=2/9$	$P(x_{12}\|A)=4/9$	$P(x_{22}\|B)=2/5$	$P(x_{23}\|B)=1/5$
$P(x_{13}\|A)=3/9$	$P(x_{11}\|B)=3/5$	$P(x_{31}\|A)=3/9$	$P(x_{32}\|A)=1/9$
$P(x_{12}\|B)=0$	$P(x_{13}\|B)=2/5$	$P(x_{33}\|A)=5/9$	$P(x_{31}\|B)=3/5$
$P(x_{21}\|A)=2/9$	$P(x_{22}\|A)=4/9$	$P(x_{32}\|B)=1/5$	$P(x_{33}\|B)=1/5$

4.4 Bayes -verkot luokittelijana

4.4.1 Naiivi Bayesin luokittelija

Bayes -verkko on rakennettu niin, että *Markov -oletus* on voimassa. Toisin sanoen $X_i \perp NonDesc(X_i) \mid Pa(X_i)$. Naiivi Bayesin -luokittelija perustuu siihen oletukseen, että muuttujat ovat toisistaan riippumattomia, kun luokittelumuuttuja on annettu. Tämä seuraa suoraan *Markov -oletuksest*a ja naiivin Bayesin -luokittelijan graafisesta esityksestä.

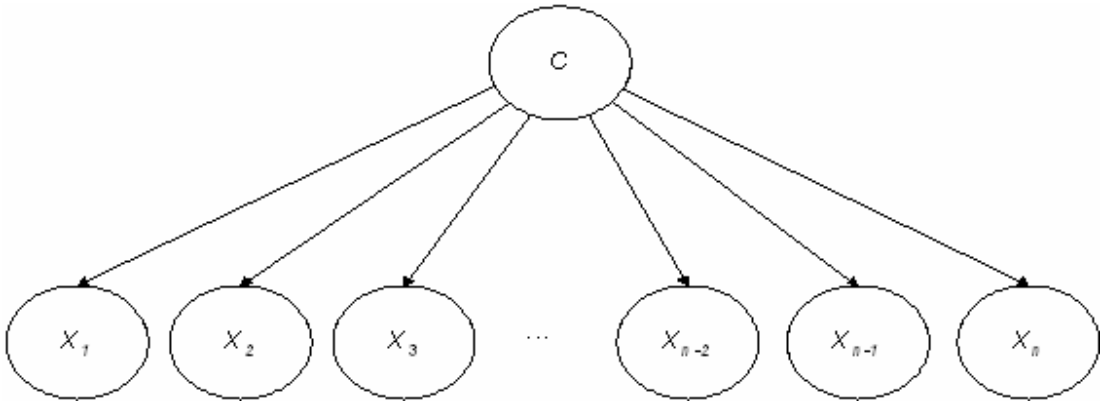
Kuvassa 7 on kuvattu naiivin Bayesin -luokittelijan rakenne. Graafi on tämän, kuten muutaman muunkin (katso seuraava luku) luokittelijan tapauksessa *puu-rakenne*. Puu on abstrakti tietotyyppi, joka tallettaa alkiot hierarkkisesti. Päällimmäistä solmua lukuun

ottamatta kaikilla solmuilla on vanhempi. Luvussa 4.1.1 määriteltiin solmu *juuri*. Kuvan 3 puussa juurisolmuna on solmu *C*. Tämä solmu on luokittelumuuttuja. Bayes -verkon avulla esitetään muuttujien yhteistodennäköisyysjakauma, joka siis luvun 4.1.3 mukaan on

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) .$$

Nyt kun verkon rakenne on tämä, että kaikille attribuuteilla on yksi ja sama vanhempi nimittäin solmu *C*, niin yhteistodennäköisyysjakauma yksinkertaistuu muotoon

$$P(X_1,...,X_n,C) = P(C)\prod_{i=1}^{n} P(X_i \mid Pa(X_i)) = P(C)\prod_{i=1}^{n} P(X_i \mid C) . \qquad (73)$$



Kuva 7 naiivin Bayesin luokittelijan puu-rakenne

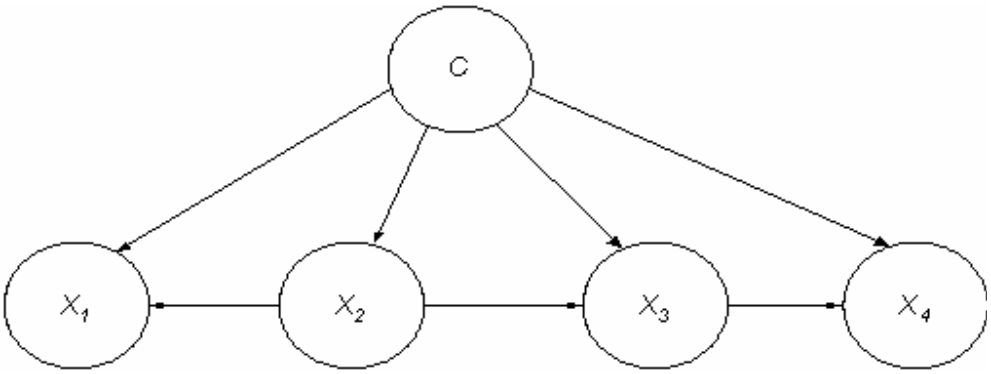
Naiivin Bayes-verkon yksinkertaisuus on yksi sen vahvuuksista. Kuten muissa luokittelijoissa naiivin luokittelijan rakenteen määräämisessä ei tarvitse käyttää hakualgoritmeja. Riittää, että asetetaan luokittelijamuuttujan kaikkien muiden muuttujien vanhemmiksi, jonka jälkeen opitaan parametrit datasta.

Naiivi Bayesin luokittelija on erittäin käytännöllinen Bayesin oppimismenetelmä. Tämä luokittelija on joissain tapauksissa suorituskyvyltään parempi kuin neuroverkot ja päätöspuut [4]. Monissa tapauksissa kuitenkin naiivin Bayesin luokittelijan *luokkaehdollinen riippumattomuusoletus* on selkeä rajoite. Toisin sanoen on tilanteita, joissa tämä oletus ei ole voimassa. [2]

4.4.2 TAN

Naiivin Bayesin luokittelijan luokkaehdollinen riippumattomuusoletus kaikkien muuttujien suhteen on oletus, joka on harvoin voimassa. *TAN* (*Tree Augmented Naive-Bayes*) -menetelmä tuottaa verkkorakenteen, jossa jokaiselle attribuutille sallitaan toinen vanhempi juurivanhemman lisäksi. Kuvassa 8 on yksinkertainen *TAN* -rakenne, jossa kaikilla muuttujilla on vähintään luokittelumuuttuja vanhempanaan. [10]

TAN –rakenteen, toisin sanoen ylimääräisten kaarien oppiminen datasta pohjautuu tunnettuun Chown ja Liun vuonna 1968 esittämään menetelmään puutyyppisten Bayes-verkkojen oppimiseen. Algoritmi *TAN* -rakenteen oppimiseen käyttää *luokkaehdollista keskinäistä informaatio -testiä* testatessaan solmujen välisiä riippuvuuksia.



Kuva 8 Esimerkki TAN-rakenteesta

Määritelmä 2 (entoropia) Diskreetin satunnaismuuttujan X entropia $H(X)$ määritetään

$$H(X) = -\sum_{x} p(x)\log p(x)\,. \tag{74}$$

Satunnaismuuttujan X entropia on mitta, joka kertoo satunnaismuuttujan satunnaisuudesta. Mitä enemmän muuttujassa on satunnaisuutta, sitä suurempi otos tarvitaan, jotta datasta saataisiin luotettava kuva.

Määritelmä 3 (keskinäinen informaatio) Olkoon X ja Y diskreettejä satunnaismuuttujia. Näiden välinen informaatio $I(X,Y)$ määritellään

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,, \end{aligned} \tag{75}$$

missä kahden muuttujan välinen ehdollinen entropia,

$$H(Y \mid X) = -\sum_{x,y} p(x,y) \log p(y \mid x), \tag{76}$$

on muuttujan X satunnaisuuden mitta kun muuttuja Y on annettu. Toisin sanoen tämä mitta kertoo sen, kuinka paljon X:n satunnaisuudesta on jäljellä, Y:n havaitsemisen jälkeen. Mitta $I(X,Y)$ on informaation määrä, jonka muuttuja Y muuttujasta X paljastaa.

Määritelmä 4 (ehdollinen informaatio) Olkoon X, Y ja Z diskreettejä satunnaismuuttujia. Muuttujien X ja Y välinen informaatio, kun Z on annettu, määritetään

$$\begin{aligned} I(X,Y \mid Z) &= H(X \mid Z) - H(X \mid Y,Z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \end{aligned} . \tag{77}$$

Mitta $I(X,Y|Z)$ on informaation määrä, jonka muuttuja Y muuttujasta X tuottaa, kun Z on havaittu.

TAN –oppimisproseduuri on seuraavanlainen:

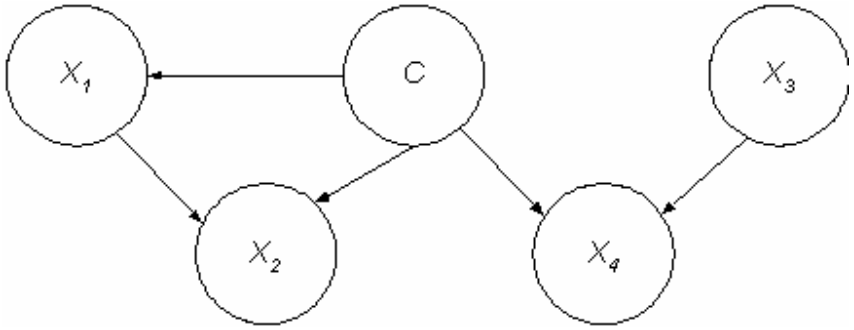
- Ota opetusjoukko ja $X\backslash\{C\}$ syötteeksi, missä C on luokittelusolmu.
- Kutsu modifoitua Chow-Liu algortimia
- Aseta solmu C vanhemmaksi jokaiselle attribuutille x_i, $1 \leq i \leq n$
- Opi parametrit ja tulosta *TAN*

Modifoitu Chow-Liu-algoritmi on seuraavanlainen [10]:

- Laske $I(A_i;A_j|C)$ jokaiselle attribuuttiparille, $i \neq j$
- Rakenna täysi suuntaamaton graafi, jossa solmut ovat attribuutit $A_1,\ldots,A_n$. Määrää kaarelle (A_i,A_j) paino ehdollisen informaation $I(A_i, A_j|C)$ mukaan.
- Rakenna maksimaalisesti virittävä puu
- Muuta saatu suuntaamaton puu suunnatuksi valitsemalla juurisolmu ja asettamalla kaarien suunnat poispäin tästä juurisolmusta
- Rakenna *TAN*-malli lisäämällä solmu C puuhun ja lisäämällä kaaret tästä solmusta jokaiseen solmuun A_i

4.4.3 Yleinen Bayes -verkko luokittelussa

Edellä esitetyt luokittelijat asettivat luokittelusolmun *C* kaikkien muiden solmujen vanhemmaksi. *GBN* (*general bayesian network*) käsittelee solmua *C* kuten mitä tahansa solmua, toisin sanoen se ei aseta välttämättä solmua *C* muiden vanhemmaksi. Kuvassa 9 on vielä kertaalleen esitetty yleisen Bayes -verkon rakenne. Kuvasta näkee, että solmu *C* on kuten mikä solmu tahansa.



Kuva 9 Esimerkki yleisestä Bayes-verkosta

Yleisen Bayes -verkon rakenteen oppimisessa voidaan käyttää pistemääräperustaista lähestymistapaa, kuten luvussa 4.2.1 Pistemääräfunktioihin perustuvat menetelmät tai lähestymistapa on esitelty. Toinen mahdollinen lähestymistapa on rajoiteperustainen lähestymistapa. Tästä lähestymistavasta on mainittu lyhyesti luvussa 4.2.2. Tässä työssä käytetään pistemääräperustaista lähestymistapaa Bayes -verkon rakenteen oppimisessa.

Naiivissa luokittelijassa luokittelumuuttuja C on muiden solmujen vanhempi, eikä muita solmuja sallita. Johtuen tästä ja Markov -oletuksesta yhteistodennäköisyys on yksikertaisuudessaan kuten kaavassa (73) on esitetty. Nyt kun verkko on kuvan 9 verkko, niin yhteistodennäköisyys on muotoa

$$P(X_1, X_2, X_3, X_4, C) = P(C)P(X_1 \mid C)P(X_2 \mid C, X_1)(X_4 \mid C, X_3)P(X_3). \quad (78)$$

Havainto luokiteltaisiin kuuluvan siihen luokkaan, joka maksimoi kaavassa (78) olevan posteriorijakauman. Jos kuvan 9 verkossa olisi solmulla X_3 vanhempaan esimerkiksi solmu X_5, niin tätä solmua ei käytettäisi luokittelutehtävään. Tämä johtuu siitä, että solmu X_5 ei olisi luokittelumuuttujan *C* Markov -peitossa [10].

5 Potilaiden luokittelu tautiryhmiin Bayes -verkkojen avulla

5.1 *Korvalääketieteellinen aineisto*

Otoneurologinen aineisto on kerätty Helsingin yliopistollisen keskussairaalan korvaklinikalla. Otoneurologiset asiantuntijat ovat todenneet, että tutkituista potilaista 815:lla on jokin kuudesta yleisestä otologisesta sairaudesta, johon liittyy vertigo eli huimaus. Nämä sairaudet ovat akustikus neurinoma, bening positional vertigo (hyvänlaatuinen asentohuimaus), Menièren tauti, äkillinen kuulon menetys, traumaattinen vertigo ja vestibular neuritis. Yleisesti, jos potilaalla sanotaan olevan sairaus, johon liittyy vertigo, voidaan sanoa, että potilas on niin sanottu huimauspotilas (engl. *vertigo; dizziness*). [21, 22]

Menièren taudin määritelmä vaihtelee maittain. Kentalan mukaan Menièren taudin oireita voivat olla vertigo, tinnitus ja vaihteleva (*fluctuating*) kuulon menetys tai mikä tahansa näiden yhdistelmä [22]. Akustikus neurinoman (*vestibular schawannoma*) pääoireiksi määritetään yllättävä kuuroutuminen tai etenevä kuulon huononeminen ja jatkuva yksipuolinen tinnitus. Muita mahdollisia akustikus neurinoman oireita ovat tasapainohäiriöt, päänsärky, kasvoalueen tunnottomuus, näköhäiriöt ja korvakipu. Bening positional vertigo on yleinen vertigon muotoa vanhemmilla ihmisillä. Oireina ovat jaksottaiset huimauskohtaukset. Korkean iän, passiivisuuden ja muiden korvasairauksien katsotaan edesauttavan taudin kehittymistä. Vestibular neuritis on oireyhtymä, johon liittyy äkillinen sisäkorvan elinten toimimattomuus ilman jatkuvaa kuulonmenetystä. Traumaattinen vertigo on ryhmänimeke labyrintin reunaosien tai keski-vestibulaariselle oireyhtymälle, jossa oireet alkavat välittömästi pään vammautumisen jälkeen. Oireet johtuvat labyrintin vauriosta tai vestibulaarisen (tasapaino) hermon vauriosta tai keski-verstibulaarisen rakenteen vauriosta. Äkillisen kuuroutumisen diagnosointi on vaikeaa johtuen tämän taudin vaihtelevasta kliinisestä kuvasta. Äkillisen kuuroutumisen esiintyminen vaihtelee eri ikäryhmissä siten, että korkein esiintymisprosentti on 50-60 -vuotiailla henkilöillä. Miehet ja naiset sairastuvat yhtä usein äkilliseen kuurouteen. Vaskulaarinen etiologia johtaa kaikkein todennäköisimmin äkilliseen kuurouteen. Muita

äkillisen kuurouden oireisiin johtavia tiloja ovat esim. kasvain, metaboliset oireyhtymät, vamma ja sisäkorvan kalvon rikkoutuminen. [22]

Taulukko 2 Muuttujan diagnoosi frekvenssit

Diagnoosi	Frekvenssi	Prosenttiosuus
akustikus neurinoma	130	15,95
bppv	146	17,91
meniere	313	38,40
sudden deafness	41	5,03
traumatic vertigo	65	7,98
vestibular neuritis	120	14,72
Yhteensä	815	100

Taulukossa 2 on 815 potilaan frekvenssit muuttujasta diagnoosi. Nähdään, että suurin potilasryhmä (38,4 %) on Menièren taudista kärsivä ryhmä. Vajaalla 18 %:lla potilaista on Beningin positionaalinen vertigo. Akustikus neurinoma ja vestibular neuritis on noin 15 %:lla potilaista. Traumaattisesta vertigosta kärsivien osuus on noin kahdeksan prosenttia ja äkillisestä kuuroudesta kärsivien osuus on viisi prosenttia. Otoneurologinen aineisto käsitti alun perin 170 eri muuttujaa; Potilaat täyttivät kysymyslomakkeen, jossa kysyttiin oireista, aikaisemmista sairauksista, mahdollisista tapaturmista ja päihteiden käytöstä [22]. Potilasjoukolle suoritettiin myös erilaisia kliinisiä mittauksia, joiden katsottiin olevan tarpeellisia diagnostisessa mielessä.

Aikaisemmissa tutkimuksissa on löydetty 38 muuttujan joukko, jonka katsotaan olevan tärkeä luokittelutehtävän kannalta [22, 28]. Taulukossa 3 ovat nämä 38 muuttujaa. Liitteessä 1 on lueteltuna näiden muuttujien arvot, sekä muuttujien lyhennetyt nimet. Luvussa 5, käytän näitä lyhenteitä viitatessani muuttujiin. Yhdeksän muuttujaa on määritelty tärkeiksi diagnoosin tekemisen kannalta (Taulukossa 3 vahvistetulla fontilla) [24]. Aineiston keränneet otoneurologiset asiantuntijat ovat määrittäneet viisi avainmuuttujaa, joita ilman diagnoosia ei voida tehdä [25]. Nämä viisi avainmuuttujaa (merkitty *:llä Taulukossa 3) ovat *huimausoireiden kesto*, *huimauskohtausten esiintyvyysfrekvenssi, huimauskohtauksen kesto, kuulo-oireiden kesto* ja *päävamma*. Muuttuja *type of hearing loss* on jälkikäteen koodattu kolmeksi dummy -muuttujaksi; kysymyksiin onko kuulon heikkeneminen ollut *äkkinäinen, hiljalleen kehittyvää vai molempia* vastaukset ovat kyllä tai ei. Näin ollen todellinen tässä työssä käsiteltävien selittävien muuttujien määrä nousi 40:een.

Taulukko 3 Muuttujaluettelo

Age at first symptoms	Ear ilness
Duration of vertigo symptoms*	Ear operation
Duration of vertigo attacks*	**Head injury***
Frequency of vertigo attacks*	Ear trauma
Intensity of vertigo attacks	Noise injury
Score for rotational vertigo	Chronic noise exposure
Score for floating vertigo	Spontaneus nystagmus
Score for position-induced vertigo	Spontaneus nystagmus degree /second
Unsteadiness	Caloric asymmetry
Sudden falls	Caloric asymmetry with response 44C left
Duration of hearing loss*	Caloric asymmetry with response 44C right
Type of hearing loss	Posturography eyes open
Location of hearing loss-Right ear	Posturography eyes closed
Location of hearing loss-Left ear	Pursuit eye movements Gain Amplitude
Fluctuation in hearing loss	Pursuit eye movements Gain Latency
Intensity of tinnitus	Tone burst audioemtry 500 Hz right
Duration of tinnitus	Tone burst audioemtry 500 Hz left
Severity of Nausea	Tone burst audioemtry 2000 Hz right
Light-headedness	Tone burst audioemtry 2000 Hz left

5.2 Puuttuvan tiedon käsittely

Keskimäärin puuttuvaa tietoa näiden 40 muuttujan osalta on noin 11 %. Muuttujajoukossa, joka käsittää erilaiset testitulokset, on eniten puuttuvaa tietoa. Tämä johtuu muun muassa siitä, että jokaiselle potilaalle on ollut turha tehdä kaikkia testejä; toiselle potilaalle on pystytty diagnoosi varmistamaan ilman jotain tiettyä testiä, kun taas toiselle potilaalle on jouduttu tekemään tämä testi diagnoosin varmistamiseksi. Yllättäen myös joidenkin kysymyslomakkeen kysymysten vastauksissa on puuttuvaa tietoa. Kysymyksiin *type of hearing loss, fluctuating in hearing, spontanic nystagmus* ja *ear operation* on potilaiden ollut hankala vastata, koska nämä kysymykset vastausvaihtoehtoineen on koettu epäselviksi [23]. Osa muuttujista oli alun perin jatkuvia. Tässä työssä käsitellään diskreettejä verkkoja, ja siksi jatkuvat muuttujat on diskretoitu.

Tämän työn kontekstissa käsitellään verkkoja, joissa puuttuvaa tietoa ei sallita. Puuttuvan tiedon käsittely on oma laaja alueensa, johon tässä työssä ei tarkemmin puututa. Esittelen seuraavaksi yleisellä tasolla puuttuvan tiedon korvaamiseen liittyviä asioita.

Se miten puuttuvaa tietoa pitäisi käsitellä, riippuu mekanismista, joka on johtanut puuttuvan tiedon syntymiseen [29]. Olkoon $\mathbf{X}=\{x_{ij}\}$ $(N\times V)$ –datamatriisi siten, että x_{ij} on tilastoyksikön i saama arvo muuttujalle j; $i=1,\ldots, N$; $j=1,\ldots, V$. Olkoon $\mathbf{M}=\{m_{ij}\}$ puuttuvan tiedon $(N\times V)$ –indikaattorimatriisi, siten että $m_{ij}=1,$ jos x_{ij} puuttuu ja $m_{ij}=0,$ jos x_{ij} ei puutu. Matriisi **M** kuvaa siis puuttuvan tiedon takana olevaa mallia. Olkoon nyt $P(\mathbf{M}|\mathbf{X}, \Phi)$ **M**:n ehdollinen todennäköisyys, kun on annettu **X** ja parametrit Φ, jotka karakterisoivat vastausprosentteja. Puuttuvan tiedon takana oleva mekanismi on *MCAR* (*Missing Completely At Random*), jos

$$P(\mathbf{M}|\mathbf{X}, \Phi) = P(\mathbf{M}|\Phi) \ \forall \ \mathbf{X}. \text{ [29,30,31]} \qquad (79)$$

Jos puuttuvien havaintojen voidaan katsoa olevan yksinkertainen satunnaisotos kaikista havaintoarvoista, niin puuttuvan tiedon takana oleva mekanismi on *MCAR*. Esimerkkinä kysely, jossa kysytään mm. henkilön vuosituloja. Jos voidaan olettaa, että henkilöillä, jotka eivät ilmoita tulojaan, on keskimäärin yhtä suuret tulos kuin henkilöillä, jotka ovat vuositulonsa ilmoittaneet, niin tällöin puuttuvan tiedon takana on *MCAR*- prosessi.

Merkitään nyt matriisin **X** havaittua osaa matriisilla $\mathbf{X}_{obs}$ ja puuttuvaa osaa matriisilla $\mathbf{X}_{mis}$ siten, että $\mathbf{X}=(\mathbf{X}_{obs}, \mathbf{X}_{mis})$. Puuttuvan datan takan prosessi on *MAR* jos

$$P(\mathbf{M}|, \mathbf{X}_{obs}, \mathbf{X}_{mis}\ \Phi) = P(\mathbf{M}\ |\mathbf{X}_{obs}, \Phi) \ \forall \ \mathbf{X}_{mis}. \text{ [29,30,31]} \qquad (80)$$

Vähemmän formaalimmin ilmaistuna *MAR* -prosessi tarkoittaa seuraavaa: Todennäköisyys, että havainto puuttuu saattaa johtua matriisista $\mathbf{X}_{obs}$ mutta ei matriisista $\mathbf{X}_{mis}$. *MAR* –prosessistaa esimerkkinä kysely, jossa kysytään vastaajilta ikää ja vuosituloja. Jos ikä-muuttuja on täysin havaittu ja vuositulo-muuttujassa on puuttuvaa tietoa, ja jos tiedon puuttuminen johtuu iästä, niin tällöin kysymyksessä on *MAR* –prosessi.

Kuten kappaleen alussa on mainittu, niin se miten puuttuvaa tietoa pitäisi käsitellä, riippuu siitä minkä prosessin katsotaan olevan puuttuvan tiedon takana. Yksi tapa lähestyä asiaa on se, että tilastollisissa tarkasteluissa ei oteta huomioon tapauksia, joissa on puuttuvaa tietoa. Tämän menetelmän etuna on se, että se on helppo toteuttaa ja voi johtaa tyydyttävään tulokseen jos puuttuvaa tietoa on vähän. Yleisesti kuitenkaan tämän menetelmän käyttöä ei suositella, koska se voi johtaa harhaisuuteen [29]. Yleisesti

käytetty menetelmä on käyttää imputointia puuttuvan tiedon korvaamiseen. Niin sanottu *ad hoc* -imputointi tarkoittaa sitä, että puuttuva tieto korvataan keskiluvullaan. Toisin sanoen, jos muuttuja on jatkuva, niin puuttuva tieto korvataan keskiarvollaan, ja jos muuttuja on järjestysasteikollinen, niin se korvataan mediaanillaan, ja jos muuttuja on nominaaliasteikollinen, niin se korvataan moodillaan. Termiin *ad hoc* (Lat. *tapauskohtainen, yhteen tarkoitukseen soveltuva*) liittyy yleensä ajatus, että asia tai menetelmä, josta käytetään ilmaisua *ad hoc*, ei ole yleiskäyttöinen. Tämä konteksti ei ole poikkeus. Puuttuvan tiedon korvaaminen keskiluvullaan on järkevää vain silloin kun puuttuvan tiedon takana oleva prosessi on *MCAR*. Kun tarkastellaan aineiston puuttuvaa tietoa, ensisijaisena toiveena yleensä on, että *MCAR* -oletus olisi voimassa. Näin ei yleensä kuitenkaan ole; *MCAR* -oletus on valitettavasti vain harvoin voimassa. Puuttuvan tiedon korvaamista keskiluvuillaan käytetään kuitenkin usein, mikä johtuu varmaan suurimmaksi osaksi tämän imputoinnin helppoudesta. Muita tapoja imputoida puuttuva tieto on ns. regressio-imputointi, joka tarkoittaa, että muuttujan puuttuvat arvot estimoidaan käyttäen hyväksi muuttujan havaittuja arvoja [29]. Myös regressio-imputoinnissa on oletuksena *MCAR* -mekanismi. Kun puuttuvan tiedon takana on *MAR* -prosessi, niin järkevä tapa hoitaa puuttuvan tiedon korvaaminen on käyttää niin sanottua malli-perustaista menetelmää nimeltä *EM* -algoritmi (*Expectation-Maximation Algorithm*). *EM* -algoritmi on kaksivaiheinen algoritmi mallin parametrien estimoimiseen. Perusalgoritmi sisältää kaksi askelta: *E*-askeleen (*Expectation*) ja *M*-askeleen (*Maximation*). Aluksi data jaetaan kahteen osaan; ensimmäisessä osassa on data, jossa ei ole puuttuvaa tietoa ja toisessa osassa data, jossa on puuttuvaa tietoa. Aluksi myös alustetaan parametrien lähtöarvot. *E*-askeleessa käyttäen parametreja, lasketaan ennustetut arvot puuttuvalle datalle, jonka jälkeen *M*-askeleessa, käyttäen saatuja ennusteita, maksimoidaan uskottavuusfunktiota tavoitteena saavuttaa uusia parametreja. Prosessia toistetaan, kunnes saavutetaan suppeneminen. *EM*-menetelmän käytön huonoin puoli on se, että sitä varten ei ole olemassa helposti käytettävissä olevaa valmista ohjelmaa; puuttuvan tiedon kanssa pähkäilevän keskiverron tilastotieteilijän, jolla ei ole koodaustaitoja, on käytännössä vaikea toteuttaa *EM* -algoritmia aineistollaan.

Tämän työn otoneurologisen datan puuttuvan tiedon takana katsotaan olevan *MAR*-prosessi [23]. Laurikkala et al. ovat tutkineet otoneurologisella aineistolla (*N=564*)

puuttuvan tiedon käsittelyn merkitystä luokittelutarkkuuksiin, kun luokittelu on suoritettu käyttäen diskriminanttianalyysia [23]. Puuttuva tieto korvattiin potilasryhmittäin käyttäen *ad hoc* -menetelmää, regressio-menetelmää ja *EM*-menetelmää. Järkevintä olisi olettaa, että *EM*-menetelmällä käsitelty data on kaikkein luotettavinta johtuen datan *MAR* -oletuksesta. Luokittelutarkkuudet ovat kuitenkin todella lähellä toisiaan näillä kolmella imputointi-menetelmällä. Keskimääräiset luokittelutarkkuudet ovat 96, 95 ja 94 prosenttia *ad hoc* -menetelmällä, regressio-menetelmällä ja *EM*-menetelmällä. Tässä työssä puuttuva tieto on korvattu käyttäen *ad hoc*-menetelmää; puuttuva tieto on korvattu käyttäen tautiryhmien keskilukuja. Syy tämän imputointi-menetelmän valintaan on sen helppous ja edellä mainitun tutkimuksen tulokset.

5.3 Weka 3

jBNC on Java-työkalu Bayes-verkkoihin pohjautuvien luokittelijoiden oppimiseen, testaamiseen ja soveltamiseen [33]. Weka 3 on tiedonlouhintaohjelma, jonka toiminnallisuuteen kuuluvat muun muassa datan esikäsittely, luokittelu, regressio, klusterointi, assosiaatiosäännöt ja datan visualisointi. Ohjelma on kehitetty Waikaton yliopiston tietojenkäsittelyn laitoksella (http://www.cs.waikato.ac.nz) koneoppimisen projektin yhteydessä. jBNC-Weka on välikappale, joka mahdollistaa jBNC:n toimintojen käyttämisen Weka 3:n käyttöliittymän kautta. Wekassa on toteutettuna naiivi luokittelija ja GBN. Lisäämällä jBNC käyttöön saadaan myös muita Bayes -verkkoon pohjautuvia luokittelijoita kuten TAN ja BAN. Yleisen Bayes -verkon rakenne voidaan löytää käyttäen joko pistemääräperustaista menetelmää tai rajoiteperustaista menetelmää. Käytännössä riippumattomuustesteihin pohjautuvaa rajoiteperustaista menetelmää on hankalampi toteuttaa kun muuttujia on paljon; suurella muuttujajoukolla muistiongelmat ovat väistämättömiä.

5.4 Käytetyt luokittelijat, testiasetelma ja luokittelutarkkuus

Tässä työssä sovelletaan naiivia luokittelijaa, *TAN* -luokittelijaa ja *GBN* -luokittelijaa. *GBN* -luokittelijan verkkorakenteen löytämiseen käytetään pistemääräperustaista menetelmää, jossa hakualgoritmina on vuorikiipeilyalgoritmi. Käytetyt

pistemääräfunktiot ovat *MDL* -mitta, Bayesin pistemäärää ja *AIC*. Käytetyt muuttujajoukot on mainitut viiden, yhdeksän ja 40 muuttujan joukot (ks. Luku 5.1.1).

Weka 3 -ohjelmassa verkon hakualgoritmin valinnan yhteydessä on valittavana, että onko aloitusverkko naiivi-verkko vai satunnaisesti tuotettu verkko. Lisäksi on valittavissa niin sanottu Markov -peittokorjaus (*Markov blanket correction*). Jos tämän valinnan tekee, niin tällöin verkon kaikki solmut ovat luokittelusolmun Markov -peitossa. Hakualgoritmin valinnan yhteydessä voi määrätä solmujen vanhempien maksimilukumäärän. Kun tämän luvun asettaa suureksi, paljon suuremmaksi kuin muuttujajoukko on, esimerkiksi 100 000, tulee varmistettua, että saatu verkko on varmasti rajoittamaton. Lisäksi valittavissa on, että sallitaanko kaarien kääntäminen (*arc reversal*). Vuorikiipeily-hakualgoritmin mahdollisia operaatioitahan ovat kaaren lisäys, poisto ja kääntö. Pakko ei kuitenkaan ole käyttää kääntö-operaatiota. Hakiessani parhaita mahdollisia luokittelijoita, kokeilin, että kummalla tavalla tulee parempia luokittelijoita – kaaren käännön kanssa vai ilman. Lopuksi oli valittavana haluttu estimaattori, jonka avulla saa määritettyä luokkaehdolliset todennäköisyystaulut (priorijakaumat) rakenteen oppimisen jälkeen. Wekassa on tällä hetkellä rajoitettu valikoima estimaattoreita ja tästä syystä tässä työssä käytetään yksinkertaista estimaattoria (Wekassa SimpleEstimator). Olen työssäni esittänyt, kuinka priorijakaumien estimointiin on käytettävissä *ML*-estimaattori, niin sanottu *M*-estimaattori, Laplace-estimaattori, Bayes-estimaattori jne. Wekan käyttämä yksinkertainen estimaatti luokkaehdollisista todennäköisyyksistä on muotoa [32]:

$$P(x_i = k \mid Pa(x_i) = j) = \frac{m_{ijk} + m_{ijk}'}{\sum_k m_{ijk} + m_{ij}'} ,$$

mikä on itse asiassa m-estimaatti. Oletuksena Wekassa on, että m_{ijk}' on 0,5. Arvon 0,5 voidaan tulkita tarkoittavan, että muuttujien oletetaan olevan binäärisiä, jolloin siis estimaatti olisikin Laplace-estimaatti. Itse luokittelijoita rakentaessani tulin siihen tulokseen, että parhaita luokittelijoita syntyi, kun $m_{ijk}' = 1/6$. Jos arvoksi laittaisi nollan, niin kysymyksessähän olisi *ML* -estimaatti.

Verkon oppimiseen käytin ristiinvalidointi-menetelmää kymmenen jaolla (*cross-validation with 10-folds*). Tässä menetelmässä data jaetaan satunnaisesti kymmeneen osajoukkoon, joista jokainen osajoukko on vuoronperään testijoukko ja loppu 90% datasta on opetusjoukko. Opetusjoukon tehtävänä on siis oppia luokittelija, jonka kyvykkyys sitten testataan testijoukon avulla. Lopullinen tulos on agregaatti näiden kymmenen kokeen tuloksista.

Lisäksi olen myös oppinut verkkorakenteen niin, että käytin koko dataa, eli 815 havaintoa opetusjoukkoja ja myös samaa joukkoa testijoukkoa. Yksinään tätä menetelmää ei saisi käyttää, koska yleensä on niin, että opetusjoukon ollessa testijoukko verkon hyvyyden arvioinnissa käytetyt tunnusluvut ovat paremmat kuin jos testijoukkona käytetään ennen näkemätöntä, toisin sanoen verkon opetuksessa käyttämätöntä dataa. Tuloksia, jotka on saatu menetelmällä jossa opetusjoukko on sama kuin testijoukko, voidaan käyttää esimerkiksi arviointiin, että onko testijoukot edustava otos koko datasta [28].

Määritetään seuraavaksi luokittelijan hyvyyden määrittämisessä tarvittavat suureet. Oikealla positiiviselle tarkoitetaan sitä, että potilas on luokiteltu siihen tautiryhmään johon hän oikeasti kuuluu. Väärällä positiivisella vastaavasti tarkoitetaan sitä, että potilas luokitellaan kuuluvaksi väärään tautiryhmään. Merkitään oikeiden positiivisten lukumäärää suureella *tp* ja väärien positiivisten lukumäärää suureella *fp*. Oikeiden negatiivisten lukumäärää merkitään suurella *tn* ja väärien negatiivisten lukumäärä merkitään suureella *fn*. Tunnistamistarkkuus *r* (*recognition accuracy*), ennustamistarkkuus *p* (*prediction accuracy*) ja kokonaistarkkuus *t* (*total accuracy*) määritetään seuraavasti:

$$r = \frac{tp}{tp + fn} 100\%, \tag{81}$$

$$p = \frac{tp}{tp + fp} 100\%, \tag{82}$$

$$t = \frac{tp + tn}{tp + tn + fp + fn} 100\%. \tag{83}$$

Havainnollisuuden vuoksi esittelen ristiintaulukon eräästä kokeiluajosta, josta esitellyt suureet saa helposti poimittua. Taulukon 4 vasemmasta sarakkeesta nähdään todellinen

tautiryhmä. Luvut viereisissä sarakkeissa kertovat, miten luokittelija (Naiivi luokittelija, 40 muuttujaa) on potilaat luokitellut. Nyt esimerkiksi nähdään, että 34 potilasta on luokiteltu oikein tautiryhmään *Akustikus neurinoma*, toisin sanoen tämän taudin osalta *tp=34*. Nähdään myös, että potilaista, jotka kuuluvat oikeasti tautiryhmään *Akustikus neurinoma*, kolme on luokiteltu tautiryhmään *Bppv* (Beningin positionaalinen vertigo), viisi tautiryhmään *Menière* ja yksi tautiryhmään *Vestibular Neuritis*, toisin sanoen *fn=9*. Väärien positiivisten lukumäärä on *fp=0* ja oikeiden negatiivisten lukumäärä *tn=209*. Nyt taudin Akustikus Neurinoman osalta tarkkuudet ovat

$$r = \frac{34}{34+9}100\% = 79{,}1\%$$

$$p = \frac{34}{34+0}100\% = 100\%$$

$$t = \frac{34+209}{34+209+0+9}100\% = 96{,}4\%$$

Taulukko 4 Esimerkki luokittelufrekvensseistä

	Akustikus Neurinoma	Bppv	Meniere	Sudden Deafness	Traumatic Vertigo	Vestibular Neuritis
Akustikus neurinoma	34	3	5	0	0	1
Bppv	0	44	3	0	0	1
Meniere	0	3	102	1	1	0
Sudden Deafness	0	0	0	13	0	0
Traumatic Vertigo	0	0	4	0	20	1
Vestibular Neuritis	0	0	3	0	0	39

5.5 *Tulokset*

5.5.1 Kymmenen luokittelijaa

Kuten edellä on kerrottu, niin tarkoituksena oli luoda luokittelevia verkkoja, kun muuttujien lukumäärä on 40, 9 ja 5 ja kun verkkotyypit ovat Naiivi, *TAN* ja *GBN*. Lisäksi *GBN* -verkon yhteydessä käytin kolmea eri pistemääräfunktiota (Bayes, *MDL* ja *AIC*). Kaiken kaikkiaan tuloksena on kymmenen eri verkkoa. Testiasettelussa käytin ristiinvalidointimenetelmää.

5.5.2 Muuttujia 40

Muuttujajoukon ollessa 40 päästiin parhaisiin tarkkuuksiin kaikilla luokittelijoilla. Taulukossa 5 on tunnistamistarkkuudet, ennustamistarkkuudet ja kokonaistarkkuudet luokittelijoilla Naiivi, *TAN*, GBN_1 (pistemäärä Bayes), GBN_2 (pistemäärä *MDL*) ja GBN_3 (pistemäärä *AIC*). Nähdään, että *MDL* -pistemääräfunktiolla luotu luokittelija on verrattuna muihin luotuihin verkkoihin paras. Yleisesti ottaen ennustamistarkkuudet ja kokonaistarkkuudet ovat hyvät, ja tunnistamistarkkuuksissakin päästään 40 muuttujalla kohtalaisiin tuloksiin. Huonoin tunnistamistarkkuus on tautityypillä Akustikus neurinoma.

Taulukko 5 Tarkkuusluvut - 40 muuttujaa

40 muuttujaa		NAIIVI	TAN	GBN_1	GBN_2	GBN_3
Akustikus neurinoma	**r**	0,854	**0,885**	0,854	0,854	0,862
	p	0,991	0,983	0,957	**1**	0,974
	t	0,976	0,98	0,971	**0,977**	0,975
Bppv	**r**	0,863	0,856	**0,877**	0,87	0,849
	p	0,9	0,893	0,895	**0,901**	0,879
	t	0,959	0,957	**0,96**	**0,96**	0,953
Menière	**r**	0,952	0,949	0,949	**0,962**	0,936
	p	0,876	0,887	0,889	**0,896**	0,867
	t	0,931	0,935	0,936	**0,943**	0,922
Sudden Deaffness	**r**	0,902	**0,927**	**0,927**	**0,927**	**0,927**
	p	0,949	0,905	0,927	**0,95**	**0,95**
	t	0,993	0,991	0,993	**0,994**	**0,994**
Traumatic Vertigo	**r**	0,908	0,831	0,877	**0,938**	0,862
	p	0,881	0,857	**0,919**	0,871	0,875
	t	0,983	0,976	**0,984**	**0,984**	0,979
Vestibular Neuritis	**r**	0,9	**0,917**	0,908	**0,917**	0,908
	p	0,923	0,932	0,916	**0,94**	0,932
	t	0,975	0,978	0,975	**0,979**	0,977

Tautityyppien Akustikus neurinoman ja Bppv:n tunnistamistarkkuudet ovat selkeästi huonommat kuin muilla, joten tutkin asiaa vähän tarkemmin. Taulukosta 6 selviää mihin tautityyppeihin havainnot on luokiteltu luokittelijalla GBN_2. Taulukosta käy ilmi, että yhteensä 111 havaintoa on luokiteltu tautiryhmään Akustikus neurinoma kuuluvaksi. Kaksi näistä havainnosta ei todellisuudessa kuulu tähän tautiryhmään. Lisäksi on nähtävissä, että 19 havaintoa, jotka oikeasti kuuluvat tähän tautiryhmään, on luokiteltu kuuluvaksi muihin tautiryhmiin. Näistä 19: sta havainnosta on itse asiassa luokiteltu 11 kuuluvaksi tautiryhmään Menièren tauti. Myös muilla luokittelijoilla on samankaltainen tulos. Luokittelijalla GBN_1 on luokiteltu 12 potilasta tautiryhmään Menièren tauti, kun he oikeasti kuuluvat tautiryhmään Akustikus neurinoma. Vastaavat luvut luokittelijoilla *TAN* ja GBN_3 ovat 10 ja 16. Naiivilla luokittelijalla on 15 luokiteltu kuuluvaksi virheellisesti tautiryhmään Menièren tauti, tapauksista jotka oikeasti kuuluvat tautiryhmään Akustikus neurinoma. Artikkelissa Discovering Diagnostic Rules from a Neurotologic Database with Genetic Algorithms (Kentala et al.) esitetään, että taudin Akustikus Neurinoman osalta tärkeitä muuttujia ovat *duration of vertigo attacks*, *duration of hearing loss*, *occurance of head injury*, *intensity of vertigo attack*, *score for positional vertigo*, *score for ro-*

tational vertigo, *score for floating vertigo*, *intensity of nausea* ja *presence of sudden falls*. Ensiksi ajattelin, että jos onkin niin, että osa tai jokin näistä muuttujista ei kuulukaan luokittelumuuttujan Markov -peittoon. Jos näin olisi, niin nämä muutujat eivät osallistuisi luokittelutehtävään. Tästä ei ole kuitenkaan kysymys, koska verkkoja rakennettaessa Markov -korjauksella varmistin sen, että kaikki muuttujat kuuluvat luokittelumuuttujan Markov -peittoon. Itse asiassa ilman Markov -korjausta vain muuttuja *HL_TYPE3 (Hearing loss both sudden and progressive)* puuttui luokittelumuuttujan Markov -peitosta. Kun muuttuja otettiin mukaan, niin ehdollinen todennäköisyys *P(HL_TYPE3=0|REMID_MD)*, riippumatta tautiryhmästä, on lähellä ykköstä. Tämän muuttujan vaikutuksen voi tulkita kaikille samanlaiseksi, toisin sanoen se olisi luokittelutehtävässä turha muuttuja. Menièren taudin osalta kaikkein tärkeimmäksi muuttujaksi artikkelissa on esitetty *severity of tinnitus*. Muita tärkeitä muuttujia tämän taudin luokittelun kannalta katsottiin olevan *duration of vertigo*, *occurance of head or ear injury, response with 44 C° in electonystamography* ja *audiometry founding at 500 Hz*. Kävin läpi verkkoja ja niiden ehdollisia todennäköisyyksiä tavoitteena selvittää, että löytäisinkö jotain merkittävyyttä näiden tautien luokittelun kannalta oleellisiksi esitettyjen muuttujien osalta. Esimerkiksi, että ovatko ehdolliset todennäköisyydet samat tautiryhmillä Akustikus Neurinoma ja Menièren tauti näillä merkitsevillä muuttujilla? Tällä tarkastelutavalla en päässyt puusta pitkään, ja varsinkin kun luokittelussa käytetään kaikkia luokittelumuuttujan Markov -peitossa olevia muuttujia, niin on hiukan vaikeaa lähteä näin maallikkona arvailemaan, mitkä ovat niitä muuttujia, jotka saattavat edesauttaa Akustikus neurinomaa sairastavien luokittelua Menièren taudista kärsiviksi.

Tautinryhmän Beningin positionaalinen vertigo osalta ilmenee sama asia kuin tautiryhmän Akustikus neurinoman vääriksi negatiiviksi luokiteltujen osalta: Suurin osa vääriksi negatiivisiksi luokitelluista on luokiteltu tautiryhmään Menièren tauti. Luvut luokitelijoilla Naiivi, *TAN*, GBN_1, GBN_2 ja GBN_3 ovat 15, 17, 16, 15 ja 19. Menièren tautiryhmään kuuluvista luokittelijoilla Naiivi ja GBN_2 ei luokiteltu yhtään kuuluvaksi tauryhmään Akustikus Neurinoma. Luokittelijoilla GBN_1, GBN_3 ja *TAN* luvut ovat 3, 1 ja 1. Menièren tautiryhmän vääristä negatiivisista on luokiteltu isohko osa kaikilla luokittelijoilla tautiryhmään Beningin positionaalinen vertigo. Luvut järjestyksessä Naiivi, *TAN*, GBN_1, GBN_2 ja GBN_3 ovat 8, 8, 6, 6, 12.

Taulukko 6 GBN_2 - 40 muuttujaa - luokittelufrekvenssit

	Akustikus Neurinoma	Bppv	Menière	Sudden Deafness	Traumatic Vertigo	Vesibular Neuritis
Akustikus neuri-noma	111	3	14	0	1	1
Bppv	0	127	15	0	2	2
Menière	0	6	301	1	3	2
Sudden Deaffness	0	1	2	38	0	0
Traumatic Vertigo	0	1	1	0	61	2
Vestibular Neuritis	0	3	3	1	3	110

Vielä palaten taulukkoon 5, jossa esitellään 40 muuttujan luokittelijoilla tunnistamistarkkuudet, ennustamistarkkuudet ja kokonaistarkkuudet. Yleisesti ottaen tunnistamistarkkuudet ovat hyviä, niin kuin voikin olettaa 40 muuttujalla.

5.5.2.1 40 muuttujaa – Naiivi luokittelija

Naiivi luokittelija 40 muuttujalla ei ole kaikkein huonoin esitellyistä viidestä luokittelijasta. Naiivin luokittelijahan olettaa, että kaikki muuttujat ovat ehdollisesti toisista riippumattomia, kun luokittelumuuttuja on annettu. Voisi olettaa että riippumattomuusoletus olisi selkeä rajoite, mutta 40 muuttujan tapauksessa naiivi-luokittelija pärjää mielestäni varsin hyvin.

Taulukossa 7 on laskettuna 40 muuttujan tapauksessa keskiarvot tarkkuuksista yli tautien. Taulukosta nähdään, että tunnistamistarkkuuden keskiarvo naiivilla 40 muuttujan luokittelijalla on 0,897, ennustamistarkkuuden keskiarvo on 0,92 ja kokonaistarkkuuden keskiarvo on 0,97.

Taulukko 7 Keskiarvot tarkkuusluvuista yli tautiryhmien - 40 muuttujaa

	NAIIVI	TAN	GBN_1	GBN_2	GBN_3
r	0,897	0,894	0,899	0,911	0,891
p	0,92	0,91	0,917	0,926	0,913
t	0,97	0,97	0,97	0,973	0,967

5.5.2.2 40 muuttujaa – TAN

Naiivin luokittelijan laajennusta – *TAN* -luokittelijaa on tutkittu useilla tahoilla (mm. Nir Fridman et al. Bayesian Network Classifiers, Jie Chang et al. Comparing Bayesian Networks) ja on esitetty, että tämä luokittelija olisi tietyissä tapauksissa parempi kuin yleiset Bayes -verkot. *TAN* -luokittelija sallii luokittelumuuttujan lisäksi toisen vanhemman solmuille, toisin sanoen solmuilla on maksimissaan kaksi vanhempaa siten, että luokittelumuuttuja on juurisolmu (ks. luku 4.4.2).

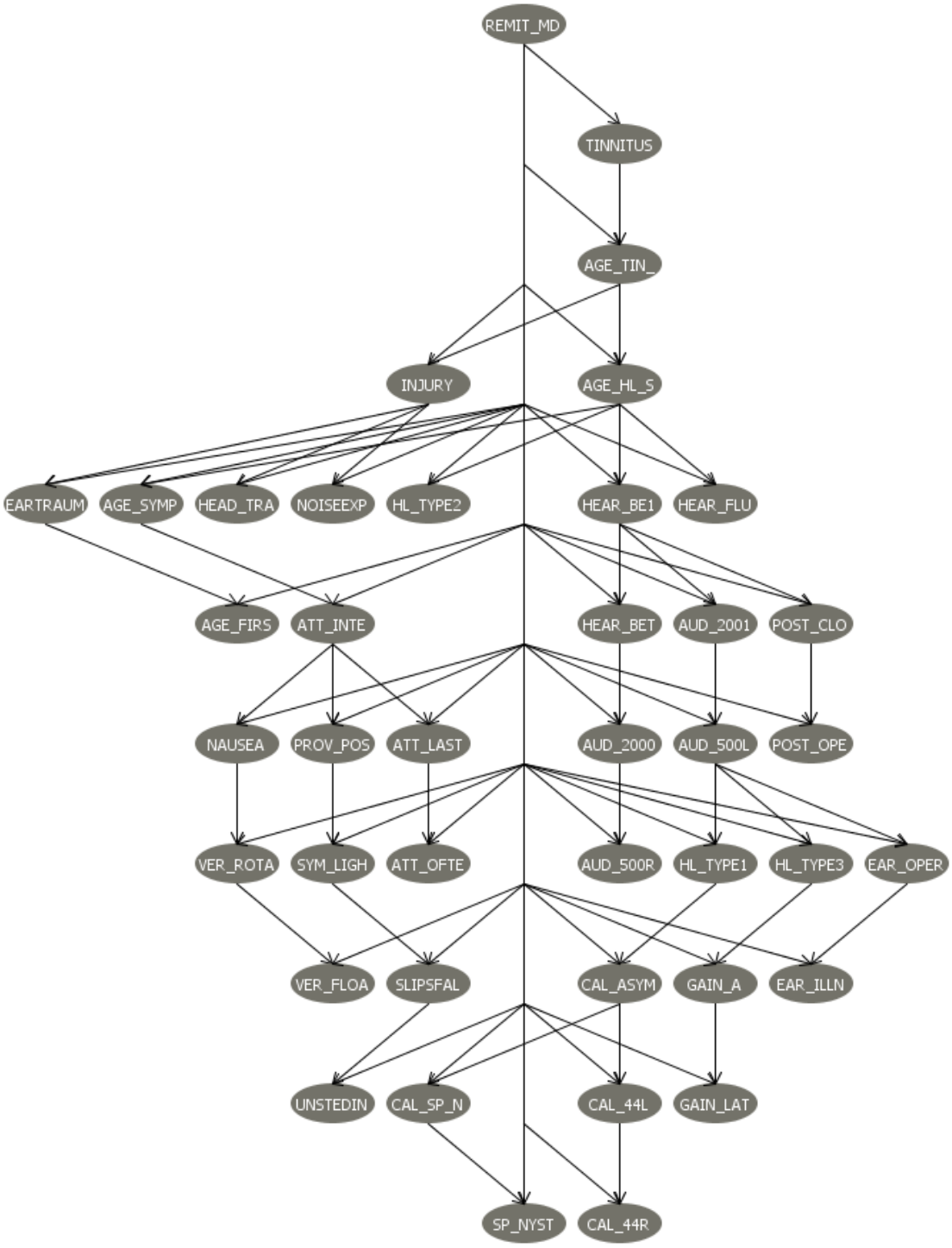
Taulukosta 7, jossa siis ovat keskiarvot yli tautiryhmien, tunnistamistarkkuuksista, ennustamistarkkuuksista ja kokonaistarkkuuksista nähdään, kuinka *TAN* pärjää luokittelutehtävässä, kun selittäviä muuttujia on 40. Oma hypoteesini oli, että TAN pärjäisi paremmin. Olettamukseni oli, että *TAN* olisi ainakin parempi kuin Naiivi-luokittelija, mikä ei tällä aineistolla pidä paikkaansa. Taulukosta 8 nähdään, että *TAN* -luokittelijalla on paras tunnistamistarkkuus tautiryhmän Akustikus Neurinoman osalta. Sitä vastoin taudin traumatic vertigo tunnistamistarkkuus on näistä viidestä luokittelijasta *TAN* -luokittelijalla heikoin, 0,831. Alla on taulukko, josta nähdään, että oikeiden positiivisten lukumäärä tautiryhmän Traumatic vertigo on 54. Väärien negatiivisten lukumäärä on 11 ja näistä 11 väärästä negatiivisesta kuusi on luokiteltu ryhmään Menièren tauti, 2 on luokiteltu ryhmään Bppv, yksi on luokiteltu ryhmään Sudden deaffness ja kaksi luokiteltu ryhmään Vestibular Neuritis. Vääriä positiivisia tautiryhmän Traumatic Vertigo on 9. Tähän tautiryhmään on luokiteltu kaikista muista tautiryhmistä lukuunottamatta tautiryhmää Sudden Deafness. Lyhyesti voidaan sanoa, että väärien negatiivisten ja väärien positiivisten kohteet/lähteet ovat varsin heterogeenisia.

Taulukko 8 TAN - 40 muuttujaa - luokittelufrekvenssit

	Akustikus Neurinoma	Bppv	Menière	Sudden Deafness	Traumatic Vertigo	Vesibular Neuritis
Akustikus neuri-noma	115	2	10	0	1	2
Bppv	0	125	17	0	2	2
Menière	1	8	297	2	3	2
Sudden Deaffness	0	1	2	38	0	0
Traumatic Vertigo	0	2	6	1	54	2
Vestibular Neuritis	1	2	3	1	3	110

Kuvassa 10 on luokittelijan *TAN* graafi, josta nähdään, kuinka luokittelijan muuttuja REMID_MD on kaikkien muiden solmujen vanhempi ja lisäksi solmuilla on maksimissaan yksi solmu lisäksi vanhempana. Kuinka graafia on tulkittava? Graafista on luettavissa, kuinka muuttujat ovat luokkaehdollisesti toisistaan riippumattomia. *Solmu on ehdollisesti riippumaton graafin muista, ei jälkeläissolmuistaan, kun solmun välitön vanhempi on annettu* - näin kuuluu ehdollisen riippumattomuuden määritelmä Bayes -verkkokontekstissa. Esimerkiksi graafista saa helposti selville seuraavan luokkaehdollisen riippumattomuuden: *P(INJURY,|AGE_HL_S, AGE_TIN, REMID_MD)=P(INJURY|AGE_TIN, REMID_MD)*, joka tarkoittaa siis sitä, että muuttuja *INJURY* on riippumaton muuttujasta *AGE_HL_S* riippumaton, kun muuttujat *INJURY* vanhemmat, muuttujat *AGE_TIN* ja *REMID_MD* on annettu. Voidaan sanoa, että muuttujan *AGE_TIN* arvojen havainnointi ei tuo lisäinformaatiota muuttujan *INJURY* arvojen ennustamiseen, kun muuttujien *REMIT_MD* ja *TINNITUS* arvot on annettu. Mitä muuta *TAN* -verkosta voi päätellä? Verkosta mielestäni näkee hyvin sen, kuinka muuttujat, joissa arvot ovat tuloksia erilaisista kliinisistä mittauksista ovat ryhmittyneet.

Muuttujasta *INJURY* on kaari muuttujiin *HEAD_TRAUMA, EAR_TRAUMA* ja *NOISEEXP*. Tämä on täysin järkeenkäypää, kun muuttujan *INJURY* arvot 0 ja 1 vastaavat kysymykseen, onko potilaalla päävamma, korvavamma tai kuulovaurio. Graafin muuttujien ja kaarien sijoittumista toisiinsa nähden voi tähän tapaan tulkita auki, mutta tämän työn puitteissa ja lääketieteellisen taustan puuttumisen vuoksi minun ei ole järkevää yrittää tulkita koko verkkoa solmu solmulta.



Kuva 10 *TAN* - 40 muuttujaa

5.5.2.3 40 muuttujaa – GBN_1

GBN_1 -verkko on siis rakennettu käyttäen hyväksi Bayes -pistemäärää. Paras tunnistamistarkkuus on taudin Sudden deaffness osalta, luvun ollessa 0,927, joka on itse asiassa sama tulos kuin muillakin *GBN*:illä. *GBN_1* on paras tunnistamistarkkuus kaikista luokittelijoista tautiryhmän Bppv osalta, 0,877. Taulukosta 9 on nähtävissä mihin ryhmään potilaat on luokiteltu ja heidän todellinen tautiryhmänsä. Tautiryhmän vestibular väärät negatiiviset on luokiteltu tasaisesti muihin tautiryhmiin. Taulukossa 7 on laskettuna keskiarvot yli tautiryhmien tunnistamistarkkuuksille, ennustamistarkkuuksille ja kokonaistarkkuuksille ja *GBN_1* luokittelijalla nämä luvut ovat tässä järjestyksessä 0,899, 0,917 ja 0,7

Taulukko 9 *GBN_1*- 40 muuttujaa - luokittelufrekvenssit

	Akustikus Neurinoma	Bppv	Menière	Sudden Deafness	Traumatic Vertigo	Vesibular Neuritis
Akustikus neuri-noma	111	2	12	1	0	4
Bppv	0	128	16	0	0	2
Menière	3	6	297	1	3	3
Sudden Deaffness	0	1	2	38	0	0
Traumatic Vertigo	0	4	3	0	57	1
Vestibular Neuritis	2	2	4	1	2	109

Kuvassa 11 on luokittelijan *GBN_1* verkkorakenne. Tälle luokittelijalle on tehty Markov -korjaus, joten kaikki muuttujat ovat luokittelijasolmun *REMID_MD* Markov -peitossa. Käytännössä tämä Markov -korjaus on useimmiten tarkoittanut sitä, että parhaimman pistemäärän tuottaman verkon löytymisen jälkeen jos solmu ei ole ollut luokittelumuuttujan Markov -peitossa, niin se on laitettu kuulumaan siihen lisäämällä kaari luokittelumuuttujasta ko. solmuun. Markov -korjauksella halutaan varmistaa se, että kaikki muuttujat osallistuvat luokittelutehtävään. Yleisessä Bayes -verkossahan luokittelumuuttujan ei ole välttämätöntä olla muiden solmujen vanhempi, ja nyt Bayes -pistemäärän avulla rakennettu onkin sellainen, että kaikki muuttujat eivät ole luokittelumuuttujan lapsia. Muuttujat *EAR_TRAUMA*, *INJURY*, *HEAD_TRAUMA* ja *HL_TYPE3* eivät ole tässä verkossa luokittelumuuttujan REMID_MD lapsia. Muuttujat

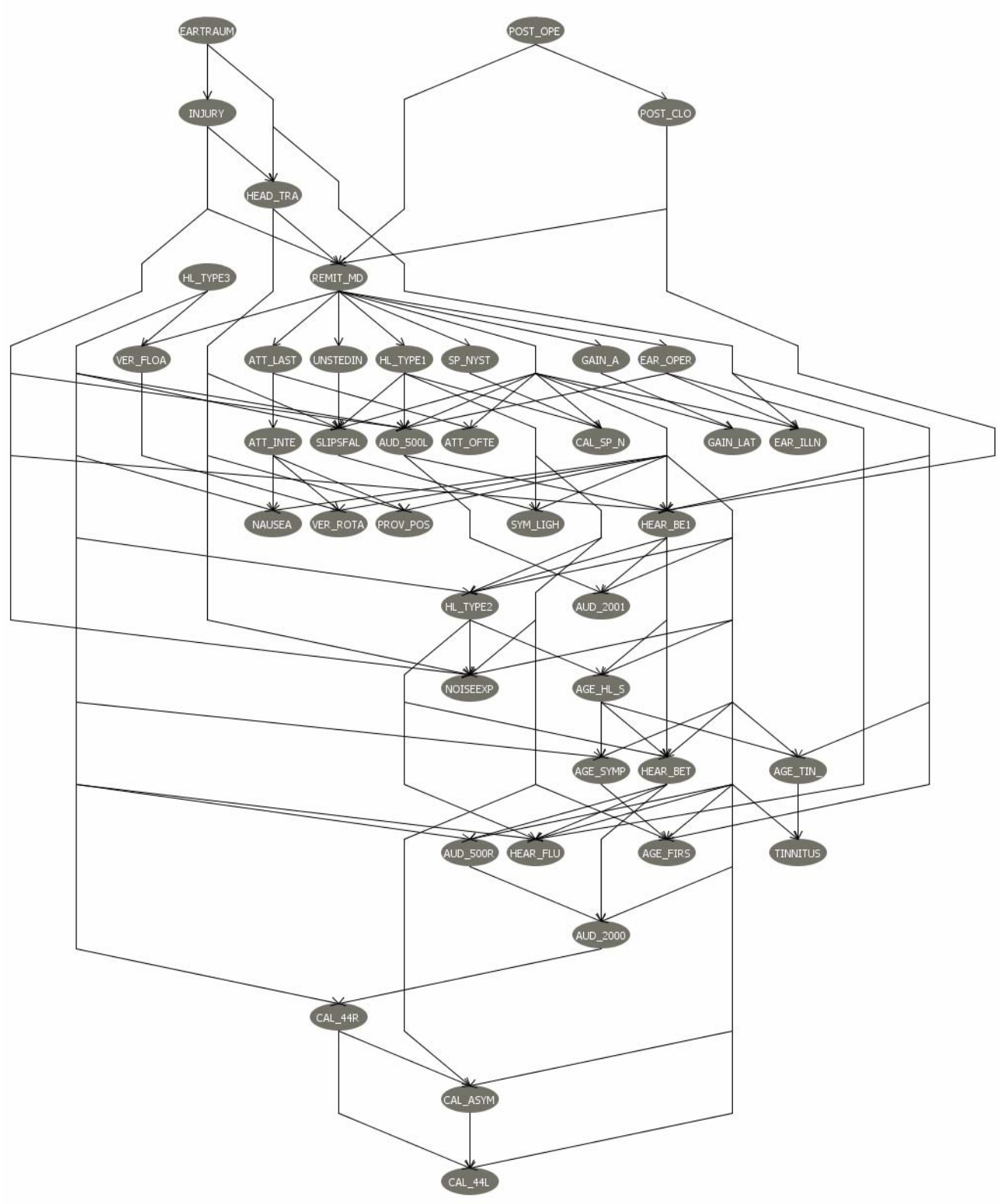
POST_OPE ja *POST_CLO* eivät myöskään ole luokittelumuuttujat lapsia, itse asiassa ne ovat luokittelumuuttujan vanhempia. Tämä johtunee Markov -korjauksesta.

Solmujen väliset ehdolliset riippumattomuudet eivät aina ole luettavissa, jos ohjenuorana käyttää tosiasiaa ”*Solmu on ehdollisesti riippumaton graafin muista, ei jälkeläissolmuistaan, kun solmun välitön vanhempi on annettu”*. Tätä tosiasiaa käytetään toki pohjana ehdollisten riippumattomuuksien selvittämiseen verkosta, mutta aina nämä riippumattomuudet eivät ole tuosta vaan nähtävissä. Kun katsotaan kuvassa 11 olevaa verkkoa, niin ovatko seuraavat ehdolliset riippumattomuudet voimassa:

1. *Ind(HEAD_TRAUMA;HL_TYPE3|INJURY)?*
2. *Ind(ATT_LAST;NAUSEA|ATT_INTE)?*
3. *Ind(INJURY;HL_TYPE1|SLIPSFAL)?*
4. *Ind(INJURY;HL_TYPE1|SYMLIGH)?*

Kysymykseen yksi vastaus on ”kyllä”. **D-separaation avulla** tämän saa selvittettyä kuvan 11 verkosta helpohkosti. Perustelun ”kyllä”-vastaukselle ovat seuraavanlaiset: Koska solmulla *HL_TYPE3* ei ole vanhempia, sitä ja solmua *HEAD_TRAUMA* yhdistävät polut kulkevat solmun *HL_TYPE3* alapuolelta. Näiden polkujen alimman solmun täytyy olla konvergoituva solmu, koska nuolien suunta on ylhäältä alaspäin. Koska solmu *INJURY* ei voi olla mikään näistä konvergoituvista solmuista eikä niiden jälkeläisistä, ehdollistaminen sen suhteen ei aiheuta riippuvuutta *HL_TYPE3*:n ja *HEAD_TRAUMA*:n välillä. (vrt. määritelmä 1, kohta 2). Kysymykseen kaksi vastaus on kyllä. Kysymyksessä on ns. sarjakytkentä; polku kulkee solmusta *ATT_LAST* evidenssiolmun *ATT_INTE* kautta solmuun *NAUSEA* ja muut polut solmujen *ATT_LAST* ja *NAUSEA* välillä kulkevat solmun *NAUSEA* alapuolella sijaitsevien konvergoituvien solmujen kautta. Koska *ATT_INTE* ei voi olla mikään näistä solmuista eikä niiden jälkeläisistä, 2. riippumattomuus on voimassa. Jos solmun *ATT_INTE* arvoa ei ole havaittu, niin tällöin solmut *ATT_LAST* ja *NAUSEA* ovat riippuvia toisistaan. Kysymykset kolme, neljä ja viisi koskevat konvergoituvaa polkua. Kysymyksiin kolme ja neljä vastaus on ”ei”, ehdolliset riippumattomuudet eivät ole voimassa. Koska solmujen välillä on konvergoituva haara *INJURY* → *SLIPSFAL* ← *HL_TYPE1* ja solmu *SLIPSFAL* on evidenssisolmu, niin

ehdollinen riippumattomuus ei ole voimassa. D-separaatiomenetelmän mukaan, jos konvergoituvan solmun yksikin jälkeläinen on havaittu, niin tällöin ehdollinen riippumattomuus ei ole voimassa – tämä on perustelu kysymyksen neljä ”ei”-vastaukselle.



Kuva 11 GBN_1- 40 muuttujaa

5.5.2.4 40 muuttujaa – GBN_2

GBN_2 -verkko on rakennettu käyttäen hyväksi *MDL* -pistemäärää. Tämä luokittelija 40 muuttujalla pärjäsi parhaiten luokittelutehtävässä. Taulukossa 10 on esitettynä keskiarvot yli tautiryhmien tunnusluvuista tunnistarkkuus, ennustamistarkkuus ja kokonaistarkkuus – ja tällä luokittelijalla on parhaat luvut kaikista luokittelijoista. Kun katsotaan taulukkoa, jossa näkee tarkkuudet jokaisesta tautiryhmästä erikseen, niin nähdään, että kaikissa muissa tautirymissä, paitsi Akustikus neurinoma ja Bppv, tällä luokittelijalla on parhaat tunnistamistarkkuudet. Tautiryhmän Traumatic vertigo osalta on melko huono ennustamistarkkuus, kun vertaa muihin luokittelijoihin – tämä selittyy verrattuna muihin luokittelijoihin suuremmalla *fp*-luvulla.

Kuvassa 12 on luokittelijan GBN_2 verkkorakenne. Verkon rakenne on mielestäni erittäin mielenkiintoinen. Suurimmalla osalla solmuista on vain yksi vanhempi ja se solmu on luokittelumuuttuja *REMID_MD*. Voisi sanoa, että tämä verkkorakenne muistuttaa paljon Naiivin luokittelijan verkkorakennetta. Luokittelijan GBN_2 verkon rakentamisessa on myös käytetty Markov -korjausta. Markov -korjauksen kanssa luokittelijasta tuli himpun verran parempi kuin ilman Markov –korjausta ja siksi tässä esimerkiksi taulukoissa 5 ja 7 esitellyt tulokset ovat Markov -korjauksella saadun luokittelijan tuloksia. Ilman Markov -korjausta verkosta jäi kokonaan ulkopuolelle muuttuja *HL_TYPE3*. Lisäksi Markov-peiton ulkopuolelle jäivät muuttujat *POST_OPE*, *UNSTEDIN*, *GAIN_LAT*, *AGE_FIRST_AUD_500L*, *EAR_ILLNESS*, *CAL_44R*, *CAL_44L*, *EAR_OPER*, *CAL_SP_N*, *AUD_2000* ja *AUD500R*. Eli yhteensä 13 muuttujaa jäi Markov -peiton ulkopuolelle, kun verkko rakennettiin ilman Markov -korjausta. Luokittelutehtävässähän käytetään vain muuttujia, jotka ovat luokittelumuuttujan *REMID_MD* Markov -peitossa. Se että muuttuja jää Markov -peiton ulkopuolelle ei tarkoita sitä, että muuttuja olisi ollut luokittelutehtävän kannalta merkityksetön muuttuja.

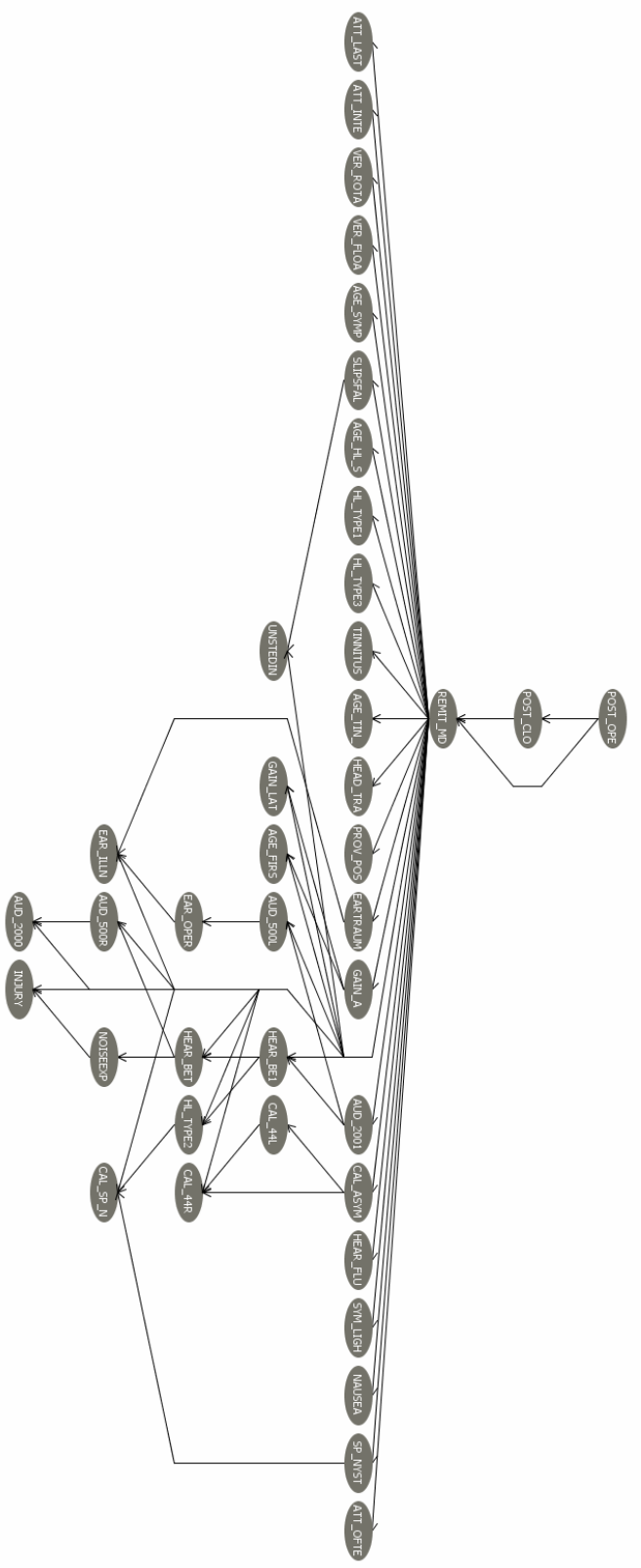
Kun rakennetaan luokittelevaa Bayes-verkkoa, niin on esitetty, että *MDL* -pistemäärä ei ole välttämättä paras pistemäärä verkon rakenteen löytämiseen. Tässä siis tarkoitetaan luokittelevaa Bayes -verkkoa, jonka verkkorakenne on rajoittamaton (ks. määritelmä luvusta 4.4.4). Nir Friedman et al. artikkelissaan *Bayesian Network Classifier* esittävät, että suurilla datoilla, ts. datoilla, joissa on suuri muuttujajoukko, verkkorakenteen

löytäminen korkealla *MDL* -pistemäärällä ei takaa sitä, että verkko olisi hyvä luokittelja [10]. He painottavat, että on ymmärrettävä ero hyvän ennustamistarkkuuden ja hyvän *MDL* -pistemäärän välillä. Heidän tutkimuksensa yksi tulos oli, että yleiset Bayes -verkot, joiden verkkorakenne oli rakennettu käyttäen *MDL* –pistemäärää, ja joissa muuttujamäärä oli yli 15, olivat huonoja luokittelijoita. Heidän aineistoissaan, joissa oli 35 ja 36 muuttujaa, *MDL* -pistemäärällä rakennettu verkko käytti luokittelussa vain viittä muuttujaa, toisin sanoen vain viisi muuttujaa oli luokittelumuuttujan Markov -peitossa. Friedman esittää, että muuttujien valinta (engl. *feature selection*) on toisissa aineistossa hyvä asia, sillä saadaan erotettua turhat muuttujat pois. Toisinaan tuo menetelmä saattaa kuitenkin jättää huomiotta muuttujia, jotka ovat oleellisia luokittelun kannalta. Friedman et al. esittävät, että myös muilla pistemääräfunktioilla esiintyy samaa ongelmaa.

Tällä aineistolla, kun ei käytetty Markov -korjausta, päästiin likimain samoihin tuloksiin kuin Markov -korjauksen käytöllä. Taulukosta 10 löytyy tarkkuudet tautiryhmittäin, kun **a)** Markov-korjausta ei ole käytetty ja **b)** Markov-korjausta on käytetty. Muuttujia on edelleen 40 ja pistemäärä siis *MDL*. Keskiarvot yli tautiryhmien näistä tarkkuusluvuista ovat Markov -peiton kanssa r_k=0,908, p_k=0,922 ja t_k=0,971. Markov -peiton kanssa keskiarvot ovat r_k=0,911, p_k=0,926 ja t_k=0,973. Tällä aineistolla muuttujien valinta mielestäni toimii. Otoneurologian asiantuntijat voivat olla toki toista mieltä siitä, että ovatko nuo mainitsemani 13 muuttujaa tarpeettomia luokittelutehtävässä.

Taulukko 10 *GBN_2* Tarkkuusluvut a) ilman ja b) Markov – korjauksen kanssa

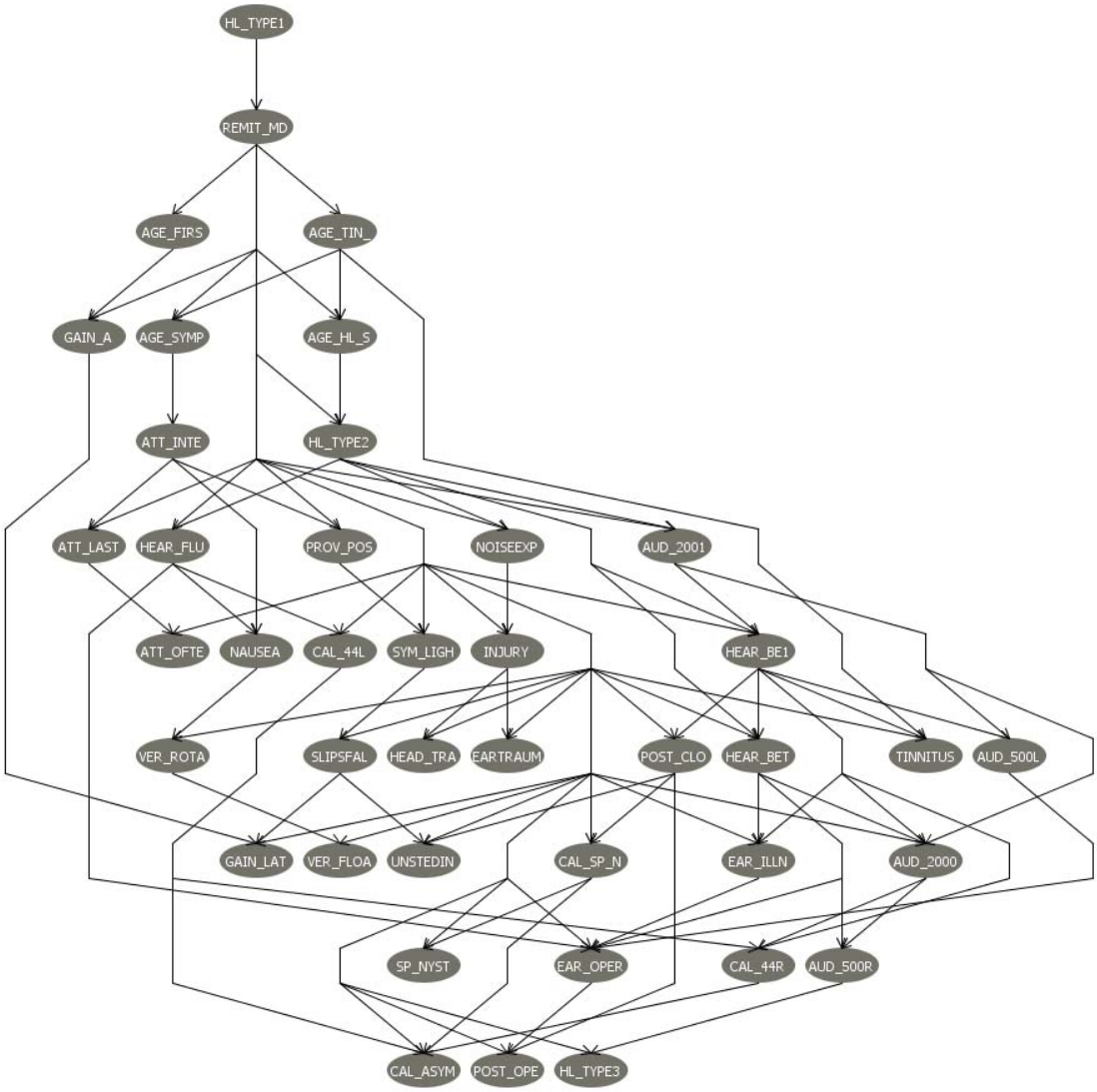
	R_a	R_b	P_a	P_b	T_a	T_b
Akustikus neuri-noma	0,831	0,854	1	1	0,974	0,977
Bppv	0,87	0,87	0,894	0,901	0,959	0,960
Menière	0,955	0,962	0,89	0,896	0,938	0,943
Sudden Deaffness	0,951	0,927	0,951	0,95	0,995	0,994
Traumatic Vertigo	0,923	0,938	0,882	0,871	0,984	0,984
Vestibular Neuritis	0,917	0,917	0,917	0,94	0,976	0,979



Kuva 12 GBN_2 - 40 muuttujaa

5.5.2.5 40 muuttujaa –GBN_3

Tämän luokittelijan rakentamiseen on käytetty *AIC* -pistemäärää. Taulukkojen 5 ja 7 luvuista nähdään, että tämä luokittelija suoriutuu yleisesti kaikkein huonoiten luokittelutehtävästä. Parhaimpiin tarkkuuslukuihin tällä luokittelijalla päästiin tautiryhmällä Sudden Deafness. Samoihin parhaimpiin tarkkuuksiin tämän tautiryhmän osalta ylsi myös luokittelija GBN_2: tunnistamistarkkuus on 0,927, ennustamistarkkuus on 0,95 ja kokonaistarkkuus on 0,994. Kuvassa 13 on graafi luokittelijasta GBN_3 ja siitä voidaan sanoa, että graafin rakenne on melko samantyyppinen kuin mitä luokittelijan GBN_1 graafi on.



Kuva 13 GBN_3 - 40 luokittelijaa

5.5.3 Yhdeksän muuttujaa

Luvussa 5.1.1 Aineiston tausta kerroin, kuinka yhdeksän muuttujan joukko on määritetty tärkeäksi luokittelutehtävän kannalta [24]. Rakensin näillä yhdeksällä selittävällä muuttujalla Naiivin luokittelijan, *TAN* -luokittalijan sekä yleiset Bayes -verkot GBN_1, GBN_2 ja GBN_3. Nämä yhdeksän tärkeää muuttujaa ovat *AGE_SYM*, *ATT_OFTE*, *ATT_LAST*, *ATT_INTE*, *TINNITUS*, *HEAD_TRAUMA*, *AGE_TIN_SYM*, *AGE_HL_SYM* ja *SYM_LIGH*.

Taulukossa 11 on nähtävissä tarkkuusluvut näillä luokittelijoilla tautiryhmittäin. Yleisesti yhdeksällä muuttujalla näillä luokittelijoilla kaikki tarkkuusluvut laskivat verrattuna 40 muuttujan luokittelijoihin. Tämä oli odotettavissa, ja tämä on normaali tulos. Tautiryhmän Sudden Deafness osalta tunnistamistarkkuus ja ennustamistarkkuus ovat parhaimmillaankin niin alhaiset kuin 0,439 ja 0,486. Kentala et al. [28] esittävät, että tämän taudin osalta ei voi sanoa, että juuri tietyt muuttujat olisivat ratkaisevia luokittelun kannalta. Nyt kun muuttujien määrää tiputettiin radikaalisti 40 muuttujasta yhdeksään muuttujaan, niin mielestäni ei ole ihmekään, että vaikeasti diagnosoitavan taudin luokittelussa tarkkuusluvut ovat noin alhaiset. Kokonaistarkkuudet ovat yleisesti hyvät – tämä johtuu oikeiden negatiivisten (t=(tp+tn)/(tp+tn+fp+fn)) korkeasta lukumäärästä. Taulukon 11 tarkkuuksista nähdään, kuinka Naiivi luokittelija on selviytynyt luokittelutehtävästä yhdeksällä muuttujalla varsin hyvin.

Taulukko 11 Tarkkuusluvut - yhdeksän muuttujaa

		NAIIVI	TAN	GBN_1	GBN_2	GBN_3
Akustikus neurinoma	**r**	0,738	**0,785**	**0,785**	0,731	0,738
	p	**0,889**	0,836	0,879	0,888	**0,889**
	t	0,946	0,943	**0,95**	0,945	0,946
Bppv	**r**	**0,829**	0,801	0,822	**0,829**	0,815
	p	**0,858**	0,83	0,833	**0,858**	0,804
	t	0,946	0,937	0,94	0,946	0,933
Menière	**r**	**0,917**	0,853	0,891	**0,917**	0,863
	p	**0,813**	0,797	**0,813**	**0,813**	0,801
	t	**0,891**	0,868	0,88	**0,891**	0,872
Sudden Deaffness	**r**	0,39	0,39	0,415	0,39	**0,439**
	p	0,485	0,412	**0,486**	0,471	0,45
	t	**0,95**	0,944	**0,95**	0,949	0,946
Traumatic Vertigo	**r**	**0,895**	0,769	0,785	0,815	0,8
	p	**0,869**	0,847	0,85	**0,869**	0,839
	t	0,976	0,971	0,972	0,976	0,972
Vestibular Neuritis	**r**	**0,892**	0,842	0,858	**0,892**	0,883
	p	**0,899**	0,842	0,88	**0,899**	0,883
	t	**0,97**	0,954	0,963	**0,97**	0,966

Taulukossa 12 on keskiarvot tarkkuuksista yli tautiryhmien. Naiivilla luokittelijalla keskiarvot ovat kaikkein parhaimmat. Luokittelijalla GBN_2, jossa pistemääränä siis on *MDL*, on toiseksi parhaimmat keskiarvot tarkkuuksista. Itse asiassa, kun vuorikiipeilyalgortimissa ei käytetty kaaren kääntöä, niin GBN_2:n verkkorakenne on sama kuin Naiivilla luokittelijalla! Tästä aiheesta tarkemmin luvussa 5.5.3.1.

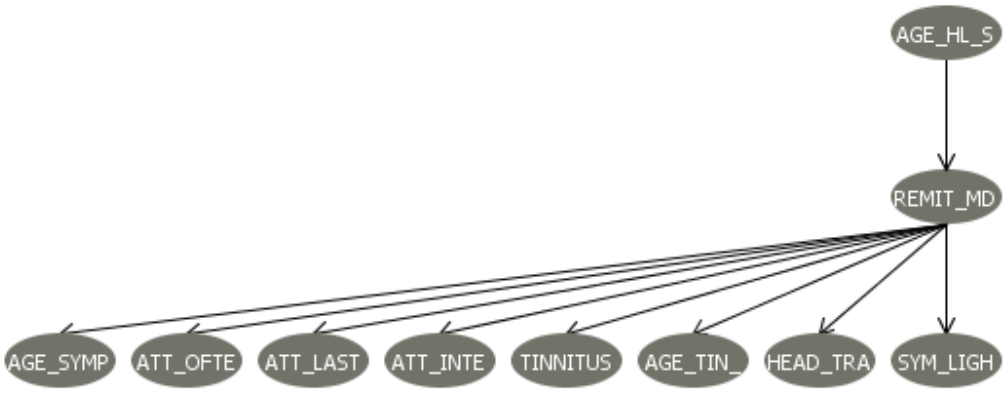
Taulukko 12 Yhdeksän muuttujaa - keskiarvot tarkkuusluvuista yli tautiryhmien

	NAIIVI	TAN	GBN_1	GBN_2	GBN_3
r	0,78	0,74	0,76	0,76	0,76
p	0,80	0,76	0,79	0,80	0,78
t	0,95	0,94	0,94	0,95	0,94

5.5.3.1 Yhdeksän muuttujaa – luokittelijat Naiivi ja GBN_2

Kuten edellisessä luvussa tuli mainittua, niin Naiivi luokittelija ja GBN_2 pärjäsivät luokittelutehtävässä parhaiten yhdeksällä muuttujalla. Kuvassa 13 on yhdeksän selittävän muuttujan GBN_2, jossa siis pistemäärämä on *MDL*, hakualgoritmi on vuorikiipeilyalgoritmi, aloitusverkko on naiivi ja Markov -korjaus on otettu mukaan. Rakenne on yhtä solmua lukuun ottamatta, kuten Naiivilla luokittelijalla on. Kokeilin

rakentaa verkkoa GBN_2 siten, että vuorikiipeilyalgoritmissa sallitut kaarien toimenpiteet ovat vain lisäys ja poisto, eli kaaren kääntämistä ei sallita. Tällä tavoin rakennetun luokittelijan rakenne on täysin Naiivia vastaava. Koska sain tällaisen tuloksen, niin käsittelen Naiivia ja GBN_{2b} (=GBN_2 ilman kaaren kääntöä) yhtenä ja saman luokittelijana.

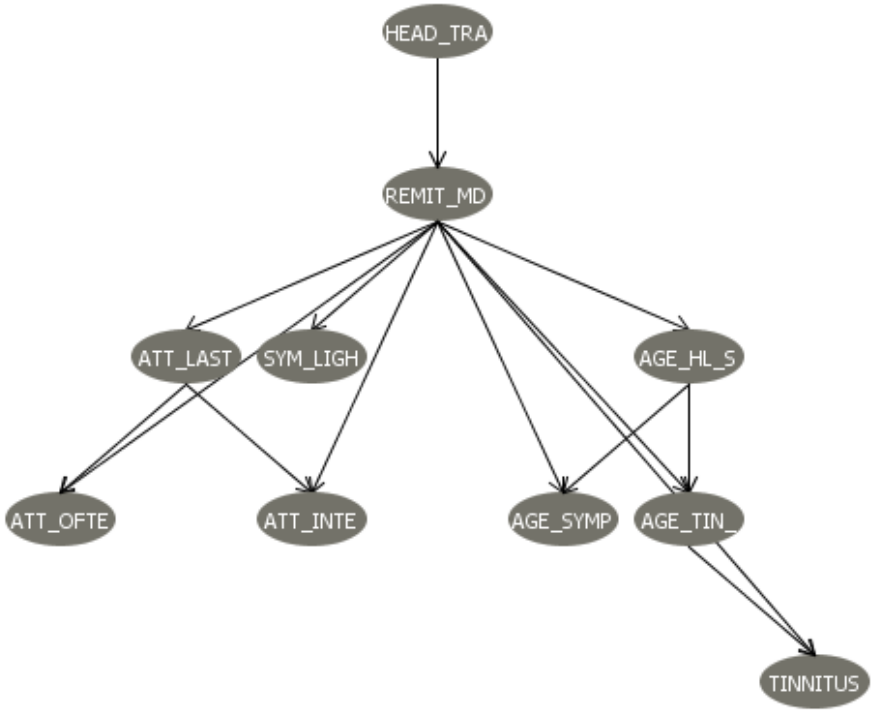


Kuva 14 GBN_2 - Yhdeksän muuttujaa

Liitteessä 2 on Naiivin luokittelijan todennäköisyysjakaumat (posteriorijakaumat). Jakaumista nähdään, että on erittäin todennäköistä, että jos sairastaa tautia vestibular neuritis, että potilas ei kärsi kuulonmenetyksestä. Nimittäin todennäköisyys P(*AGE_HL_S*=0|*REMID_MD*=Vestibular Neuritis)=0,884. Sama pätee myös kyseisen taudin osalta tinnitukseen; P(*TINNITUS*=0|*REMID_MD*=Vestibular neuritis)=0,722. Mielestäni näistä jakaumista on hyvin nähtävissä se, kuinka todennäköisyysjakaumien perusteella ei voi sanoa kovinkaan paljoa taudista sudden deaffness. Luokkaehdollisissa jakaumissa tästä tautiryhmistä arvon ”ei oireita tai ei vammaa” saa 30 - 40 % kun loput havainnot jakautuvat suhteellisen tasaisesti loppuihin luokkiin. Poikkeuksena ovat toki dikotomiset muuttujat *HEAD_TRAUMA* ja *SYM_LIGH*, joissa prosentit menevät 99.6%/0.4% ja 56%/44% – näidenkään lukujen perusteella ei voi sanoa juuri mitään taudista Sudden deaffness. Muuttujan *HEAD_TRAUMA* osalta muillakin tautiryhmillä on samanlaiset todennäköisyydet.

5.5.3.2 Yhdeksän muuttujaa – luokittelijat GBN_1 ja GBN_3

Taulukossa 12, jossa on esitettynä keskiarvot tarkkuuksista yli tautiryhmien nähdään, että näillä kahdella luokittelijalla on suunnilleen samat lukemat. Nämä kaksi luokittelijaa selviytyivät muita paremmin tautiryhmän Sudden Deaffness luokittelussa, mikä nähdään taulukosta 11. Kuvassa 15 on luokittelijan GBN_1 graafi. Tässä luokittelumuuttuja *REMID_MD* on kaikkien muiden paitsi solmun *HEAD_TRAUMA* vanhempi. Muuttuja *AGE_HL_S*, joka kertoo kuulo-oireiden kestosta, on muuttujien *AGE_TIN* (kuinka kauan tinnitus kestänyt) ja *AGE_SYMP* (kuinka kauan oireet yleensä ovat kestäneet) vanhempina. Nyt siis *AGE_TIN* ja *AGE_SYMP* ovat toisistaan ehdollisesti riippumattomia, kun *REMID_MD* ja *AGE_HL_S* on annettu. Tämä vaikuttaa järkeenkäyvältä; kun on tiedossa tautiryhmä ja kuulo-oireiden kesto, niin tinnituksen keston tietäminen ei toisi lisäinformaatiota oireiden keston yleensä ennustamiseen.

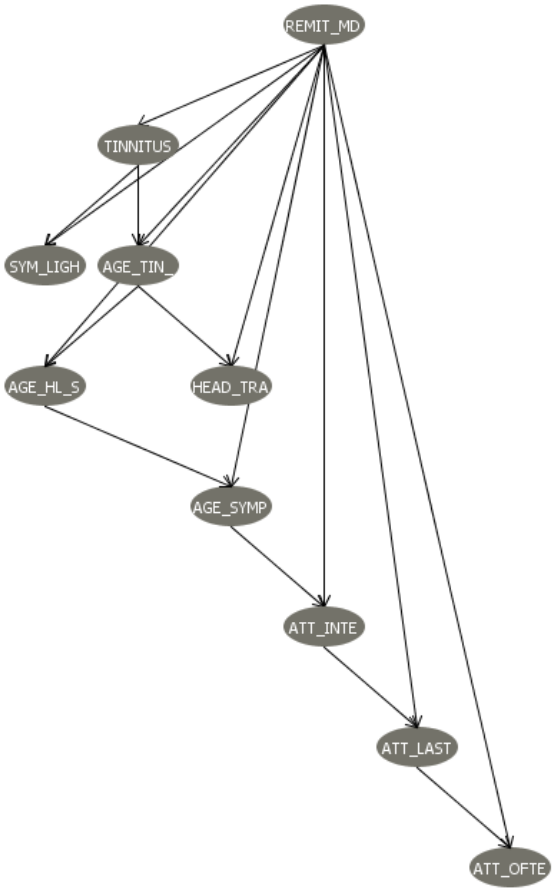


Kuva 15 GBN_1 - Yhdeksän muuttujaa

5.5.3.3 Yhdeksän muuttujaa – TAN-luokittelija

Yleisesti ottaen yhdeksän selittävän muuttujan *TAN* -luokittelija on huonoin luokittelija, kun vertaillaan taulukossa 12 olevia keskiarvoja tarkkuusluvuista yli tautiryhmien. Kun katsotaan taulukkoa 11, jossa on tautiryhmäkohtaiset tarkkuusluvut, niin nähdään, että *TAN* -luokittelijalla on kaikista luokittelijoista parhain tunnistamistarkkuus taudin Akustikus Neurinoma osalta. Ennustamistarkkuus ja kokonaistarkkuus eivät sitä vastoin ole parhaita tämän tautiryhmän osalta *TAN* -luokittelijalla johtuen suuresta väärien positiivisten lukumäärästä (*fp*). Kuvassa 16 on *TAN* -luokittelijan graafi. Nähdään kuinka solmu *TINNITUS* on solmun *AGE_TIN* vanhempi ja kuinka edelleen solmu *AGE_TIN* on solmun *HEAD_TRAUMA* vanhempi. Liitteessä 3 ovat *TAN* -luokittelijan luokkaehdolliset todennäköisyysjakaumat (posteriorijakaumat).

Todennäköisyydet *P(HEAD_TRAUMA=1|AGE_TIN=0, REMID_MD=5)=0,836* ja *P(HEAD_TRAUMA=1|AGE_TIN=1, REMID_MD=5)=0,5.* Jos siis on tiedossa, että potilas on kärsinyt tinnituksesta muutamia päiviä, ja potilaalla on tauti traumatic vertigo, niin todennäköisyys, että potilaalla on ns. takanaan HEAD_TRAUMA, on 0,5. Sitä vastoin, jos potilas ei kärsi tinnituksesta ja potilaalla on tauti traumatic vertigo, niin päävamman todennäkösyys kasvaa 86,6 prosenttiin. Muilla muuttujan *AGE_TIN* (tinnitus kestänyt viikoja - neljä vuotta tai enemmän) arvoilla tuossa tautiryhmässä päävamman todennäköisyys on n. 80 - 100 %. Tautiryhmillä lukuunottamatta tauteja traumatic vertigo ja akustikus neurinoma, riippumatta muuttujan *AGE_TIN* arvosta, päävamman todennäköisyys on pieni, maksimissaan luku on 18 %. Tästä voi päätellä sen, että näillä tautiryhmillä päävamma ei ole kovin yleistä. Tautiryhmän Akustikus neurinoma osalta on niin, että jos tästä taudista kärsivä potilas on kärsinyt tinnituksesta muutamia päiviä, niin todennäköisyydellä 0,5 hänellä on päävamma. Muilla muuttujan *AGE_TIN* arvoilla, ja kun kysymyksessä on tauti akustikus neurinoma, päävamman todennäköisyys on pieni (muuttujan *AGE_TIN* arvoilla 0, 3,4,5,6 päävamman todennäköisyyydet ovat 0,003 - 0,016, arvolla 2 todennäköisyys on 0,125).



Kuva 16 *TAN* - Yhdeksän muuttujaa

5.5.4 Viisi muuttujaa

Kuten luvussa 5.1.1 Aineiston tausta on kerrottu, niin viisi avainmuuttujaa, joita ilman diagnoosia ei voi tehdä, ovat *HEAD_TRAUMA*, *AGE_HL_SYM*, *ATT_LAST*, *AGE_SYMP* ja *ATT_OFTE*. Näillä muuttujilla rakennetiin viisi luokittelijaa, joista parhaiten pärjäsi Naiivi-luokittelija ja GBN_{2b}, joista jälkimmäinen yleinen verkko, *MDL* -pistemäärällä, kun kaaren kääntämistä ei ole sallittu vuorikiipeily-algoritmissa. Itse asiassa, kuten yhdeksän muuttujan verkoissa, niin nämä kaksi luokittelijaa ovat samat. Taulukossa 14 on keskiarvot tarkkuuksista yli tautiryhmien, josta nähdään muun muassa, että *TAN* on myös viidellä muuttujalla huonoin luokittelija. Taulukossa 15 on tarkkuudet tautiryhmittäin. Huomion arvoista on, että tarkkuusluvut eivät ole paljoa laskeneet yhdeksän muuttujan luokittelijoiden tarkkuuksista. Itse asiassa joidenkin tautiryhmien

kohdalla tunnistamistarkkuudet ovat nousseet tietyillä luokittelijoilla, kun vertaa viiden muuttujan luokitelijoita yhdeksän muuttujan luokittelijoihin.

Tautiryhmien Sudden Deaffness ja Bppv osalta tunnistamistarkkuusluvut nousivat luokittelijoilla Naiivi, *TAN* ja GBN_2. Tautiryhmän Traumatic Vertigo osalta tunnistamistarkkuus nousi luokittelijalla GBN_3. Kaikilla luokittelijoilla on parempi tunnistamistarkkuus tautiryhmän Vestibular Neuritis osalta. Lisäksi myös kaikissa tautiryhmissä nousi ennustamistarkkuudet ja kokonaistarkkuudet tietyillä luokittelijoilla.

Taulukko 13 Viisi muuttujaa keskiarvot tarkkuusluvuista yli tautiryhmien

	NAIIVI	TAN	GBN_1	GBN_2	GBN_3
r	0,7685	0,743333	0,751167	0,768	0,761667
p	0,803667	0,772833	0,7855	0,802833	0,801
t	0,944833	0,935333	0,9415	0,9445	0,943667

Taulukko 14 Viisi muuttujaa - tarkkuusluvut

		NAIIVI	TAN	GBN_1	GBN_2	GBN_3
Akustikus Neuri-noma	**r**	0,723	**0,754**	0,731	0,723	0,715
	p	**0,904**	0,803	0,864	**0,904**	0,903
	t	**0,946**	0,934	0,941	**0,946**	0,945
Bppv	**r**	**0,849**	0,788	0,801	**0,849**	**0,849**
	p	0,832	0,804	**0,88**	0,827	0,827
	t	0,944	0,93	0,947	0,943	0,943
Menière	**r**	**0,885**	0,85	0,895	0,882	**0,885**
	p	**0,805**	0,794	0,791	**0,805**	0,801
	t	**0,879**	0,865	0,874	0,878	0,877
Sudden Deaffness	**r**	0,415	**0,439**	0,366	0,415	0,39
	p	0,531	0,514	0,417	0,531	**0,533**
	t	**0,954**	0,952	0,944	**0,954**	**0,954**
Traumatic Vertigo	**r**	**0,831**	0,754	**0,831**	**0,831**	**0,831**
	p	0,871	**0,875**	0,885	0,871	0,871
	t	0,977	0,972	**0,978**	0,977	0,977
Vestibular Neuritis	**r**	0,908	0,875	0,883	0,908	0,9
	p	0,879	0,847	0,876	0,879	0,871
	t	0,969	0,959	0,965	0,969	0,966

5.5.4.1 Viisi muuttujaa – viisi luokittelijaa

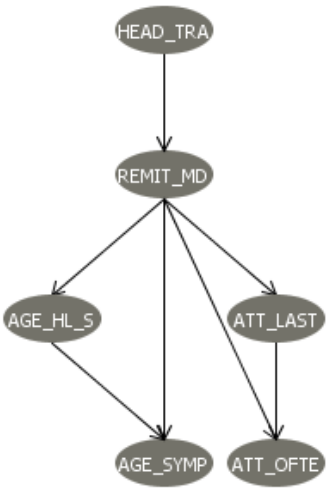
Kuten edellisessä luvussa tuli kerrottua, niin *MDL* -pistemäärällä saatu luokittelija graafeineen muistuttaa erehdyttävästi Naiivia luokittelijaa. Ero on vain siinä, että solmu *AGE_HL_SYM* on solmun *REMID_MD* vanhempi, muuten luokittelija on kuin Naiivi

luokittelija. Kun vuorikiipeilyalgorimista poistettiin kaaren käännön mahdollisuus, niin *MDL* -pistemäärällä saatiin viidellä muuttujalla Naiivi luokittelija, aivan kuten yhdeksänkin muuttujan tapauksessa. Viidellä muuttujalla tunnistamis- ja ennustamistarkkuuksien keskiarvot nousivat hieman Naiivilla luokittelijalla, kokonaistarkkuus laski sitä vastoin hieman.

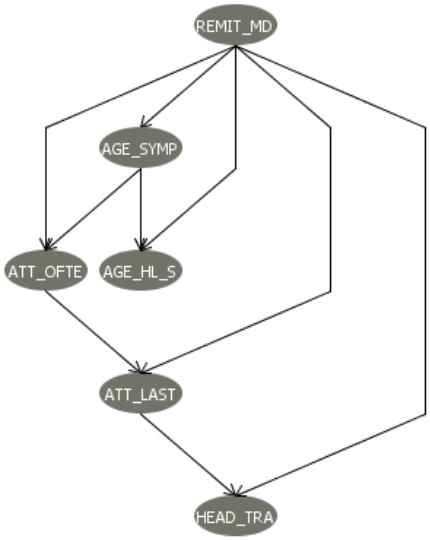
Viiden muuttujan *TAN* -luokittelijalla tilanne on sama kuin Naiivilla luokittelijalla. Keskiarvot tunnistamis – ja ennustamistarkkuuksista paranivat hiukan kokonaistarkkuuden keskiarvon laskiessa hiukan. Kuvassa 18 on *TAN* -luokittelijan graafi.

Luokittelijalla GBN_3 kaikki kolme keskiarvoa tarkkuusluvuista nousivat, kun verrataan yhdeksän muuttujan GBN_3 luokittelijaan. Tunnistamistarkkuudet ovat itse asiassa nousseet kaikissa muissa tautiryhmissä lukuunottamatta tautiryhmiä Akustikus Neurinoma ja Sudden Deaffness. Kuvassa 19 on luokittelijan GBN_3 graafi.

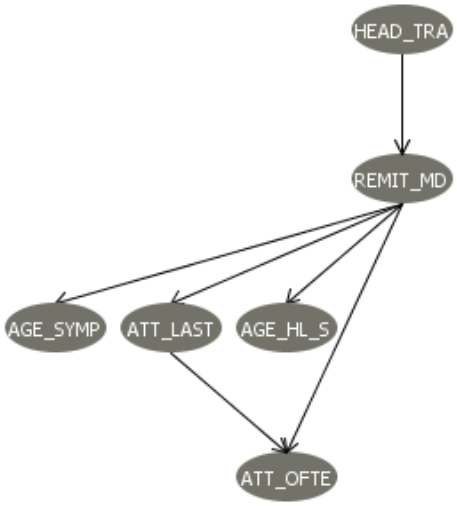
Luokittelijalla GBN_1 nousi tunnistamistarkkuus tautiryhmien Menièren tauti, Traumatic Vertigo ja Vestibular Neuritis osalta, kun verrataan yhdeksän muuttujan luokittelijaa GBN_1. Ennustamis- ja kokonaistarkkuudet nousivat tautiryhmien Bppv, Traumatic Vertigo ja Vestibular Neuritis. Kuvassa 17 on luokittelija GBN_1 graafi.



Kuva 17 GBN_1 - viisi muuttujaa



Kuva 18 *TAN* - viisi muuttujaa

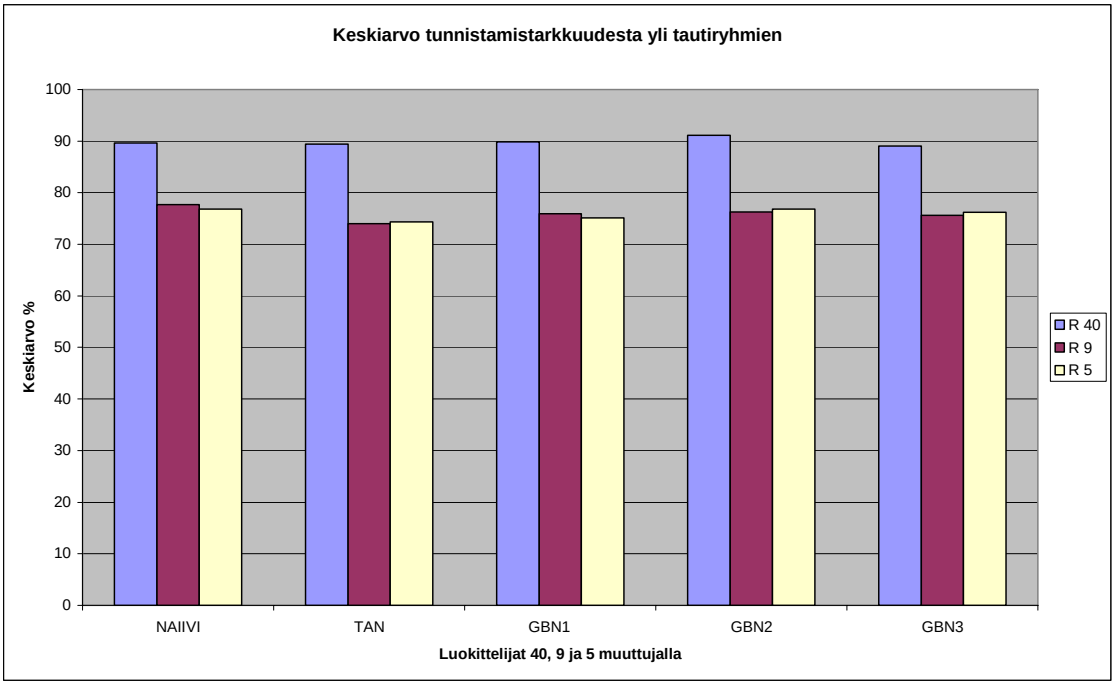


Kuva 19 GBN_3 - viisi muuttujaa

5.6 Yhteenveto tuloksista

Kuvassa 20 on esitettynä keskiarvot tunnistamistarkkuuksista yli tautiryhmien, kun selittäviä muuttujia on 40, yhdeksän ja viisi. Kun muuttujia on 40, niin paras luokittelija tällä tunnusluvulla mitattuna on GBN_2, joka on yleinen Bayes -verkko, jonka verkkorakenteen määrittämisessä on käytetty *MDL* -pistemäärää. Yhdeksällä muuttujalla

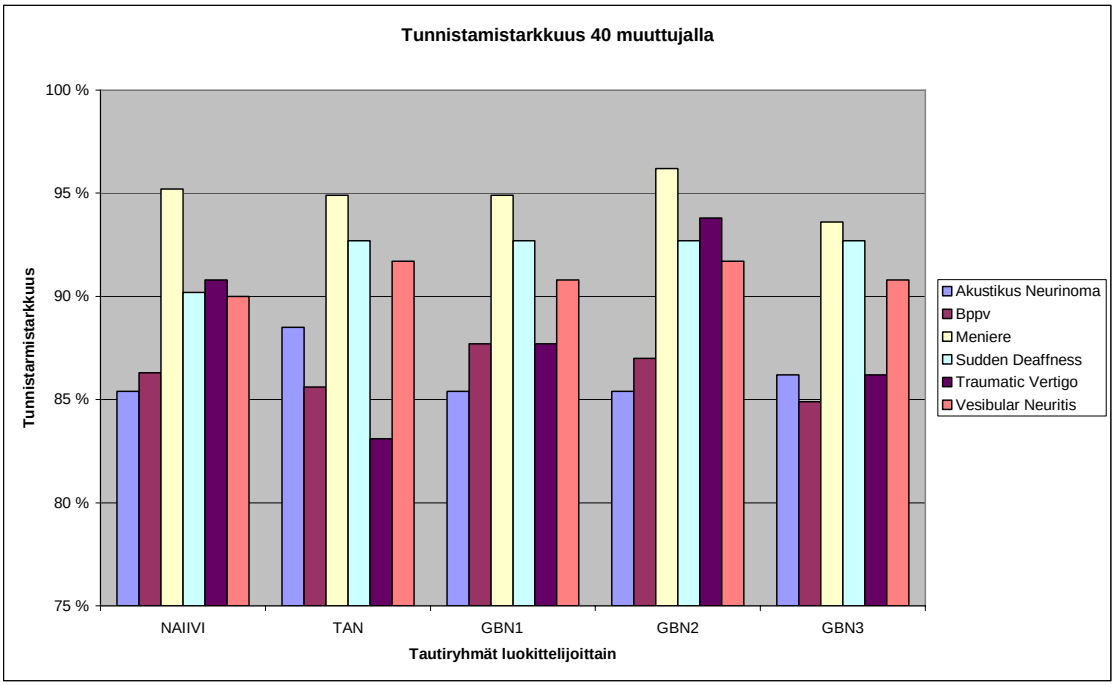
tunnistamistarkkuudet laskivat, mikä olikin odotettavissa, yli kymmenen prosenttia. Parhaiten yhdeksän muuttujan luokittelijoista selvisi Naiivi luokittelija, joka on sama kuin GBN_{2b} (=*MDL* pistämäärä ja vuorikiipeilyalgoritmi ilman kaaren kääntöä). Toiseksi paras yhdeksän muuttujan luokittelija tunnistamistarkkuudella mitattuna on GBN_2. Viiden muuttujan luokittelijoista paras luokittelija on Naiivi luokittelija. Huomionarvoista on se, että tarkkuusluvut eivät ole laskeneet juuri lainkaan, päinvastoin, joillain luokittelijoilla tarkkuusluvut nousivat.



Kuva 20 Keskiarvot tunnistamistarkkuudesta yli tautiryhmien eri luokittelijoilla eri muuttujalukumäärillä

40 muuttujalla eri tautiryhmien tunnistamistarkkuudet ovat hyvät. Kuten kuvasta 21 ilmenee, niin taudista Menièren tauti kärsivät potilaat on osattu varsin hyvin luokitella oikeaan tautiryhmään. Myös tautien Sudden Deafness ja Vestibular Neuritis osalta luokittelutehtävä on onnistunut; tarkkuusluvut ovat yli 90 prosenttia tai yli kaikilla luokittelijoilla. Lisäksi nähdään, että luokittelijoilla Naiivi ja GBN_2 on tautiryhmän Traumatic Vertigo luokittelu sujunut hyvin, luokittelijalla GBN_2 jonkin verran paremmin kuin Naiivilla. Tautiryhmien Akustikus Neurinoma ja Bppv osalta luokittelu sujui

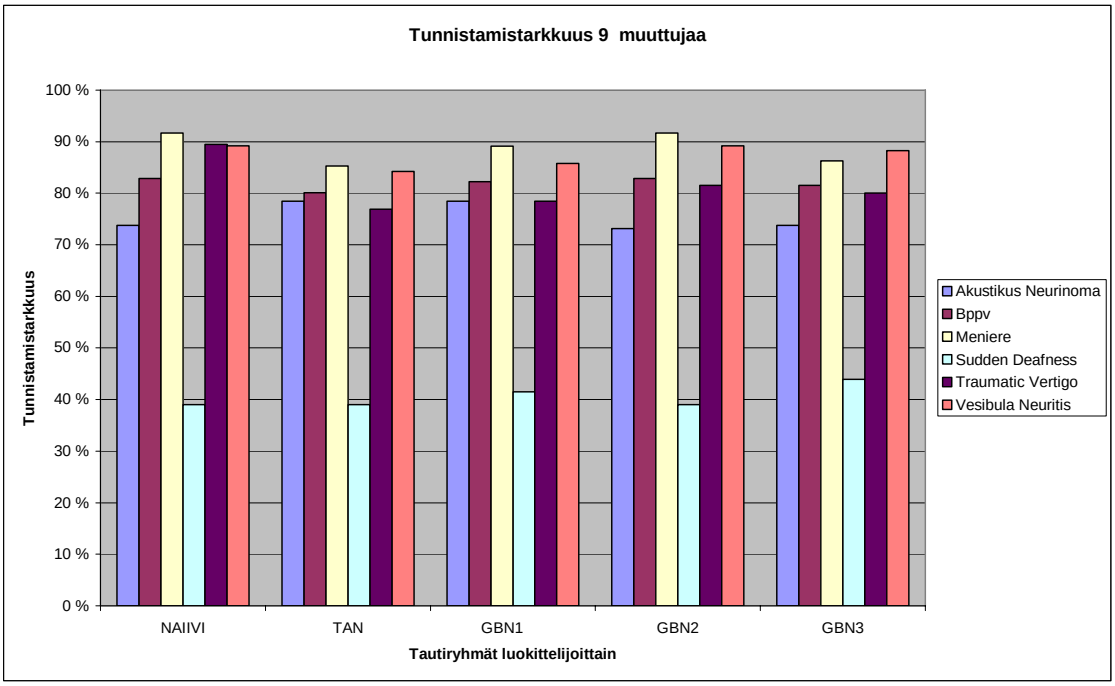
kaikilla luokittelijoilla huonoiten, kun muuttujia on 40. Tunnistamistarkkuudet ovat 85 prosetin molemmin puolin näiden tautiryhmien osalta.



Kuva 21 Tunnistamistarkkuuus 40 muuttujaa - viisi luokittelijaa

Kuvassa 22 on tunnistamistarkkuudet yhdeksän muuttujan luokittelijoilla. Tautiryhmän Sudden Deafness osalta tunnistamistarkkuudet eri luokittelijoila ovat laskeneet huomattavasti, kun lukuja verrataan 40 muuttujan luokittelijoihin. Tunnistarmistarkkuudet tämän tautiryhmän osalta eri luokittelijoila on noin 40 prosenttia, kun 40 muuttujan luokittelijalla luku on 90 prosenttia tai enemmän. Tämä selittyy sillä, että tämän tautiryhmän osalta oireet vaihtelevat varsin paljon eri potilaiden välillä. Kun muuttujamäärää lasketaan näinkin rajusti, niin tässä tapauksessa on varsin odotettavaa, että tunnistamistarkkuus laskee rajusti. Muilla tautiryhmillä tunnistarmistarkkuudet laskivat 40 muuttujan luokittelijoista maltillisemmin. Tautiryhmän Akustikus Neurinoman osalta tunnistamistarkkuus laski keskimäärin kymmenen prosenttia. Muilla tautiryhmillä tunnistamistarkkuus laski neljästä seitsemään prosenttia. Taulukossa 15 on esitelty luvut, jotka kertovat paljonko tunnistamistarkkuus on laskenut eri tautiryhmillä, kun verrataan 40 muuttujan luokittelijoita yhdeksän muuttujan luokittelijoihin. Mielestäni luvut ovat lukuunottamatta tautiryhmää Sudden deaffness hyvät. Maltilliset laskuluvut

vahvistavat sitä tietämystä, että nämä yhdeksän muuttujaa yksinään pystyvät todella hyvin auttamaan huimauspotilaiden diagnosoinnissa.



Kuva 22 Tunnistamistarkkuus - yhdeksän muuttujaa - viisi luokittelijaa

Taulukko 15 Tarkkuuslukujen ero: 40 muuttujaa - yhdeksän muuttujaa

	NAIIVI	TAN	GBN_1	GBN_2	GBN_3
Akustikus Neurinoma	0,116	0,1	0,069	0,123	0,124
Bppv	0,034	0,055	0,055	0,041	0,034
Menière	0,035	0,096	0,058	0,045	0,073
Sudden Deaffness	0,512	0,537	0,512	0,537	0,488
Traumatic Vertigo	0,013	0,062	0,092	0,123	0,062
Vestibular Neuritis	0,008	0,075	0,05	0,025	0,025

Koska viiden muuttujan luokittelijoiden kaikki tarkkuusluvut ovat suunnilleen samat kuin yhdeksän muuttujan luokittelijoiden, niin tässä ei erikseen ole esiteltynä pylväsdiagrammia tunnistamistarkkuuksista.

6 Yhteenveto

Yleinen Bayes -verkko, kun *MDL* -pistemäärää käytettiin verkon rakentamisessa, oli paras 40 muuttujan luokittelijoista. Kun muuttujien määrää laskettiin yhdeksään, niin paras luokittelija oli Naiivi luokittelija, kuten myös viiden muuttujan luokittelijoista. Huomattava on, että yleinen verkko *MDL* -pistemäärällä ilman kaaren kääntöä on sama kuin Naiivi luokittelija. Itse asiassa 40 muuttujan tapauksessa *MDL* -pistemäärällä saadun luokittelijan verkko on ns. Naiivin luokittelijan kaltainen. Yhdeksällä muuttujalla tarkkuusluvut olivat luokittelijoilla todella hyvät, kun verrataan 40 muuttujaan luokittelijoihin, lukuunottamatta tautiryhmää sudden deaffness. Tunnistamistarkkuudet laskivat maksimissaan noin kymmenen prosenttia. Viidellä muuttujan luokittelijoilla kaikki tarkkuusluvut ovat melko samat kuin yhdeksän muuttujan luokittelijoilla.

Yleisesti eri taudit tulivat hyvin tunnistettua. Tautien Akustikus Neurinoma ja Bppv osalta tunnistarkkkudet olivat heikoimmat 40 muuttujan luokittelijoilla. Kun tautia Sudden Deaffness ei oteta lukuun, niin muihin luokittelijoihin verrattuna luvut olivat heikoimmat myös yhdeksän ja viiden muuttujien tapauksessa. Kokonaistarkkuudet ovat kaikilla 40, yhdeksän ja viiden muuttujan luokittelijoilla, kaikissa tautiryhmissä yli 90 prosenttia. Korkeat luvut johtuvat pitkälti suurista oikeiden negatiivisten luvuista.

Yllättävintä mielestäni empiirisissä tuloksissa on se, kuinka hyvin Naiivi luokittelija pärjäsi. Oletin, että *TAN* olisi pärjännyt paremmin. Erot tarkkuusluvuissa eri luokittelijoiden välillä eivät olleet toki suuret, mutta oletin, että *TAN* olisi yleisesti parempi kuin muut verkot. Tällä aineistolla näin ei kuitenkaan käynyt. Mielenkiintoista on se, että 40 muuttujan tapauksessa *MDL*-pistemäärällä saatiin Naiivin kaltainen verkkorakenne. Yhdeksällä ja viidellä muuttujalla verkoista tuli naiiveja, kun kaaren kääntöä ei sallittu.

Kerroin työni teoriaosuudessa priorijakaumien määrittämisestä ja parametrien estimoimiseen liittyvästä problematiikasta. Empiriaosuudessani käyttämästäni ohjelmasta Wekasta johtuen testijoukon priorijakaumien estimaattina käytin Laplace -estimaattia. Olisi mielenkiintoista rakentaa luokittelijoita oikeasti bayesilaisella tavalla. Lisäksi olisi kiinnostava nähdä, että millaisia luokittelijoista tulisi, kun verkon rakentamiseen käytetään rajoiteperustaisia menetelmiä.

Lähteet

1. Winkler, R.H. 1972. Introduction to Bayesian Inference and Decision. USA, *Holt, Rinehart and Winston, Inc.*
2. Tom M. Mitchell. 1997. Machine Learning. The McGraw-Hill Companies,Inc.
3. William E. Pollard. 1986. Bayesian Statistics for Evaluation Research. Sage Pulications, Inc.
4. Sergios Theodoridis & Konstantinos Koutroumbas. 2001. Pattern Recognition And Neural Networks. *Machine Learning and Its Applications 169-195*.
5. Boaz Lerner & Neil D. Lawrence. 2001. A Comparison of State-of-the-Art Classification Techniques with Application to Cytogenetics. *Neural Computing & Applications, no. 10, s.39-47.*
6. K.V.Mardia, J.T.Kent, J.M.Bibby. 1979. Multivariate Analysis. *London Academic Press Inc. Ltd.*
7. Paul.J.Krause. 1998. Learning Probabilistic Networks. *The Knowledge Engineering Review* Volume 13, issue 4.
8. Dimitris Margaritis. 2003. Learning Bayesian Network Model Structure from Data. Partial Fulfillment for the Degree of Doctor of Philosophy. School of Computer Science, Carnegie Mellon University.
9. Eugene Charniak. 1991. Bayesian Network without Tears. *AI magazine, vol.12, no. 4. s.50-63*
10. Nir Friedman & Dan Geiger & Moises Goldszmit. 1997. Bayesian Network Classifiers. *Machine Learning 29, s. 131-163(1997). Kluwer Academic Publisher*
11. Judea Pearl. Graphical Models for Probabilistic and Causal Reasoning. 1997. *UCLA Cognitive Systems Laboratory, Technical Report (R-232-U)*
12. C.K Chow & C.N Liu Approximating Discrete Probability Distributions with Dependence Trees.1968. *IEEE Transactions on Information Theory , vol IT-14, No 3.*

13. Irina Rish. 2000. Advances in Bayesian Learning. *Proceedings of the 2000 International Conference on Artificial Intelligence* (IC-AI'2000), 95–102. Las Vegas, Nevada: CSREA Press.

14. Nir Friedman & Iftach Nahman & Dana Peér. 1999. Learning Bayesian Network Structure from Massive Datasets:"The Sparse Candidate Algorithm". *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 206-215.

15. Ferat Sahin 2000. A Bayesian Network Approach to the Self-organization and Learning in Intelligent Agents.

16. Pentti Huuhtanen & Arto Kalinen 1998. Matemaattinen tilastotiede, Matematiikan, tilastotieteen ja filosofian laitos, Tampereen yliopisto.

17. David Heckerman 1995. A Tutorial on LearningWith Bayesian Networks. *Technical Report MSR-TR-95-06. Microsoft Research, Advanced Technology Division.*

18. Christian P.Robert 1994. The Bayesian Choise, A Decision-Theoretic Motivation. *Springer-Verlag New York, Inc.*

19. Gregory F. Cooper & Edvard Herskovits 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning 9, s. 309-347.*

20. Richard A. Johnson & Dean W. Wichern 1992. *Applied Multivariate Statistical Analysis. USA. Prentice-Hall, Inc.*

21. Martti Juhola & Jorma Laurikkala 2003. Increasing Accuracy of Neural Networks Classification with Principal Component Analysis for Otoneurological and Female Urinary Incontinence Data. *CD-ROM of Medical Informatics Europe 2003, St. Malo, France.*

22. Erna Kentala. 1996. A Neurotologic Expert System for Vertigo and Characteristics of Six Otologic Deseases Involving Vertigo. Department of Otorhinolaryngology, University of Helsinki. Doctoral Thesis.

23. Jorma Laurikkala, Erna Kentala, Martti Juhola, Ilmari Pyykkö, Seppo Lammi. 2000. Usefulness of Imputation for the Analysis of Incomplete Otoneurological Data. *International Journal of Medical Informatics 58-59 s.235-242*

24. Martti Juhola, Kati Viikki, Jorma Laurikkala, Ilmari Pyykkö, Erna Kentala. 2001. On Classification Capability of Neural Networks: A Case Study with Otoneurological Data. *MEDINFO 2001*, *Amsterdam*.

25. Erna Kentala. 1996. Characteristics of Six Otologic Deseases Involving Vertigo. *American Journal of Otology, 17, 883-892*.

26. Kati Viikki, Erna Kentala, Martti Juhola ja Ilmari Pyykkö. 1999. Decision Tree Induction in the Diagnosis of Otoneurological Diseases. *Medical Informatics (1999), vol 24, No. 4, s. 277-289*

27. Kati Viikki, Martti Juhola, Ilmari Pyykkö ja Pekka Honkavaara. 2001. Evaluating Training Data Suitability for Decision Tree Induction. *Journal of Medical Systems, Vol. 25, No. 2*

28. Erna Kentala, Jorma Laurikkala, Ilmari Pyykkö ja Martti Juhola. 1999. Discovering Diagnostic Rules from a Neurotologic Database with Genetic Algorithms. *Annals of Otology, Rhinology & Laryngology, Vol. 108, No. 10*

29. Roderick J.A Little & Donald D. Rubin (1987) *Statistical Analysis With Missing Data. John Wiley & Sons, Inc. USA*

30. Gerhard Arminger, Clofford C. Clogg & Michael E. Sobel. 1995. Handbook of Statistical Modeling for the Social and Behavioral Sciences. P*lenum Press, New York*

31. J.L Schafer. 1997. Analysis of Incomplete Multivariate Data. *Chapman & Hall, London*

32. *http://www.cs.waikato.ac.nz/~ml/weka/*

33. *http://jbnc.sourceforge.net/*

34. Friedman & Koller. 2001. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Kluwer Academic Publishers, Netherlands*

35. S.Tong & D. Koller. 2001. Active Learning for Structure in Bayesian Networks. *Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)* (s. 863-869).

36. Jie Chen & Russell Greiner. 1999. Comparing Bayesian Networks Classifiers. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99), Sweden, Aug 1999.

LIITE 1 MUUTTUJIEN NIMET, ARVOT JA ARVOJEN LUKUMÄÄRÄT

HEAR_BET_R Hearing loss of right ear	No	511
	Yes	304
HEAR_BET_L Hearing loss of left ear	No	443
	Yes	372
Type off hearing loss;sudden?	No	770
	Yes	45
Type of hearing loss;progressive?	No	255
	Yes	560
Type of hearing loss;Both sudden and progressive?	No	814
	Yes	1
TINNITUS Severity of the tinnitus	no handicap	223
	slight handicap	283
	moderate handicap	185
	severe handicap	124
AGE_TIN_SYM When did tinnitus first occur?	no tinnitus	224
	days	33
	1-4 weeks	39
	1-4 months	77
	< 1 year	106
	1-4 years	161
	more	175
EAR_ILLNESS Ear infection(s)?	No	712
	Yes	103
EAR_OPER Any ear operation?	No	762
	Yes	53
INJURY Head or ear trauma, noise injury	No	696
	Yes	119
NOISEEXP Chronic noise exposure	No	713
	Yes	102
HEAD_TRAUMA Trauma of the head	No	752
	Yes	63

Diagnosis	akustikus neurinoma	130
	bppv	146
	meniere	313
	sudden deafness	41
	traumatic vertigo	65
	vestibular neuritis	120
AGE_FIRST Age when the symptoms appeared?	0	29
	1	276
	2	469
	3	41
AGE_SYMPTOMS When did the symptoms occur?	no symptoms	77
	days	23
	1-4 weeks	111
	1-4 months	159
	< 1 year	118
	1-4 years	155
	more	172
ATT_OFTEN How frequent are the spells?	no spells	104
	has come only once	157
	1-2 /year	62
	3-12 /year	83
	1-4 /month	99
	2-7 / weeks	127
	several times in a day	149
	constant dizziness	34
ATT_LAST How long does an attack last?	no attacks	106
	1-15 s	200
	15 s – 5 min	112
	5 min – 4 h	157
	4 –24 h	97
	1-5 days	143
ATT_INTE How severe is the attack?	no attacks	81
	weak	124
	moderate	230
	strong	219
	very strong	161
VER_ROTA rotational	0	305
	1	147
	2	104
	3	259
VER_FLOAT swinging /floating /unsteadiness	0	498
	1	188
	2	52
	3	77
SLIPSFALLS Tumarkin-type drop attacks	no handicap	285
	slight handicap	268
	moderate handicap	218
	severe handicap	44
PROV_POSIT Position induced vertigo	0	325
	1	170
	2	75
	3	245
UNSTEDINESS Unsteadiness outside attack	no	539
	slight handicap	199
	moderate handicap	62
	severe handicap	15
AGE_HL_SYM Duration of hearing symptoms?	no hearing loss	261
	days	11
	1-4 weeks	34
	1-4 months	70
	< 1 year	77
	1-4 years	143
	more	219

GAIN_A Pursuit gain by amplitude %	0	293
	1	404
	2	118
GAIN_LATENCY Pursuit latency in ms	0	54
	1	569
	2	192
CAL_44L 44ø caloric LEFT	0	45
	1	581
	2	189
CAL_44R 44ø caloric RIGHT	0	32
	1	600
	2	183
CAL_ASYM Caloric asymmetry [%]	0	511
	1	180
	2	124
CAL_SP_NYST Spontanic nystagmus	0	575
	1	240
POST_CLOSE Base line, eyes closed [cm/s]	0	506
	1	309
POST_OPEN Base line, eyes open [cm/s]	0	153
	1	662
SP_NYST Spontanic nystagmus	No	571
	Yes	244
EARTRAUMA Trauma of the ear	No	781
	Yes	34

GAIN_LATENCY Pursuit latency in ms	0	54
	1	569
	2	192
AUD_500R audiometry at 500 Hz right, dB	0	645
	1	61
	2	76
	3	33
AUD_500L audiometry at 500 Hz left, dB	0	616
	1	79
	2	77
	3	43
AUD_2000R audiometry at 2 kHz right, dB	0	618
	1	81
	2	75
	3	41
AUD_2000L audiometry at 2 kHz left, dB	0	558
	1	111
	2	88
	3	58
NAUSEA Nausea and/or vomiting	no handicap	269
	slight handicap	204
	moderate handicap	170
	severe handicap	172
HEAR_FLUCT Does the hearing fluctuate?	No	557
	Yes	258
SYM_LIGHTHEAD Lightheadness	No	313
	Yes	502

LIITE2 NAIIVI –LUOKITTELIJA –YHDEKSÄN MUUTTUJAA - TODENNÄKÖISYYSJAKAUMAT

REMID_MD

Akustikus neurinoma	Bppv	Menierin tauti	Sudden Deaffness	Traumatic Vertigo	Vestibular Neuritis
0,16	0,179	0,384	0,05	0,08	0,147

	TINNITUS			
REMID_MD	0(=no handicap)	1(=slight handicap)	2(=moderate handicap)	3(=severe handicap)
Akustikus neurinoma	0,2	0,46	0,277	0,062
Bppv	0,458	0,315	0,144	0,083
Menierin tauti	0,02	0,367	0,319	0,294
Sudden Deaffness	0,268	0,388	0,244	0,1
Traumatic Vertigo	0,398	0,292	0,201	0,109
Vestibular Neuritis	0,722	0,225	0,043	0,01

	AGE_HL_S						
REMID_MD	0(=no hearing loss)	1(=days)	2(=1-4 weeks)	3(=1-4 months)	4(= < 1year)	5(= 1-4 year)	6(more)
Akustikus neurinoma	0,047	0,001	0,032	0,047	0,169	0,421	0,283
Bppv	0,755	0,001	0,001	0,015	0,001	0,076	0,151
Menierin tauti	0,032	0,02	0,036	0,118	0,147	0,204	0,443
Sudden Deaffness	0,004	0,123	0,241	0,455	0,075	0,051	0,051
Traumatic Vertigo	0,411	0,003	0,123	0,078	0,093	0,154	0,139
Vestibular Neuritis	0,884	0,001	0,01	0,01	0,001	0,01	0,084

	AGE_SYMP						
REMID_MD	0(=no symptoms)	1(=days)	2(=1-4 weeks)	3(=1-4 months)	4(= < 1year)	5(= 1-4 years)	6(=more)
Akustikus neurinoma	0,504	0,001	0,009	0,062	0,108	0,0222	0,093
Bppv	0,001	0,022	0,11	0,334	0,198	0,137	0,198
Menierin tauti	0,001	0,029	0,055	0,131	0,144	0,265	0,376
Sudden Deaffness	0,265	0,075	0,217	0,265	0,028	0,075	0,075
Traumatic Vertigo	0,003	0,003	0,139	0,214	0,29	0,244	0,108
Vestibular Neuritis	0,001	0,067	0,488	0,298	0,034	0,034	0,026

	ATT_INTE				
REMID_MD	0(=no attacks)	1(=weak)	2(=moderate)	3(=strong)	4(=very strong)
Akustikus neurinoma	0,529	0,169	0,162	0,07	0,07
Bppv	0,001	0,171	0,512	0,205	0,11
Menierin tauti	0,001	0,144	0,23	0,348	0,278
Sudden Deaffness	0,291	0,267	0,195	0,171	0,076
Traumatic Vertigo	0,003	0,185	0,504	0,23	0,078
Vestibular Neuritis	0,001	0,076	0,175	0,407	0,341

	AGE_TIN						
REMID_MD	0(=no tinnitus)	1(=days)	2(=1-4 weeks)	3(=1-4 months)	4(= < year)	5(=1-4 years)	6(more)
Akustikus neurinoma	0,169	0,001	0,009	0,078	0,199	0,36	0,184
Bppv	0,45	0,11	0,035	0,055	0,055	0,151	0,144
Menierin tauti	0,02	0,032	0,048	0,125	0,179	0,236	0,36
Sudden Deaffness	0,265	0,075	0,241	0,194	0,099	0,028	0,099
Traumatic Vertigo	0,38	0,003	0,063	0,108	0,139	0,199	0,108
Vestibular Neuritis	0,777	0,034	0,034	0,043	0,026	0,034	0,051

	ATT_LAST					
REMID_MD	0(=no attacks)	1(=1-15 s)	2(=15 s - 5min)	3(=5 min-4h)	4(=4-24h)	5(=1-5days)
Akustikus neurinoma	0,719	0,085	0,055	0,085	0,024	0,032
Bppv	0,001	0,586	0,362	0,022	0,001	0,028
Menierin tauti	0,001	0,15	0,109	0,408	0,214	0,118
Sudden Deaffness	0,29	0,385	0,075	0,099	0,123	0,028
Traumatic Vertigo	0,003	0,548	0,199	0,109	0,048	0,093
Vestibular Neuritis	0,001	0,034	0,018	0,034	0,158	0,753

	ATT_OFTE							
REMID_MD	0(=no spells)	1(=has come only once)	2(=1-2/year)	3(=3-12/year)	4(=1-4/month)	5(=2-7/weeks)	6(=several times in a day)	7(=constant dizziness)
Akustikus neurinoma	0,702	0,093	0,077	0,032	0,001	0,032	0,047	0,016
Bppv	0,001	0,008	0,042	0,089	0,089	0,225	0,537	0,008
Menierin tauti	0,001	0,058	0,125	0,188	0,249	0,204	0,112	0,064
Sudden Deaffness	0,287	0,311	0,098	0,028	0,028	0,051	0,169	0,028
Traumatic Vertigo	0,003	0,048	0,018	0,078	0,093	0,349	0,289	0,123
Vestibular Neuritis	0,001	0,908	0,018	0,01	0,01	0,01	0,026	0,018

	HEAD_TRAUMA	
REMID_MD	0(=no)	1(=yes)
Akustikus neurinoma	0,999	0,001
Bppv	0,992	0,008
Menierin tauti	0,99	0,01
Sudden Deaffness	0,996	0,004
Traumatic Vertigo	0,11	0,89
Vestibular Neuritis	0,99	0,01

	SYM_LIGH	
REMID_MD	0(=no)	1(=yes)
Akustikus neurinoma	0,822	0,178
Bppv	0,179	0,821
Menierin tauti	0,272	0,728
Sudden Deaffness	0,56	0,44
Traumatic Vertigo	0,293	0,707
Vestibular Neuritis	0,442	0,558

LIITE 3 TAN – YHDEKSÄN MUUTTUJAA – TODENNÄKÖISYYSJAKAUMAT

		HEAD_TRAUMA	
REMID_MD	AGE_TIN	0(=no)	1(=yes)
Akustikus neurinoma	0(=no tinnitus)	0,9925	0,0075
Akustikus neurinoma	1(=days)	0,5000	0,5000
Akustikus neurinoma	2(=1-4 weeks)	0,8750	0,1250
Akustikus neurinoma	3(=1-4 months)	0,9839	0,0161
Akustikus neurinoma	4(= < year)	0,9937	0,0063
Akustikus neurinoma	5(=1-4 years)	0,9965	0,0035
Akustikus neurinoma	6(more)	0,9932	0,0068
Bppv	0(=no tinnitus)	0,9975	0,0025
Bppv	1(=days)	0,9898	0,0102
Bppv	2(=1-4 weeks)	0,9688	0,0312
Bppv	3(=1-4 months)	0,9800	0,0200
Bppv	4(= < year)	0,8600	0,1400
Bppv	5(=1-4 years)	0,9925	0,0075
Bppv	6(more)	0,9922	0,0078
Menierin tauti	0(=no tinnitus)	0,8158	0,1842
Menierin tauti	1(=days)	0,9839	0,0161
Menierin tauti	2(=1-4 weeks)	0,9891	0,0109
Menierin tauti	3(=1-4 months)	0,9958	0,0042
Menierin tauti	4(= < year)	0,9615	0,0385
Menierin tauti	5(=1-4 years)	0,9978	0,0022
Menierin tauti	6(more)	0,9985	0,0015
Sudden Deaffness	0(=no tinnitus)	0,9853	0,0147
Sudden Deaffness	1(=days)	0,9500	0,0500
Sudden Deaffness	2(=1-4 weeks)	0,9839	0,0161
Sudden Deaffness	3(=1-4 months)	0,9800	0,0200
Sudden Deaffness	4(= < year)	0,9615	0,0385
Sudden Deaffness	5(=1-4 years)	0,8750	0,1250
Sudden Deaffness	6(more)	0,9615	0,0385
Traumatic Vertigo	0(=no tinnitus)	0,1645	0,8355
Traumatic Vertigo	1(=days)	0,5000	0,5000
Traumatic Vertigo	2(=1-4 weeks)	0,0385	0,9615
Traumatic Vertigo	3(=1-4 months)	0,1591	0,8409
Traumatic Vertigo	4(= < year)	0,2321	0,7679
Traumatic Vertigo	5(=1-4 years)	0,0125	0,9875
Traumatic Vertigo	6(more)	0,0227	0,9773
Vestibular Neuritis	0(=no tinnitus)	0,9876	0,0124
Vestibular Neuritis	1(=days)	0,9615	0,0385
Vestibular Neuritis	2(=1-4 weeks)	0,9615	0,0385
Vestibular Neuritis	3(=1-4 months)	0,9688	0,0312
Vestibular Neuritis	4(= < year)	0,9500	0,0500
Vestibular Neuritis	5(=1-4 years)	0,9615	0,0385
Vestibular Neuritis	6(more)	0,9737	0,0263

	TINNITUS			
REMID_MD	0(=no handicap)	1(=slight handicap)	2(=moderate handicap)	3(=severe handicap)
Akustikus neurinoma	0,2003	0,4605	0,2768	0,0625
Bppv	0,4580	0,3148	0,1443	0,0830
Menierin tauti	0,0197	0,3672	0,3193	0,2938
Sudden Deaffness	0,2680	0,3880	0,2440	0,1000
Traumatic Vertigo	0,3985	0,2919	0,2005	0,1091
Vestibular Neuritis	0,7224	0,2251	0,0428	0,0097

		SYM_LIGH	
REMID_MD	TINNITUS	0(=no)	1(=yes)
Akustikus neurinoma	0(=no handicap)	0,8418	0,1582
Akustikus neurinoma	1(=slight handicap)	0,7818	0,2182
Akustikus neurinoma	2(=moderate handicap)	0,8853	0,1147
Akustikus neurinoma	3(=severe handicap)	0,7400	0,2600
Bppv	0(=no handicap)	0,2252	0,7748
Bppv	1(=slight handicap)	0,1763	0,8237
Bppv	2(=moderate handicap)	0,1016	0,8984
Bppv	3(=severe handicap)	0,0946	0,9054
Menierin tauti	0(=no handicap)	0,3421	0,6579
Menierin tauti	1(=slight handicap)	0,3916	0,6084
Menierin tauti	2(=moderate handicap)	0,2708	0,7292
Menierin tauti	3(=severe handicap)	0,1209	0,8791
Sudden Deaffness	0(=no handicap)	0,8971	0,1029
Sudden Deaffness	1(=slight handicap)	0,4388	0,5612
Sudden Deaffness	2(=moderate handicap)	0,4032	0,5968
Sudden Deaffness	3(=severe handicap)	0,5000	0,5000
Traumatic Vertigo	0(=no handicap)	0,5000	0,5000
Traumatic Vertigo	1(=slight handicap)	0,1121	0,8879
Traumatic Vertigo	2(=moderate handicap)	0,3125	0,6875
Traumatic Vertigo	3(=severe handicap)	0,0227	0,9773
Vestibular Neuritis	0(=no handicap)	0,4599	0,5401
Vestibular Neuritis	1(=slight handicap)	0,4085	0,5915
Vestibular Neuritis	2(=moderate handicap)	0,2187	0,7813
Vestibular Neuritis	3(=severe handicap)	0,8750	0,1250

		AGE_HL_S						
REMID_MD	AGE_TIN	0(=no hearing loss)	1(=days)	2(=1-4 weeks)	3(=1-4 months)	4(= < 1year)	5(= 1-4 year)	6(more)
Akustikus neurinoma	0(=no tinnitus)	0,0504	0,0072	0,0072	0,0072	0,1799	0,4388	0,3094
Akustikus neurinoma	1(=days)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Akustikus neurinoma	2(=1-4 weeks)	0,0769	0,0769	0,5385	0,0769	0,0769	0,0769	0,0769
Akustikus neurinoma	3(=1-4 months)	0,1045	0,0149	0,0149	0,4627	0,1940	0,1045	0,1045
Akustikus neurinoma	4(= < year)	0,0061	0,0061	0,0061	0,0429	0,4110	0,3374	0,1902
Akustikus neurinoma	5(=1-4 years)	0,0657	0,0035	0,0657	0,0035	0,0450	0,7301	0,0865
Akustikus neurinoma	6(more)	0,0464	0,0066	0,0066	0,0066	0,1258	0,0066	0,8013
Bppv	0(=no tinnitus)	0,8660	0,0025	0,0025	0,0174	0,0025	0,0025	0,1067
Bppv	1(=days)	0,7087	0,0097	0,0097	0,0097	0,0097	0,1262	0,1262
Bppv	2(=1-4 weeks)	0,6757	0,0270	0,0270	0,0270	0,0270	0,0270	0,1892
Bppv	3(=1-4 months)	0,7818	0,0182	0,0182	0,0182	0,0182	0,0182	0,1273
Bppv	4(= < year)	0,6727	0,0182	0,0182	0,0182	0,0182	0,1273	0,1273
Bppv	5(=1-4 years)	0,5252	0,0072	0,0072	0,0504	0,0072	0,2230	0,1799
Bppv	6(more)	0,5489	0,0075	0,0075	0,0075	0,0075	0,1429	0,2782
Menierin tauti	0(=no tinnitus)	0,1628	0,0233	0,0233	0,0233	0,0233	0,1628	0,5814
Menierin tauti	1(=days)	0,1045	0,2836	0,1940	0,0149	0,1940	0,1940	0,0149
Menierin tauti	2(=1-4 weeks)	0,0103	0,0103	0,4433	0,1340	0,0722	0,0103	0,3196
Menierin tauti	3(=1-4 months)	0,0788	0,0290	0,0041	0,7261	0,0539	0,0041	0,1037
Menierin tauti	4(= < year)	0,0029	0,0204	0,0204	0,0029	0,5802	0,1603	0,2128
Menierin tauti	5(=1-4 years)	0,0554	0,0155	0,0022	0,0554	0,0820	0,5876	0,2018
Menierin tauti	6(more)	0,0102	0,0015	0,0102	0,0190	0,0190	0,0715	0,8686
Sudden Deaffness	0(=no tinnitus)	0,0137	0,0137	0,2603	0,5890	0,0137	0,0959	0,0137
Sudden Deaffness	1(=days)	0,0400	0,7600	0,0400	0,0400	0,0400	0,0400	0,0400
Sudden Deaffness	2(=1-4 weeks)	0,0149	0,1045	0,4627	0,3731	0,0149	0,0149	0,0149
Sudden Deaffness	3(=1-4 months)	0,0182	0,0182	0,0182	0,8909	0,0182	0,0182	0,0182
Sudden Deaffness	4(= < year)	0,0323	0,0323	0,0323	0,0323	0,6129	0,0323	0,2258
Sudden Deaffness	5(=1-4 years)	0,0769	0,0769	0,0769	0,0769	0,0769	0,5385	0,0769
Sudden Deaffness	6(more)	0,0323	0,2258	0,4194	0,0323	0,0323	0,0323	0,2258
Traumatic Vertigo	0(=no tinnitus)	0,5796	0,0064	0,1592	0,0828	0,0446	0,0828	0,0446
Traumatic Vertigo	1(=days)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Traumatic Vertigo	2(=1-4 weeks)	0,2258	0,0323	0,2258	0,0323	0,0323	0,2258	0,2258
Traumatic Vertigo	3(=1-4 months)	0,5102	0,0204	0,0204	0,3878	0,0204	0,0204	0,0204
Traumatic Vertigo	4(= < year)	0,2131	0,0164	0,1148	0,0164	0,5082	0,1148	0,0164
Traumatic Vertigo	5(=1-4 years)	0,2941	0,0118	0,0118	0,0118	0,0118	0,4353	0,2235
Traumatic Vertigo	6(more)	0,1429	0,0204	0,2653	0,0204	0,0204	0,0204	0,5102
Vestibular Neuritis	0(=no tinnitus)	0,9475	0,0018	0,0018	0,0018	0,0018	0,0018	0,0438
Vestibular Neuritis	1(=days)	0,6129	0,0323	0,0323	0,0323	0,0323	0,0323	0,2258
Vestibular Neuritis	2(=1-4 weeks)	0,4194	0,0323	0,2258	0,0323	0,0323	0,0323	0,2258
Vestibular Neuritis	3(=1-4 months)	0,6757	0,0270	0,0270	0,1892	0,0270	0,0270	0,0270
Vestibular Neuritis	4(= < year)	0,5200	0,0400	0,0400	0,0400	0,0400	0,0400	0,2800
Vestibular Neuritis	5(=1-4 years)	0,6129	0,0323	0,0323	0,0323	0,0323	0,2258	0,0323
Vestibular Neuritis	6(more)	0,4419	0,0233	0,0233	0,0233	0,0233	0,0233	0,4419

		AGE_TIN						
REMID_MD	TINNITUS	0(=no tinnitus)	1(=days)	2(=1-4 weeks)	3(=1-4 months)	4(= < year)	5(=1-4 years)	6(more)
Akustikus neurinoma	0(=no handicap)	0,8160	0,0061	0,0061	0,0429	0,0061	0,0798	0,0429
Akustikus neurinoma	1(=slight handicap)	0,0027	0,0027	0,0191	0,1008	0,2643	0,3951	0,2153
Akustikus neurinoma	2(=moderate handicap)	0,0045	0,0045	0,0045	0,0583	0,1928	0,4888	0,2466
Akustikus neurinoma	3(=severe handicap)	0,0182	0,0182	0,0182	0,1273	0,3455	0,3455	0,1273
Bppv	0(=no handicap)	0,9707	0,0171	0,0024	0,0024	0,0024	0,0024	0,0024
Bppv	1(=slight handicap)	0,0035	0,3004	0,1095	0,1095	0,0883	0,2367	0,1519
Bppv	2(=moderate handicap)	0,0075	0,0526	0,0075	0,0977	0,1880	0,2782	0,3684
Bppv	3(=severe handicap)	0,0127	0,0127	0,0127	0,0886	0,0127	0,3924	0,4684
Menierin tauti	0(=no handicap)	0,7209	0,0233	0,0233	0,0233	0,0233	0,0233	0,1628
Menierin tauti	1(=slight handicap)	0,0100	0,0617	0,0789	0,1736	0,1736	0,2769	0,2253
Menierin tauti	2(=moderate handicap)	0,0016	0,0214	0,0511	0,1104	0,2488	0,2685	0,2982
Menierin tauti	3(=severe handicap)	0,0018	0,0125	0,0125	0,0877	0,1199	0,1628	0,6029
Sudden Deaffness	0(=no handicap)	0,9178	0,0137	0,0137	0,0137	0,0137	0,0137	0,0137
Sudden Deaffness	1(=slight handicap)	0,0097	0,1262	0,2427	0,3010	0,1845	0,0097	0,1262
Sudden Deaffness	2(=moderate handicap)	0,0149	0,1045	0,3731	0,2836	0,1045	0,1045	0,0149
Sudden Deaffness	3(=severe handicap)	0,0323	0,0323	0,4194	0,0323	0,0323	0,0323	0,4194
Traumatic Vertigo	0(=no handicap)	0,9264	0,0061	0,0061	0,0429	0,0061	0,0061	0,0061
Traumatic Vertigo	1(=slight handicap)	0,0083	0,0083	0,1074	0,2066	0,2562	0,2562	0,1570
Traumatic Vertigo	2(=moderate handicap)	0,0118	0,0118	0,1529	0,0118	0,1529	0,4353	0,2235
Traumatic Vertigo	3(=severe handicap)	0,0204	0,0204	0,0204	0,2653	0,2653	0,2653	0,1429
Vestibular Neuritis	0(=no handicap)	0,9887	0,0019	0,0019	0,0019	0,0019	0,0019	0,0019
Vestibular Neuritis	1(=slight handicap)	0,2544	0,1124	0,1479	0,1479	0,0414	0,1479	0,1479
Vestibular Neuritis	2(=moderate handicap)	0,0270	0,1892	0,0270	0,1892	0,3514	0,0270	0,1892
Vestibular Neuritis	3(=severe handicap)	0,0769	0,0769	0,0769	0,0769	0,0769	0,0769	0,5385

		AGE_SYMP						
REMID_MD	AGE_HL_S	0(=no symptoms)	1(=days)	2(=1-4 weeks)	3(=1-4 months)	4(= < 1year)	5(= 1-4 years)	6(=more)
Akustikus neurinoma	0(=no hearing loss)	0,4419	0,0233	0,0233	0,3023	0,1628	0,0233	0,0233
Akustikus neurinoma	1(=days)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Akustikus neurinoma	2(=1-4 weeks)	0,2258	0,0323	0,2258	0,0323	0,4194	0,0323	0,0323
Akustikus neurinoma	3(=1-4 months)	0,4419	0,0233	0,0233	0,4419	0,0233	0,0233	0,0233
Akustikus neurinoma	4(= < 1year)	0,6978	0,0072	0,0072	0,0504	0,0935	0,0504	0,0935
Akustikus neurinoma	5(= 1-4 year)	0,3591	0,0030	0,0030	0,0386	0,1098	0,3591	0,1276
Akustikus neurinoma	6(more)	0,6070	0,0044	0,0044	0,0044	0,0830	0,2140	0,0830
Bppv	0(=no hearing loss)	0,0015	0,0282	0,1352	0,3492	0,1798	0,1263	0,1798
Bppv	1(=days)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Bppv	2(=1-4 weeks)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Bppv	3(=1-4 months)	0,0526	0,0526	0,0526	0,3684	0,3684	0,0526	0,0526
Bppv	4(= < 1year)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Bppv	5(= 1-4 year)	0,0137	0,0137	0,0137	0,2603	0,2603	0,3425	0,0959
Bppv	6(more)	0,0072	0,0072	0,0504	0,2662	0,2230	0,0935	0,3525
Menierin tauti	0(=no hearing loss)	0,0149	0,1045	0,1045	0,3731	0,0149	0,3731	0,0149
Menierin tauti	1(=days)	0,0233	0,3023	0,1628	0,1628	0,0233	0,1628	0,1628
Menierin tauti	2(=1-4 weeks)	0,0137	0,0959	0,4247	0,1781	0,0959	0,0959	0,0959
Menierin tauti	3(=1-4 months)	0,0044	0,0830	0,0830	0,5022	0,1354	0,1354	0,0568
Menierin tauti	4(= < 1year)	0,0035	0,0035	0,0035	0,0247	0,5548	0,3216	0,0883
Menierin tauti	5(= 1-4 year)	0,0026	0,0179	0,0026	0,0793	0,1253	0,5243	0,2481
Menierin tauti	6(more)	0,0012	0,0083	0,0511	0,0654	0,0369	0,1653	0,6718
Sudden Deaffness	0(=no hearing loss)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Sudden Deaffness	1(=days)	0,3514	0,1892	0,1892	0,0270	0,0270	0,1892	0,0270
Sudden Deaffness	2(=1-4 weeks)	0,1045	0,0149	0,6418	0,1045	0,0149	0,0149	0,1045
Sudden Deaffness	3(=1-4 months)	0,2562	0,1074	0,0579	0,4545	0,0083	0,0579	0,0579
Sudden Deaffness	4(= < 1year)	0,2800	0,0400	0,0400	0,0400	0,2800	0,2800	0,0400
Sudden Deaffness	5(= 1-4 year)	0,6842	0,0526	0,0526	0,0526	0,0526	0,0526	0,0526
Sudden Deaffness	6(more)	0,0526	0,0526	0,0526	0,3684	0,0526	0,0526	0,3684
Traumatic Vertigo	0(=no hearing loss)	0,0059	0,0059	0,1124	0,2189	0,2899	0,3609	0,0059
Traumatic Vertigo	1(=days)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Traumatic Vertigo	2(=1-4 weeks)	0,0182	0,0182	0,5636	0,0182	0,1273	0,1273	0,1273
Traumatic Vertigo	3(=1-4 months)	0,0270	0,0270	0,0270	0,8378	0,0270	0,0270	0,0270
Traumatic Vertigo	4(= < 1year)	0,0233	0,0233	0,1628	0,1628	0,5814	0,0233	0,0233
Traumatic Vertigo	5(= 1-4 year)	0,0149	0,0149	0,0149	0,1045	0,2836	0,2836	0,2836
Traumatic Vertigo	6(more)	0,0164	0,0164	0,0164	0,1148	0,3115	0,2131	0,3115
Vestibular Neuritis	0(=no hearing loss)	0,0015	0,0570	0,5008	0,2974	0,0940	0,0293	0,0200
Vestibular Neuritis	1(=days)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Vestibular Neuritis	2(=1-4 weeks)	0,0769	0,0769	0,5385	0,0769	0,0769	0,0769	0,0769
Vestibular Neuritis	3(=1-4 months)	0,0769	0,0769	0,0769	0,5385	0,0769	0,0769	0,0769
Vestibular Neuritis	4(= < 1year)	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429	0,1429
Vestibular Neuritis	5(= 1-4 year)	0,0769	0,0769	0,5385	0,0769	0,0769	0,0769	0,0769
Vestibular Neuritis	6(more)	0,0149	0,1940	0,2836	0,2836	0,0149	0,1045	0,1045

		ATT_INTE				
REMID_MD	AGE_SYM	0(=no attacks)	1(=weak)	2(=moderate)	3(=strong)	4(=very strong)
Akustikus neurinoma	0	0,9900	0,0025	0,0025	0,0025	0,0025
Akustikus neurinoma	1	0,2000	0,2000	0,2000	0,2000	0,2000
Akustikus neurinoma	2	0,0909	0,0909	0,6364	0,0909	0,0909
Akustikus neurinoma	3	0,1321	0,4717	0,2453	0,0189	0,1321
Akustikus neurinoma	4	0,0112	0,4831	0,2135	0,1461	0,1461
Akustikus neurinoma	5	0,0391	0,2737	0,3743	0,1397	0,1732
Akustikus neurinoma	6	0,0909	0,2468	0,3247	0,2468	0,0909
Bppv	0	0,2000	0,2000	0,2000	0,2000	0,2000
Bppv	1	0,0435	0,0435	0,0435	0,8261	0,0435
Bppv	2	0,0099	0,1287	0,3069	0,3069	0,2475
Bppv	3	0,0033	0,2441	0,5251	0,1639	0,0635
Bppv	4	0,0056	0,2067	0,6089	0,1732	0,0056
Bppv	5	0,0080	0,0560	0,7760	0,1520	0,0080
Bppv	6	0,0056	0,1397	0,3408	0,2067	0,3073
Menierin tauti	0	0,2000	0,2000	0,2000	0,2000	0,2000
Menierin tauti	1	0,0169	0,1186	0,3220	0,4237	0,1186
Menierin tauti	2	0,0093	0,1215	0,4019	0,3458	0,1215
Menierin tauti	3	0,0040	0,2191	0,2191	0,3147	0,2430
Menierin tauti	4	0,0036	0,2000	0,2655	0,2655	0,2655
Menierin tauti	5	0,0020	0,1690	0,2406	0,3479	0,2406
Menierin tauti	6	0,0014	0,0856	0,1781	0,3801	0,3548
Sudden Deaffness	0	0,9437	0,0141	0,0141	0,0141	0,0141
Sudden Deaffness	1	0,0435	0,3043	0,3043	0,0435	0,3043
Sudden Deaffness	2	0,0169	0,4237	0,2203	0,3220	0,0169
Sudden Deaffness	3	0,0141	0,3521	0,3521	0,2676	0,0141
Sudden Deaffness	4	0,0909	0,0909	0,0909	0,0909	0,6364
Sudden Deaffness	5	0,0435	0,0435	0,3043	0,3043	0,3043
Sudden Deaffness	6	0,3043	0,5652	0,0435	0,0435	0,0435
Traumatic Vertigo	0	0,2000	0,2000	0,2000	0,2000	0,2000
Traumatic Vertigo	1	0,2000	0,2000	0,2000	0,2000	0,2000
Traumatic Vertigo	2	0,0169	0,3220	0,3220	0,3220	0,0169
Traumatic Vertigo	3	0,0112	0,2135	0,4157	0,2135	0,1461
Traumatic Vertigo	4	0,0084	0,1092	0,5630	0,2101	0,1092
Traumatic Vertigo	5	0,0099	0,2475	0,5446	0,1287	0,0693
Traumatic Vertigo	6	0,0213	0,0213	0,5319	0,4043	0,0213
Vestibular Neuritis	0	0,2000	0,2000	0,2000	0,2000	0,2000
Vestibular Neuritis	1	0,0189	0,0189	0,1321	0,2453	0,5849
Vestibular Neuritis	2	0,0028	0,0529	0,1365	0,4540	0,3538
Vestibular Neuritis	3	0,0045	0,1403	0,2489	0,3303	0,2760
Vestibular Neuritis	4	0,0154	0,1077	0,2000	0,4769	0,2000
Vestibular Neuritis	5	0,0345	0,0345	0,0345	0,4483	0,4483
Vestibular Neuritis	6	0,0435	0,0435	0,3043	0,3043	0,3043

		ATT_LAST					
REMID_MD	ATT_INTE	0(=no attacks)	1(=1-15 s)	2(=15 s - 5min)	3(=5 min-4h)	4(=4-24h)	5(=1-5days)
Akustikus neurinoma	0(=no attacks)	0,9738	0,0024	0,0024	0,0167	0,0024	0,0024
Akustikus neurinoma	1(=weak)	0,3986	0,3116	0,1377	0,1377	0,0072	0,0072
Akustikus neurinoma	2(=moderate)	0,4167	0,1894	0,1439	0,1894	0,0076	0,0530
Akustikus neurinoma	3(=strong)	0,3167	0,0167	0,1167	0,0167	0,3167	0,2167
Akustikus neurinoma	4(=very strong)	0,5167	0,0167	0,0167	0,3167	0,0167	0,1167
Bppv	0(=no attacks)	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667
Bppv	1(=weak)	0,0064	0,7756	0,1987	0,0064	0,0064	0,0064
Bppv	2(=moderate)	0,0022	0,6338	0,3311	0,0154	0,0022	0,0154
Bppv	3(=strong)	0,0054	0,3925	0,5215	0,0376	0,0054	0,0376
Bppv	4(=very strong)	0,0098	0,3627	0,4216	0,0686	0,0098	0,1275
Menierin tauti	0(=no attacks)	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667
Menierin tauti	1(=weak)	0,0036	0,5254	0,1558	0,2210	0,0254	0,0688
Menierin tauti	2(=moderate)	0,0023	0,2078	0,2489	0,3721	0,0708	0,0982
Menierin tauti	3(=strong)	0,0015	0,0379	0,0833	0,4742	0,2924	0,1106
Menierin tauti	4(=very strong)	0,0019	0,0473	0,0019	0,4451	0,3314	0,1723
Sudden Deaffness	0(=no attacks)	0,9359	0,0128	0,0128	0,0128	0,0128	0,0128
Sudden Deaffness	1(=weak)	0,0139	0,6806	0,1806	0,0972	0,0139	0,0139
Sudden Deaffness	2(=moderate)	0,0185	0,7963	0,0185	0,0185	0,0185	0,1296
Sudden Deaffness	3(=strong)	0,0208	0,1458	0,1458	0,3958	0,2708	0,0208
Sudden Deaffness	4(=very strong)	0,0417	0,0417	0,0417	0,0417	0,7917	0,0417
Traumatic Vertigo	0(=no attacks)	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667
Traumatic Vertigo	1(=weak)	0,0128	0,6282	0,1667	0,0897	0,0128	0,0897
Traumatic Vertigo	2(=moderate)	0,0049	0,5637	0,2108	0,1225	0,0343	0,0637
Traumatic Vertigo	3(=strong)	0,0104	0,5729	0,1979	0,0729	0,0729	0,0729
Traumatic Vertigo	4(=very strong)	0,0278	0,0278	0,1944	0,1944	0,1944	0,3611
Vestibular Neuritis	0(=no attacks)	0,1667	0,1667	0,1667	0,1667	0,1667	0,1667
Vestibular Neuritis	1(=weak)	0,0167	0,2167	0,0167	0,0167	0,1167	0,6167
Vestibular Neuritis	2(=moderate)	0,0076	0,0985	0,0076	0,0985	0,0530	0,7348
Vestibular Neuritis	3(=strong)	0,0033	0,0033	0,0233	0,0433	0,0833	0,8433
Vestibular Neuritis	4(=very strong)	0,0040	0,0040	0,0278	0,0040	0,3135	0,6468

		ATT_OFTE							
REMID_MD	ATT_LAST	0(=no spells)	1(=has come only once	2(=1-2/year)	3(=3-12/year)	4(=1-4/month)	5(=2-7/weeks)	6(=several times in a day)	7(=constant dizziness)
Akustikus neurinoma	0(=no attacks)	0,9458	0,0122	0,0017	0,0017	0,0017	0,0017	0,0122	0,0227
Akustikus neurinoma	1(=1-15 s)	0,0135	0,2568	0,0135	0,0135	0,0135	0,3378	0,3378	0,0135
Akustikus neurinoma	2(=15 s - 5min)	0,1400	0,1400	0,3800	0,1400	0,0200	0,0200	0,1400	0,0200
Akustikus neurinoma	3(=5 min-4h)	0,0946	0,3378	0,3378	0,1757	0,0135	0,0135	0,0135	0,0135
Akustikus neurinoma	4(=4-24h)	0,0385	0,0385	0,5000	0,2692	0,0385	0,0385	0,0385	0,0385
Akustikus neurinoma	5(=1-5days)	0,0312	0,5938	0,2188	0,0312	0,0312	0,0312	0,0312	0,0312
Bppv	0(=no attacks)	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Bppv	1(=1-15 s)	0,0019	0,0019	0,0134	0,0821	0,0477	0,2080	0,6317	0,0134
Bppv	2(=15 s - 5min)	0,0031	0,0031	0,0767	0,0767	0,1319	0,2791	0,4264	0,0031
Bppv	3(=5 min-4h)	0,0385	0,0385	0,0385	0,0385	0,5000	0,0385	0,2692	0,0385
Bppv	4(=4-24h)	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Bppv	5(=1-5days)	0,0312	0,2188	0,2188	0,4063	0,0312	0,0312	0,0312	0,0312
Menierin tauti	0(=no attacks)	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Menierin tauti	1(=1-15 s)	0,0034	0,0034	0,0862	0,1483	0,1690	0,2724	0,2931	0,0241
Menierin tauti	2(=15 s - 5min)	0,0047	0,0047	0,0613	0,1179	0,2594	0,2311	0,2877	0,0330
Menierin tauti	3(=5 min-4h)	0,0013	0,0554	0,1095	0,1869	0,2564	0,2487	0,0631	0,0786
Menierin tauti	4(=4-24h)	0,0024	0,0463	0,1195	0,2805	0,2366	0,1634	0,0463	0,1049
Menierin tauti	5(=1-5days)	0,0043	0,2130	0,2913	0,1348	0,3174	0,0043	0,0043	0,0304
Sudden Deaffness	0(=no attacks)	0,9125	0,0125	0,0125	0,0125	0,0125	0,0125	0,0125	0,0125
Sudden Deaffness	1(=1-15 s)	0,0096	0,4712	0,0096	0,0096	0,0096	0,0673	0,3558	0,0673
Sudden Deaffness	2(=15 s - 5min)	0,0385	0,2692	0,0385	0,0385	0,0385	0,2692	0,2692	0,0385
Sudden Deaffness	3(=5 min-4h)	0,0312	0,4063	0,2188	0,2188	0,0312	0,0312	0,0312	0,0312
Sudden Deaffness	4(=4-24h)	0,0263	0,3421	0,3421	0,0263	0,1842	0,0263	0,0263	0,0263
Sudden Deaffness	5(=1-5days)	0,0714	0,0714	0,5000	0,0714	0,0714	0,0714	0,0714	0,0714
Traumatic Vertigo	0(=no attacks)	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Traumatic Vertigo	1(=1-15 s)	0,0045	0,0312	0,0045	0,0580	0,0045	0,3527	0,3795	0,1652
Traumatic Vertigo	2(=15 s - 5min)	0,0116	0,0116	0,0116	0,1512	0,0814	0,5000	0,1512	0,0814
Traumatic Vertigo	3(=5 min-4h)	0,0200	0,0200	0,0200	0,0200	0,2600	0,3800	0,2600	0,0200
Traumatic Vertigo	4(=4-24h)	0,0385	0,0385	0,2692	0,2692	0,0385	0,0385	0,2692	0,0385
Traumatic Vertigo	5(=1-5days)	0,0227	0,2955	0,0227	0,0227	0,4318	0,0227	0,0227	0,1591
Vestibular Neuritis	0(=no attacks)	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Vestibular Neuritis	1(=1-15 s)	0,0312	0,2188	0,0312	0,0312	0,2188	0,0312	0,4063	0,0312
Vestibular Neuritis	2(=15 s - 5min)	0,0500	0,6500	0,0500	0,0500	0,0500	0,0500	0,0500	0,0500
Vestibular Neuritis	3(=5 min-4h)	0,0312	0,5938	0,0312	0,0312	0,0312	0,2188	0,0312	0,0312
Vestibular Neuritis	4(=4-24h)	0,0082	0,8443	0,0574	0,0082	0,0082	0,0082	0,0082	0,0574
Vestibular Neuritis	5(=1-5days)	0,0018	0,9440	0,0126	0,0126	0,0018	0,0018	0,0126	0,0126