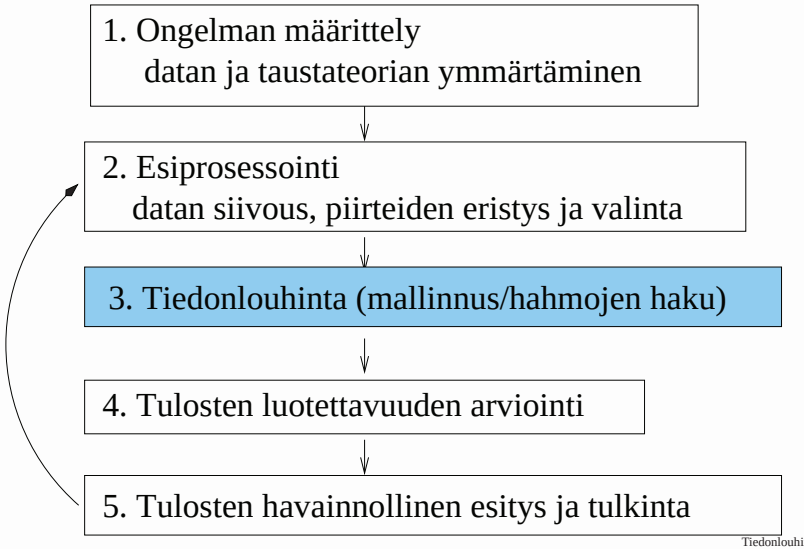
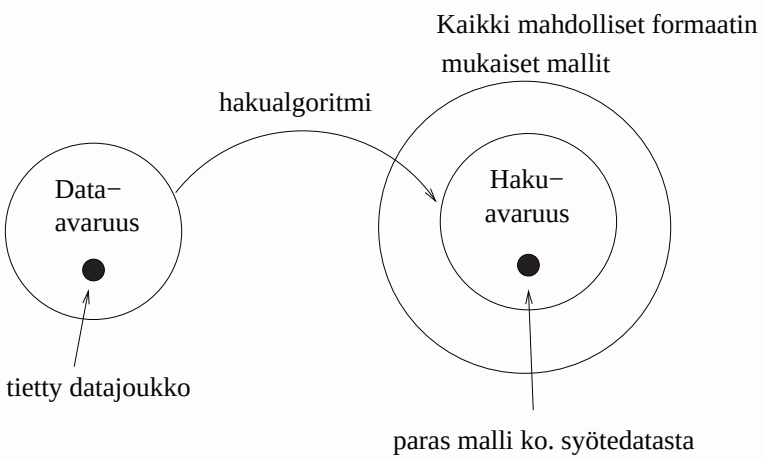
Luento 3: 1 Yleiset mallinnusperiaatteet ja 2 Riippuvuusanalyysi



Yleiset mallinnusperiaatteet

- data-avaruus
  = kaikki mahdolliset muuttujien arvokombinaatiot
- hakuavaruus
  = kaikki mahdolliset annetun formaatin mukaiset mallit/hahmot, jotka datasta voisi johtaa
- globaalisti ja lokaalisti optimaaliset hakualgoritmit
  = Ovatko löydetyt mallit/hahmot varmasti parhaita mahdollisia?
- ylisovittuminen
  Kuvaako malli liian hyvin syötedatan?

Parhaan mallin haku datasta



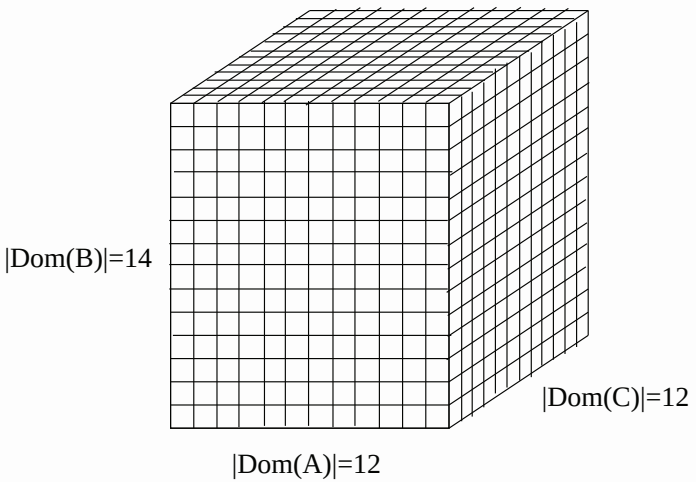
Huom! Annetut muuttujat määräävät data-avaruuden.

Data-avaruus

Olk. muuttujat $A_1, \ldots, A_k$, joiden arvoalueet $Dom(A_1), \ldots, Dom(A_k)$. Silloin $\mathcal{D} = Dom(A_1) \times \ldots \times Dom(A_k)$ määrittää data-avaruuden.

- Datan dimensio k
- Jos kaikki muuttujat diskreettejä, voi kokoa luonnehtia solujen lukumäärällä: $|\mathcal{D}| = |Dom(A_1)| \cdot \ldots \cdot |Dom(A_k)|$.
  - Nyt $|Dom(A_i)| = |\{a_{i1}, \ldots, a_{im}\}| = m$
- Jos jatkuvia muuttujia, ei voi laskea solujen lkm:ää!
  - Nyt $|Dom(A_i)| = |[a_{min}, a_{max}]| = a_{max} - a_{min}$ (voi olla ∞)

Esim. 3-ulotteisen diskreetin datan esitys



Koko $12 \cdot 12 \cdot 14 = 2016$ solua.

Hakuavaruus

- Kattaa kaikki mahdolliset mallit, jotka mallinnusmenetelmän ja data-avaruuden mukaisia ja jotka voi löytää annetulla hakualgoritmilla
- Data-avaruus (annetut muuttujat ja niiden arvoalueet) määräävät, millaisia malleja voi löytää
- Samaan mallinnusmenetelmään (“mallinnusparadigmaan”, esim. päätöspuut, neuroverkot, riippuvuussäänöt) voi liittyä erilaisia hakualgoritmeja!
  $\rightarrow$ voivat löytää eri malleja

Esimerkki: riippuvuussääntöjen hakuavaruus

Binäärimuuttujat $R = \{A_1, \ldots, A_k\}$ ja etsitään säännöt muotoa $X \rightarrow A$, $X \subsetneq R$, $A \in X$. Hakuavaruus koostuu kaikista mahdollisista säännöistä eli

1. Tutkitaan kaikki mahdolliset R:n osajoukot Y, joiden koko vähintään 2 attribuuttia ($|Y| \geq 2$)
   - R:n **potenssijoukko** (koko 2^k)

$$\mathcal{P}(R) = \{\emptyset, A_1, A_2, \ldots, A_k, A_1A_2, A_1A_3, \ldots, A_{k-1}A_k, A_1A_2A_3, \ldots, \ldots, A_1A_2 \ldots A_k\}$$

   - Ei tarvitse tutkia joukkoja $\emptyset, \{A_1\}, \ldots, \{A_k\}$.

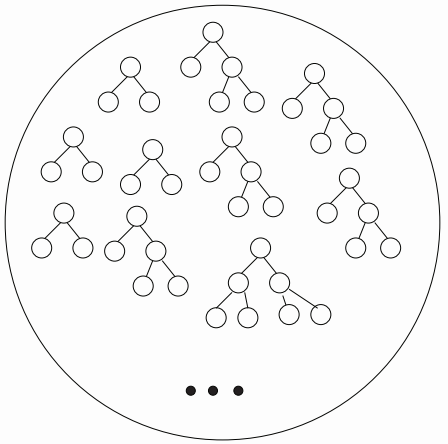
Esimerkki (jatk.)

2. Jokaisesta osajoukosta Y muodostetaan kaikki mahdolliset säännöt
   - Jos $|Y| = l$, voi muodostaa l sääntöä, $Y \setminus \{A_i\} \rightarrow A_i$ ($\forall A_i \in Y$).

Hakuavaruuden koko on siis $\sum_{i=2}^{k} i \cdot \binom{k}{i} = \mathcal{O}(k \cdot 2^k)$

Hakuavaruus (jatk.)

Jos binäärimuuttujat $R = \{A_1, \ldots, A_k\}$ ja luokkamuuttuja C. Millainen olisi kaikkien mahdollisten päätöspuiden joukko?



Mallin hyvyysmitta

- Tarvitaan jokin hyvyysmitta M, joka kertoo mallin hyvyyden.
  - esim. miten monta alkiota luokittelee oikein, miten vahva tai merkitsevä riippuvuus on, miten tiheä klusteri on tms.
  - yleensä pyrkii heijastamaan myös sitä, miten hyvin malli yleistyy mallin muodostusdatan ulkopuolelle (ts. että hahmo pätee “oikeasti” tai malli toimii tulevaisuudessa)
- Jos malli paranee, kun M kasvaa, on M “hyvyyden mukaan kasvava” (esim. luokittelun accuracy, χ^2-mitta)
- muuten taas “hyvyyden mukaan vähenevä” (esim. luokitteluvirhe, p-arvot)

Hakuongelma: 2 variaatiota

Annettu mallinnusmenetelmä, data ja hyvyysmitta M.

1. Luettelointiongelma (enumeration problem): Etsi kaikki riittävän hyvät mallit (annettu jokin M:n raja-arvo).
2. Optimointiongelma (optimization problem): Etsi K parasta mallia (ML:ssä yleensä $K = 1$)

- Loppujen lopuksi ei suurta eroa!
- Käyttäjä ei halua tulokseksi tuhansia (saati miljoonia) hahmoja!

$\rightarrow$ Keskitytään optimointiongelmaan

Hakualgoritmin optimaalisuus?

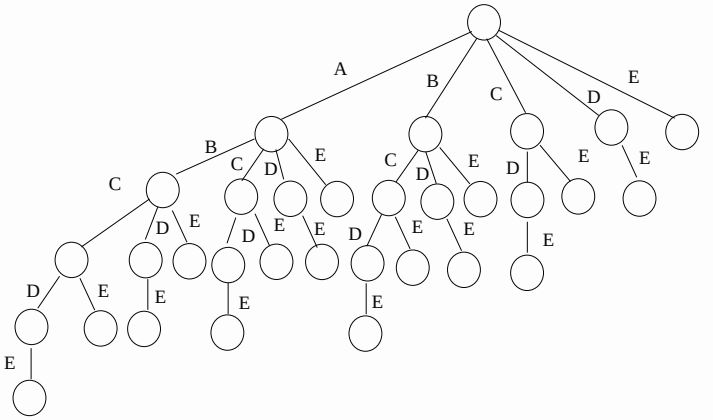
- Globaalisti optimaalinen, jos palauttaa parhaat K mallia hakuavaruudesta
  - Takuu: datasta ei mahdollista löytää näitä parempia annetun muotoisia malleja.
- Lokaalisti optimaalinen: Löydetyt mallit parhaita jossain hakuavaruuden osassa, mutta eivät välttämättä koko hakuavaruudessa.
  - algoritmi ei tutki koko avaruutta, vaan valikoi tutkittavan alueen heuristisesti
  - algoritmi voi “juuttua” lokaaliin optimiin

Heuristiikka

- Intuitiivisesti: apumenetelmä tai periaate, jota soveltamalla ongelma ratkeaa nopeammin mutta tulokset silti melko hyviä/useimmiten hyviä.
- yleensä ovat “alioptimaalisia” (suboptimal) eli eivät takaa globaalia optimia (parhaan mallin löytymistä)
- Esim. **ahne heuristiikka** (greedy heuristic): Edetään haussa sillä hetkellä parhaalta näyttävään suuntaan.
  - Tyypillistä ihmisille! Tyyliin “Väsyttää ja masentaa, hirveästi rästitöitä, paras pitää lökeripäivä.” (optimoi mielihyvää lyhyellä tähtäimellä)
- joskus heuristiikan virheellisyys merkitsee vain hidasta hakua, ei oikeiden tulosten hukkaamista (riippuu mihin heuristiikkaa käytetään)

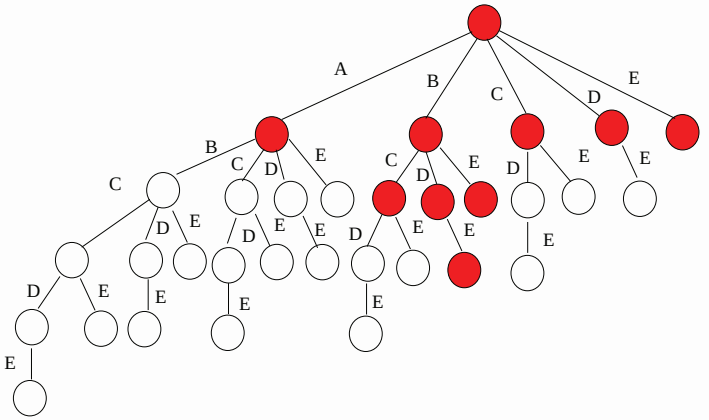
Esimerkki: Luokittelusääntöjen hakuavaruus

Olk. $R = \{A, B, C, D, E\}$ ja luokka-attribuutti F. Hakuavaruus $\mathcal{P}(R)$ voidaan esittää luettelointipuuna (enumeration tree).



Parhaan säännön ahne haku

Jos tasoittain haku ylhäältä alas ja ahne haku tuottaisi $BDE \rightarrow F$:n. Tutkisi vain 9 solmua 32:sta, mutta voi mennä metsään!



Varoituksen sana!

Aina menetelmää/sen hakualgoritmia valitessasi

- Epäile alioptimaalista heuristiikkaa!
- Ota selvää, mikä heuristiikka on ja milloin se voi hukata parhaat mallit.
- Miten todennäköistä on, että saadaan huonoja tuloksia?
- Hälytyskellojen paras soida, jos kuulet sanan “ahne heuristiikka” (vaikka joskus harvoin turha hälytys)
- Ongelma: Menetelmien kehittäjät eivät mielellään kerro, ettei menetelmä ole globaalisti optimaalinen...

Mallin ylisovittuminen (overfitting)

Intuitiivisesti: Malli/hahmo h kuvaa mallinnettavan datan d liian yksityiskohtaisesti, mukaan lukien satunnaisvaihtelun, eikä siksi yleisty d:n ulkopuolelle.

Täsmällisemmin: h ylisovittunut, jos $\exists h'$ s.e. $M(h,d) > M(h',d)$ mutta $M_{true}(h) < M_{true}(h')$, missä M_{true} on mallin "oikea" hyvyys.

- Oikeaa hyvyyttä M_{true} ei voi mitata! Koetettava arvioida heuristisesti.
- Oletus: M hyvyyden mukaan kasvava (jos ei, niin vaihda $>$ ja $<$)

Occamin partaveitsi –periaate

Tärkeä havainto: Ylisovittuminen edellyttää monimutkaista mallia.

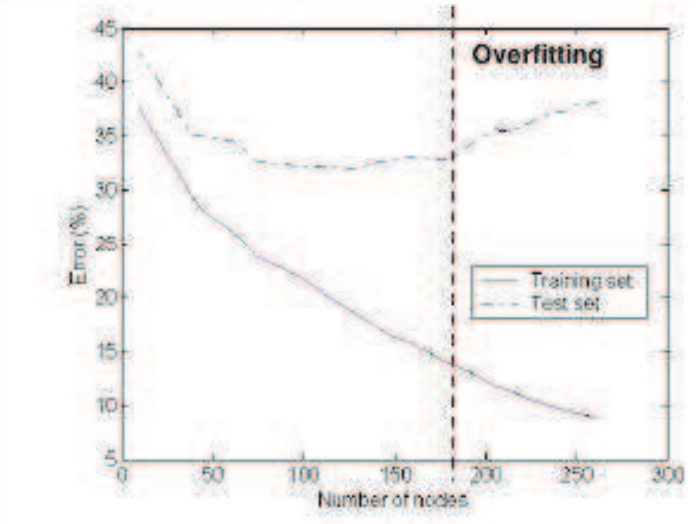
- monimutkaisella mallilla voi kuvata yksityiskohdat $\rightarrow$ tarkka kuvaus mutta harvoille (tai vain yhdelle) datajoukoille
- yksinkertaisella mallilla ei voi kuvata kaikkia yksityiskohtia $\rightarrow$ kuvaa ylimalkaisemmin mutta useampia datajoukkoja
- $\Rightarrow$ ylisovittumista voi estää kontrolloimalla mallin monimutkaisuutta!
- = Occamin partaveitsi (Occam's Razor) –periaate: (Yhtä hyvistä) malleista suosi yksinkertaisinta.
- tai rankaise monimutkaisuudesta

Alisovittuminen (underfitting)

Intuitio: Jos malli on liian yksinkertainen, se kuvaa minkä tahansa datan liian ylimalkaisesti.

- Helppo havaita! Hyvyysmitta saa huonon arvon myös opetusdatassa.
- Huom! Jos data oikeasti satunnaista, ei siitä voi muodostaa parempaa mallia!

Esimerkki: Päätöspuun ylisovittuminen



Mallin monimutkaisuus suhteutettava datan määrään!

Perusperiaate: mitä enemmän datarivejä, sitä monimutkaisemman mallin voi muodostaa.

- Peukalosääntö: oltava vähintään 5–10 datariviä jokaista malliparametria kohden
- Jos jokin hahmo esiintyy < 5 rivillä, ei mitään takeita, etteikö sattumaa
- $\Rightarrow$ Jos vähän dataa, käytettävä yksinkertaisia mallinnusmenetelmiä ja etsittävä yksinkertaisia malleja.

Mallin monimutkaisuus vs. datan määrä

- Poikkeus: tukivektorikoneet (support vector machines, SVM) hyviä välttämään ylisovittumisen, ainakin teoriassa (käytännössä valittava malliparametrit huolella!)

Tilastollisten riippuvuuksien analyysi

- Tilastollisen riippuuden määritelmä/tulkinnat
- Riippuvuussääntöjen haku kategorisesta (tai diskretoidusta) datasta
- Riippuvuuksien haku numeerisesta datasta
- Ekskursio: Bayesläinen riippuvuuksien mallinnus (Bayes-verkoilla)

Tilastollinen riippuvuus: Monta tulkintaa!

Määritelmä: Tapahtumat X ja Y ovat tilastollisesti **riippumattomia**, jos $P(XY) = P(X)P(Y)$.

Mutta: Milloin muuttujat tai muuttuja-arvokombinaatiot ovat tilastollisesti **riippuvia**?

(termi "**korrelaatio**" viittaa yleensä kahden numeerisen muuttujan väliseen lineaariseen riippuvuuteen, mutta käytetään väljästi)

Tilastollinen riippuvuus: Monta tulkintaa!

- Tilastollinen riippuvuus edellyttää siis että $P(XY) \neq P(X)P(Y)$. Mutta: pieni poikkeama $P(X)P(Y)$:stä johtuu todennäköisesti sattumasta! Miten arvioida, onko riippuvuus aitoa? $\rightarrow$ Mitat riippuvuuden voimakkuuden ja merkitsevyyden arviointiin.
- Jos A ja B binäärimuuttujia, riittää tutkia yhtä arvokombinaatiota: $P(A = 1B = 1) \neq P(A = 1)P(B = 1) \Leftrightarrow \forall a \in \{0, 1\}\forall b \in \{0, 1\}P(A = aB = b) \neq P(A = a)P(B = b)$. Entä jos A ja B moniarvoisia muuttujia? $\rightarrow$ muuttuja- ja arvoperustaiset tulkinnat
- Miten määritellä kolmen muuttujan, A, B ja C, välinen riippuvuus? $\rightarrow$ ainakin 3 vaihtoehtoista tulkintaa.

Tilastollinen riippuvuus: 3 tulkintaa

Olk. A, B, C binäärisiä. Merk. $\neg A \equiv (A = 0)$ ja $A \equiv (A = 1)$

1. Riippuvuussääntö $AB \rightarrow C$: oltava $\delta = P(ABC) - P(AB)P(C) > 0$ ja δ riittävän suuri.
2. Kokonaistodennäköisyysmalli: Lasketaan $\delta_1 = P(ABC) - P(AB)P(C)$, $\delta_2 = P(A\neg BC) - P(A\neg B)P(C)$, $\delta_3 = P(\neg ABC) - P(\neg AB)P(C)$ ja $\delta_4 = P(\neg A\neg BC) - P(\neg A\neg B)P(C)$. Jos $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, ei ainakaan riippuvuutta. Muuten päätellään jollain laskukaavalla d_i:stä ($i = 1, \ldots, 4$)
3. ABC on “korreloiva joukko” jos $\rho = P(ABC) - P(A)P(B)P(C) \neq 0$ ja ρ riittävän suuri. (Yl. olt. $\rho > 0$ eli positiivinen riippuvuus)

1. Riippuvuussäännöt

Yleisessä muodossa $X \rightarrow A = a$, $X \subsetneq R$, $A \in R \setminus X$ ja $a \in \{0, 1\}$.

- Yleensä ei siis sallita negaatioita ehto-osassa X (ei esim. $A_1\neg A_3 \rightarrow A_5$. Helppo kiertää: luo uudet binäärimuuttujat negaatioille! (esim. $A_i = B_{2i}$ ja $\neg A_i = B_{2i+1}$)
- $X \rightarrow A_i$ (tai $X \rightarrow A_i = 1$) merkitsee positiivista riippuvuutta ($P(XA_i) > P(X)P(A_i)$)
- $X \rightarrow \neg A_i$ (tai $X \rightarrow A_i = 0$) merkitsee negatiivista riippuvuutta X:n ja A_i:n välillä ($P(XA_i) < P(X)P(A_i)$ eli $P(X\neg A_i) > P(X)P(\neg A_i)$)

1. Riippuvuussäännöt: Esimerkkejä

korkea LDL $\rightarrow$ ateroskleroosi
korkea HDL $\rightarrow \neg$ ateroskleroosi
tupakoi $\rightarrow$ ateroskleroosi
paljon tupakkaa $\rightarrow$ alzheimer
kahvi, ei tupakoi $\rightarrow$ ateroskleroosi
kohtuudella alkoholia $\rightarrow \neg$ sydänsairaus
miestyypin kaljuus -> sydänsairaus
ApoE-e4 -> aivoinfarkti
ApoE-e4, kohtuudella alkoholia $\rightarrow$ alzheimer
ApoE-e4, kohtuudella tupakkaa $\rightarrow \neg$ alzheimer
ABCA1-R219K $\rightarrow \neg$ alzheimer
ABCA1-R219K, nainen $\rightarrow$ alzheimer

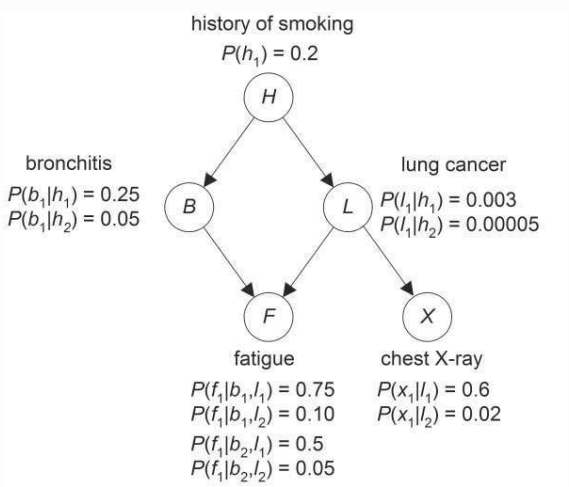
1. Riippuvuussäännöt

- lokaaleja hahmoja, eivät globaaleja malleja
- erittäin tehokkaat algoritmit! (silti globaalisti optimaalisia)
- tuhansille tai jopa kymmenille tuhansille binäärimuuttujille (riippuu datan jakaumasta, miten tehokkaita)
- ei-binäärimuuttujat edellyttävät binärisointia

2. Kokonaistodennäköisyysmalli

- esittävät kokonaistodennäköisyysjakauman $P(A_1 = a_1 \ldots A_k = a_k)$, $\forall a_i \in \{0, 1\}$
- globaaleja malleja
- **paljon** raskaampia oppia $\rightarrow$ esivalittava joukko hyviä muuttujia, silti vain lokaalisti optimaalisia
- voi esittää havainnollisesti Bayes-verkkona

2. Kokonaistodennäköisyysmalli: Esimerkki



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

3. Korreloiva joukko

Yleisessä muodossa $X \subseteq R$, missä $P(X) > \Pi_{A_i \in X} P(A_i)$

- kätevä kun ei selkeää säännön seurausosaa
- esim. esiintymisdata (ocurrence data)
  - Data: kasvilistoja eri havaintopaikoista. Hahmo: lajit jotka esiintyvät poikkeuksellisen usein yhdessä.
- monia heuristisia hahmoja ja niille hakualgoritmeja, mutta vähän hyviä (tilastollisesti mielekkäitä)
- Älä sekoita säännöllisiin joukkoihin (frequent sets)! Koostuu riittävän usein yhdessä esiintyvistä attribuuteista, joiden välillä ei välttämättä tilastollista riippuvuutta.