Luento 4

- Riippuvuussäännöt (tarkemmin)
- Riippuvuuksien haku numeerisesta datasta
- Ekskursiona Bayes-verkot

Riippuvuussäännöt

- Muuttuja- ja arvoperustainen tulkinta
- Riippuvuuden vahvuuden ja merkityksellisyyden mittaus
- Säännön redundanssi
- Tilastollisen riippuvuuden epämonotonisuus ja Simpsonin paradoksi
- Ei-binääristen muuttujien käsittely
- Varoitus assosiaatiosäännöistä

Muuttuja- vs. arvoperustainen tulkinta

Sääntö $X \to A$. (Merk. myös vastaavia muuttujia X:llä ja A:lla.)

1. Muuttujaperustainen (variable-based) tulkinta: **binäärimuuttujien** X ja A välillä on vahva positiivinen riippuvuus.
   - Sama asia kuin $\neg X \to \neg A$
   - Yhtä hyviä kuin negatiivinen riippuvuus X:n ja $\neg A$:n välillä (tai $\neg X$:n ja A:n)
2. Arvoperustainen (value-based): **arvokombinaatioiden** $X = 1$ ja $A = 1$ välillä on vahva positiivinen riippuvuus.
   - eri asia kuin $\neg X \to \neg A$, joka voi olla heikko!

Riippuvuuden vahvuus: Tavallisimmat mitat

Olk. $X \to A$

1. Muuttujaperustainen: Leverage

$$\delta(X, A) = P(XA) - P(X)P(A)$$

2. Arvoperustainen: Lift

$$\gamma(X, A) = \frac{P(XA)}{P(X)P(A)} = \frac{P(A|X)}{P(A)}$$

$P(A|X)$ on säännön “konfidenssi”

Huom! Kontingenssitaulu

	A	$\neg A$	kaikki
X	$fr(XA) = n[P(X)P(A) + \delta]$	$fr(X\neg A) = n[P(X)P(\neg A) - \delta]$	$fr(X)$
$\neg X$	$fr(\neg XA) = n[P(\neg X)P(A) - \delta]$	$fr(\neg X\neg A) = n[P\neg(X)P(\neg A) + \delta]$	$fr(\neg X)$
kaikki	$fr(A)$	$fr(\neg A)$	n

Kaikkien arvokombinaatioiden välillä siis sama $|\delta|$!
$\Leftrightarrow$ γ riippuu arvokombinaatiosta

Esim: Milloin arvoperustaisesta voisi olla iloa?

Olk. X=alleelikombinaatio, D=sairaus ja $P(X)$ pieni.

$P(D|X) = 1.0$ eli $\gamma(X, D) = P(D)^{-1}$ voi olla iso
$P(\neg D|X) = 0$ eli $\gamma(X, \neg D) = 0$
$P(D|\neg X) \approx P(D)$ eli $\gamma(\neg X, D) \approx 1$
$P(\neg D|\neg X) \approx P(\neg D)$ eli $\gamma(\neg X, \neg D) \approx 1$

$\leftrightarrow$ $\delta(X, D) = P(X)P(\neg D)$ pieni.

Ts. riippuvuus olisi vahva arvoperustaisessa tulkinnassa, mutta heikko muuttujaperustaisessa.

(Yleensä **muuttuja**perustaiset riippuvuudet “kestävämpiä”)

Merkitsevyysmitat

- Riippuvuuden vahvuus ei kerro kaikkea! Voisi johtua sattumasta $\Rightarrow$ arvioitava tilastollinen merkitsevyys
- Idea: Oletetaan että X ja A riippumattomia. Mikä on todennäköisyys p, että olisimme silloin havainneet vähintään näin vahvan riippuvuuden?
- Jos p on pieni, riippuvuus tuskin johtui sattumasta.
- Tarkka p luotettava mitta varsinkin muuttujaperustaisille säännöille! (Riippuvuudet pätevät suurin piirtein yhtä vahvoina myös tulevaisuudessa.)
- Muita tunnettuja hyvyysmittoja: χ^2 ja mutual information (MI) + paljon heuristisia mittoja!

Esimerkkejä mitoista

$$p_F = \frac{\sum_{i=0}^{\infty} \binom{fr(X)}{fr(XA)+i} \binom{fr(\neg X)}{fr(\neg X\neg A)+i}}{\binom{n}{fr(A)}}$$

$$MI = \log \frac{P(XA)^{P(XA)} P(X\neg A)^{P(X\neg A)} P(\neg XA)^{P(\neg XA)} P(\neg X\neg A)^{P(\neg X\neg A)}}{P(X)^{P(X)} P(\neg X)^{P(\neg X)} P(A)^{P(A)} P(\neg A)^{P(\neg A)}}$$

$$\chi^2 = \frac{n(P(XA) - P(X)P(A))^2}{P(X)P(\neg X)P(A)P(\neg A)}$$

Kannattaa suosia p_F:ää!

Säännön redundanssi

Intuitio: Spesifimpi sääntö $XQ \rightarrow A$ on redundantti, jos se ei lisää mitään uutta informaatiota yleisempään sääntöön $X \rightarrow A$.

esim. Jos tunnetaan $stressi \rightarrow verenpaine$, $stressi, suklaa \rightarrow verenpaine$ lisää tuskin informaatiota.

- Arvoperustaiset säännöt: Oltava $P(A|XQ) > P(A|X)$, jotta $XQ \rightarrow A$ olisi ei-redundantti. Lisäksi konfidenssin parannuksen oltava tilastollisesti merkitsevä

Säännön redundanssi

- Muuttujaperustaiset säännöt: Huomioitava myös $P(\neg A|\neg(XQ))$ vs. $P(\neg A|\neg X)$.
  - Jos $P(A|XQ) > P(A|X)$, on usein $P(\neg A|\neg(XQ)) < P(\neg A|\neg X)$.
  - Punnittava näitä: Onko $P(A|XQ)$:n tuottama parannus merkitsevämpi kuin $P(\neg A|\neg X)$:n tuottama?
  - Käytännössä selviää tutkimalla onko $M(XQ \rightarrow A) > M(X \rightarrow A)$ (kunhan M fiksu mitta)

Analogia: Suuret makeat omenat

Tarjolla on vain punaisia ja virheitä omenoita, joista osa suuria ja osa pieniä.
Tunnetaan omenoiden makeutta indikoiva sääntö $punainen \rightarrow makea$.
Milloin $punainen, suuri \rightarrow makea$ on ei-redundantti?

- “suuruuden” lisättävä makeuden todennäköisyyttä
- Entä jos pienet punaiset omenat ovat toisinaan makeita?
- toimii, jos pieniä punaisia makeita omenoita tosi vähän tai pienet punaiset ovat happamia

Tilastollisen riippuvuuden epämonotonisuus

Määr. Ominaisuus O on monotoninen (osajoukkorelaatiolle), mikäli kaikilla $Z \supseteq X$ pätee $O(X) \Rightarrow O(Z)$ ja antimonotoninen, mikäli kaikilla $X \subseteq Z$ pätee $O(Z) \Rightarrow O(X)$. Muuten O on epämonotoninen.

- esim. ominaisuus $fr(Y) \geq min_{fr}$ on antimonotoninen
- ominaisuus $fr(YC) = 0$ on monotoninen
- Hakua helpottavia ominaisuuksia!
- Mutta: Kumpikaan ei päde tilastolliselle riippuvuudelle...

Epämonotonisuus: Merkitys

- $AB \rightarrow C$ voi ilmaista merkitsevää riippuvuutta, vaikka A ja C sekä B ja C riippumattomia toisistaan.
- Pahimmassa tapauksessa datajoukon ainut riippuvuus koskee kaikkia attribuutteja $A_1 \ldots A_k$ (esim. $A_1 \ldots A_{k-1} \rightarrow A_k$)

$\Rightarrow$ 1) Ahne heuristiikka ei toimi!

$\Rightarrow$ 2) Kahden välisten riippuvuuksien tutkiminen ei paljasta kaikkea!

Simpsonin paradoksi

ABCA1-R219K $\rightarrow$ ¬alzheimer
ABCA1-R219K, nainen $\rightarrow$ alzheimer

(Jopa ABCA1-R219K $\rightarrow$ alzheimer tai ABCA1-R219K $\perp\!\!\!\perp$ alzheimer)

Positiivinen riippuvuus muuttuu negatiiviseksi, kun sääntöä spesifoidaan, tai päinvastoin.

Riippuvuuden suunta riippuu siitä, minkä osaryhmän edustajia datassa oli!

$\Rightarrow$ Lääkkeeksi edustava data ja riittävän monimutkaisien riippuvuuksien analyysi.

Riippuvuussääntöjen rajoitukset

- perusmuodossa vain binääridatalle
- yleensä ei sallita negaatioita ehto-osassa $\rightarrow$ luotava uudet attr. negaatioille
- monet ohjelmat eivät salli edes seurausosassa (Kingfisher poikkeus)
- soveltuvat sellaisenaan esiintymisdatan analyysiin!
  - Kasvi- ja eläinlajien esiintymisdata
  - Geenien (alleelien) esiintymisdata
  - Ostoskoridata

Ei-binääristen muuttujien käsittely

Perusstrategia:

- Kategorisen muuttujan kaikille arvoille luodaan uudet binääriattribuutit
  - esim. 5 pyyntipaikkaa $\rightarrow$ luodaan binääriattribuutit $paikka1, \ldots, paikka5$
  - ei hukkaa mitään riippuvuuksia!
- Numeeriset muuttujat diskretoidaan etukäteen
  - Saattaa hukata riippuvuuksia – ei takaa enää globaalia optimia!
  - Ongelma: Eri diskretointi voi olla optimaalinen eri riippuvuuksissa $\rightarrow$ olisi parempi diskretoida dynaamisesti, haun aikana (onnistuu, mutta nykymenetelmillä tehotonta)

Lisää attribuuttien muodostusvinkkejä

Hämäläinen: Thorough analysis of log data with dependency rules: Practical solutions and theoretical challenges. (kannattaa tutkia!)

`http://www.cs.uef.fi/~whamalai/research.html`

Ja algoritmeista kiinnostuneille (hyvin yleistajuisesti):

Hämäläinen: Globaali optimonti haastaa lokaalit menetelmät: luotettavampia riippuvuuksia tehosta tinkimättä.

Varoituksen sana: Assosiaatiosäännöt

- Esitetään usein samannäköisesti $X \rightarrow A$ tai $X \Rightarrow A$
- tilastotieteessä “assosiaatio”=kategoristen muuttujien välinen riippuvuus
- Tiedonlouhinnassa ei yleensä merkitse sitä!
- X ja A voivat olla tilastollisesti riippumattomia tai voi olla $P(XA) < P(X)P(A)$ (vaikka käyttäjä luulee päinvastaista)
- Assosiaatiosääntö kertoo vain, että kombinaatio XA on riittävän frekventti, ts. $P(XA) > min_{fr}$ jollain raja-arvolla $min_{fr} \in [0, 1]$

Varoituksen sana: Assosiaatiosäännöt

- Periaatteessa voi karsia jälkikäteen, jolloin jäljelle jäävät tilastollisilla mitoilla parhaat frekventit säännöt
- Ei silti korvaa haun aikana hukattuja sääntöjä!
- algoritmit yleensä niin tehottomia, että oltava suuri min_{fr} (jopa 0.6)
- jos tehokkaampia, käyttävät muita järjettömiä heuristiikkoja (esim. suljettujen joukkojen heuristiikka suosii monimutkaisia ylisovittuneita sääntöjä)
- Poikkeus: Geoff Webbin MagnumOpus etsii tilastollisilla mitoilla ja toimii jopa ilman minimifrekvenssiä

Jos käytät assosiaatiosääntöjä

- tarkista huolella, millaisia hahmoja tuottavat!
- parhaimmillaan $\approx$ riippuvuussääntöjä, pahimmillaan vain harhaanjohtavia
- Luokittelusäännöt (classification rules), joissa seurausosa luokka-attribuutti, perustuvat yleensä tilastolliseen riippuvuuteen
  - tutki silti onko lisäkriteerejä tai vaatiiko min_{fr}:ää
- Pieni min_{fr} (esim. $min_{fr} = \frac{5}{n}$) ei haittaa!

Numeeristen (ja ordinaalisten) muuttujien riippuvuudet

- Kahden muuttujan riippuvuudet
  - systemaattinen haku onnistuu
- Monen muuttujan riippuvuudet
  - ML-menetelmät: yhden–muutaman mallin oppiminen

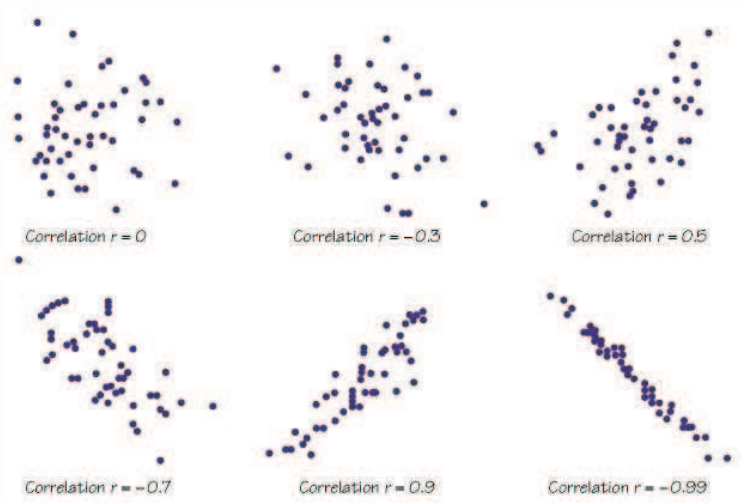
Muistutus: ordinaalisen muuttujan arvoilla järjestys, mutta eivät “oikeita” numeroita.
Esim. vatsahaavalla 4 arvoa: puuttuu, lievä, keskivaikea, vaikea.

Pearsonin korrelaatiokerroin

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

- numeerisille muuttujille X ja Y
- paljastaa lineaarisen riippuvuuden (pos. tai neg.)
- jos X ja Y riippumattomia, $r \approx 0$
- Voi olla riippuvuus, vaikka $r = 0$!
  - Jos X ja Y normaalijakautuneita, $r = 0$ merkitsee riippumattomuutta
- Varo: Hyvin herkkä kohinalle ja ulkopuolisille (outliers)!

Esimerkkejä



Lähde: `http://knottwiki.wikispaces.com/`

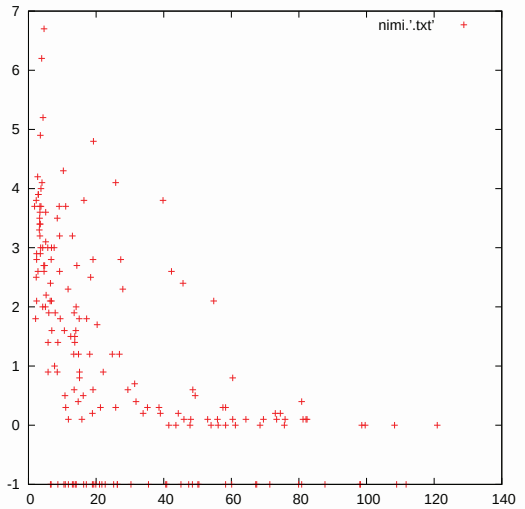
Järjestyskorrelaatiokertoimet (rank correlation coefficients)

- tunnetuimpia Kendallin τ ja Spearmanin ρ
- hyviä ordinaalimuuttujille!
- periaatteessa numeeristen muuttujien monotonisen riippuvuuden tunnistamiseen
- **Mutta:** varsinkin Spearman herkkä kohinalle/yksilölliselle vaihtelulle
- eivätkä osaa hyödyntää tietoa etäisyyksistä...

→ pieni diskretointi (**moneen** osaväliin) voisi auttaa?

Esimerkki: Infant Mortality–Physicians

$r = -0.53$, $\tau = -0.36$, $\rho = 0.03$



Kahden muuttujan mv. riippuvuus

Ehdotettuja mittoja esim.
1) Korrelaatiosuhde (correlation ratio) $\eta(X, Y)$

- X diskretoidaan ja verrataan Y:n varianssia X:n osaväleillä vs. koko datassa
- jos lineaarinen riippuvuus, $\eta^2 = r^2$
- $\eta - |r|$ kertoo epälineaarisuuden asteesta
- Ongelma: riippuu X:n diskretoinnista! (Jos paljon osavälejä, $\eta = 1$)
- Huomaa: $\eta(X, Y) \neq \eta(Y, X)$

Kahden muuttujan mv. riippuvuus

2) Mutual information $MI(X, Y)$

- Diskretoidaan X ja Y K:hon yhtä suureen osaväliin
- Lasketaan sitten

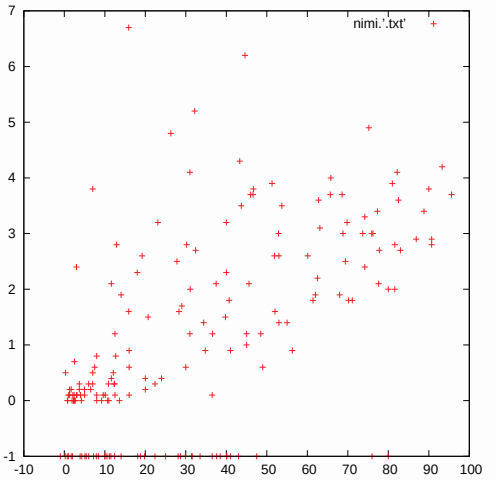
$$MI(X, Y) = \sum_{i=1}^{K} \sum_{j=1}^{K} P(X = x_i, Y = y_j) \log \frac{P(X = x_i, Y = y_j)}{P(X = x_i)P(Y = y_j)}$$

- + mahd. normalisointi

Pian päässemme testaamaan, miten hyvin toimii!

Joskus binarisointi+riippuvuussäännöt löytävät!

Internet users – Physicians



Riippuvuussäännöt

mostInternetUsers → **mostPhysicians**
fr=62 (0.3370), cf=0.816, M=-3.639e+01
leastInternetUsers → ¬ **mostPhysicians**
fr=84 (0.4565), cf=0.785, M=-3.499e+01
leastPhysicians → **leastInternetUsers**
fr=53 (0.2880), cf=0.930, M=-2.536e+01
mostInternetUsers → ¬ **leastPhysicians**
fr=72 (0.3913), cf=0.947, M=-2.473e+01

Monen muuttujan riippuvuudet: lineaariregressio (Multiple linear regression)

$$Y = \beta_1 X_1 + \ldots + \beta_k X_k + \alpha$$

- kuvaa Y:n **lineaarista** riippuvuutta muuttujista $X_1, \ldots, X_k$
- $X_1, \ldots, X_k$ selittävät muuttujat (explanatory variables)
- $\beta_1, \ldots, \beta_k$ regressiokertoimet (regression coefficients)
  - voivat kertoa myös muuttujien tärkeydestä
- koska yksinkertainen, on hyvin resistantti ylisovittumista vastaan!
  - Suosi erityisesti, jos vähän dataa

Rajoitukset

- riippuvuuden oltava lineaarinen, mutta pieni poikkeama ei vaarallista
  - Jos ennustava mallinnus, sietää suurtakin poikkeamaa!
- ongelma datan (suuret) aukot: onko lineaarinen aukkojen kohdalla?
- hyvin herkkä ulkopuolisille (outliers)!
- selittävien muuttujien oltava keskenään riippumattomia

Ongelma: Jos X_i:t korreloivat keskenään

- = “multicollinearity”
- β_i:t epätarkkoja–järjettömiä → eivät sovi ainakaan datan deskriptiiviseen mallinnukseen!
- mitä pienempi data ja enemmän kohinaa, sitä pahempi ongelma
- peukalosääntö: oltava $r(X_i, X_j) < \min\{r(X_i, Y), r(X_j, Y)\}$ $(\forall i, j)$
- periaatteessa korrelaatiot voi poistaa PCA:lla

Deskriptiivinen analyysi: Arvioi mallin validius!

1. Arvoi lineaarisuus laskemalla

$$r^2 = \frac{Var(Y) - MSE}{Var(Y)}$$

   - =multiple correlation coefficient
   - yleistää Pearsonin korrelaatiokertoimen monelle muuttujalle!
   - mitä enemmän $r^2 \to 1$, sitä lineaarisempi
2. Tutki residuaalit (virheet) $\epsilon_i = y_i - (\beta_1 x_{11} + \ldots + \beta_k x_{1k} + \alpha)$ (olt. riippumattomia, normaalijakautuneita, ka. 0 ja sama varianssi)
3. Tutki miten hyvin sopii dataan (F-testi)

Parhaan lineaariregressiomallin haku

- Valittava muuttujat X_i ja niiden parametrit β_i ja α
- yhden mallin sovitus dataan kevyttä $O(n)$ (mutta muuttujien valinta pahimmillaan $O(2^k)$)
- silti globaalisti optimaalisia menetelmiä vain pienidimensioiselle datalle → voisi varmasti kehittää tehokkaampiakin!
- valitettavan usein ahne haku
  - Aloitetaan yksinkertaisista malleista ja täydennetään aina parasta
    Tutki kaikki $\{Y = \beta_i X_i + \alpha_i\}$ → esim. $\beta_3 X_3 + \alpha_3$ paras
    Tutki kaikki $\{Y = \beta_3 X_3 + \beta_i X_i + \alpha_i\}$ ($i \neq 3$) jne.
  - tai aloita $Y = \beta_1 X_1 + \ldots \beta_k X_k + \alpha$:sta ja poista muuttujia 1 kerrallaan

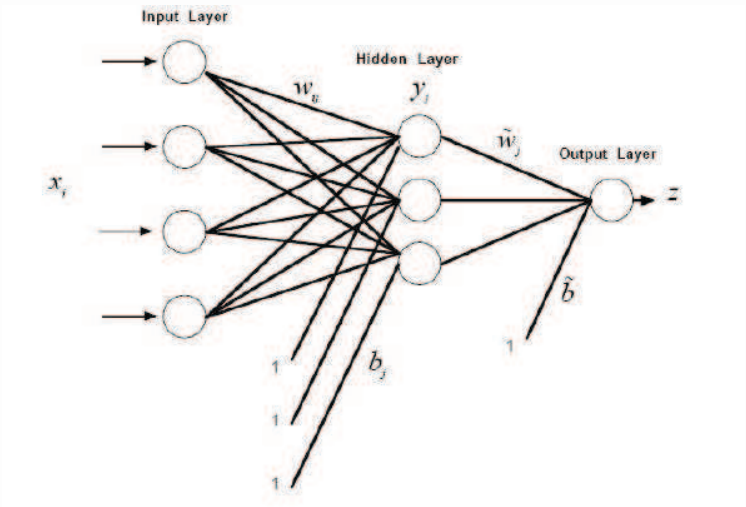
Laajennoksia

- $Y = \beta_1 f_1(X_1) + \ldots + \beta_k f_k(X_k) + \alpha$, missä f_i:t "pehemeitä" funktioita (esim. $\log$ tai $\sqrt{}$)
  - voi auttaa jos riippuvuus ei-lineaarinen
  - silti resistentti ylisovittumiselle!
  - hyvät funktiot löytyvät yleensä vain kokeilemalla
- joskus kannattaa muuntaa Y:tä, esim. $\log(Y)$ tai $\sqrt{(Y)}$
- logistinen regressio: Y on binäärinen tai ordinaalinen
- Y voi olla paloittain lineaarinen eri X_i:den yhdistelmien suhteen (piecewise linear) → malli kootaan yksinkertaisista lokaaleista hahmoista (splines)

Laajennoksia (jatkuu)

- korkeamman asteen termit, kuten $\beta_{i1} X_i$, $\beta_{i2} X_i^2$, $\beta_{i3} X_i^3$
- muuttujien vuorovaikutusta kuvaavat termit (intercation terms), kuten $\beta_{ij} X_i X_j$, $\beta_{ijl} X_i X_j X_l$, $\beta_{ijlm} X_i X_j X_l X_m$
- kaikki mahdollisia vuorovaikutustermejä yht. $2^k - 1$ kpl!
- monimutkaiset mallit, helposti ylisovittuvia!
- neuroverkoilla voi kuvata mielivaltaisia riippuvuuksia $Y = f(X_1, \ldots, X_k)$! (yleinen regressiomalli)

Yleinen regressio neuroverkolla



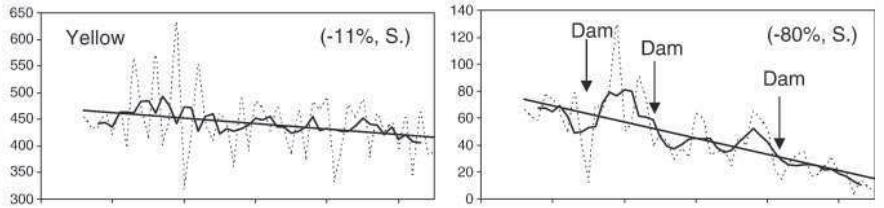
Lähde: `web.unbc.ca/~ytang/Chapter6.pdf`

Yleinen regressio neuroverkolla: Muista!

- ylisovittuu erittäin helposti!
- oppimisalgoritmi vaatii paljon ja edustavaa dataa jotta luotettavia tuloksia
- mutta ei jaksa käsitellä liian suuria datamääriä...
- eivät koskaan takaa globaalia optimia!
- Peukalosääntö: Valitse oletusarvoisesti vain 3-tasoinen verkko (vain 1 piilotaso)
  - riittää minkä tahansa funktion esittämiseen
  - auttaa ylisovittumista vastaan
- “Neuroverkko työkaluna on yhtä hyvä kuin käyttäjänsä”

Esimerkki: Trendien kuvaus lineaariregressiolla

Sademäärät ja joen vesimäärän muutokset Keltaisenjoen alueella 1945–2000



Lähde: Xu et al.: Temporal trend of precipitation and runoff in major Chinese Rivers since 1951, Global and Panetary Change 73, pp. 219–232, 2010

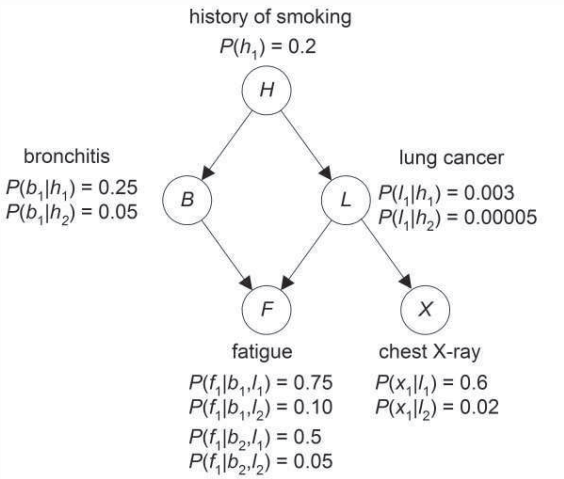
Ekskursio: Bayes-verkot

- esittää datan riippuvuusrakenteen ja implisiittisesti kokonaistodennäköisyysjakauman
- koostuu verkon rakenteesta ja solmujen todennäköisyystauluista
- Perusidea: solmut vastaavat muuttujia ja kaaret esittävät riippuvuuksia
- Jos kaksi muuttujajoukkoa X ja Y ehdollisesti riippumattomia annettuna kolmas Z, niin verkossa kaikki X:n ja Y:n solmut jonkin Z-solmun erottamia

Ekskursio: Bayes-verkot

- Juurisolmuille A_i annetaan prioritodennäköisyydet $P(A_i = a_i)$ (kaikilla A_i:n arvoilla a_i)
- Muille solmuille B_i ehdolliset todennäköisyydet $P(B_i = b_i | X = \overline{x})$ (X=solmun vanhemmat ja $\overline{x}$=niiden arvokombinaatio)

Esimerkki



Lähde: Jiang. et al. Learning genetic epistasis using Bayesian network scoring criteria, Bioinformatics 12:89 2011.

Ekskursio: Bayes-verkot

- Havainnollinen esitys riippuvuuksista (apua jopa kausaalisuhteiden keksimiseen)
- Rakenteen oppiminen laskennallisesti hankalaa!
- Ei siis takeita globaalista optimista (kuten ei yleensä)
- DM:ssä ei kuitenkaan järkeä määrittää rakennetta ad hoc