Luku 8. Diskreettejä menetelmiä laajojen 0-1 datajoukkojen analyysiin

T-61.2010 Datasta tietoon, syksy 2011

professori Heikki Mannila

Tietojenkäsittelytieteen laitos, Aalto-yliopisto

28.11.2011

Kurssin loppuosa

- Kurssin alkuosassa muuttujat olivat pääsääntöisesti reaalilukuarvoisia. Kahdessa viimeisessä luvussa tarkastellaan *diskreettejä hahmoja*.
- Luku 8: Diskreettejä menetelmiä laajojen 0-1 datajoukkojen analyysiin
  - Kattavat joukot ja niiden etsintä tasoittaisella algoritmilla
- Luku 9: Relevanttien sivujen etsintä verkosta: satunnaiskulut verkossa
  - Linkkikeskukset ja auktoriteetit (“hubs and authorities”) -algoritmi
- Kutakin asiaa vain raapaistaan; kaikki merkittäviä tutkimusalueita

Tämän luvun sisältö

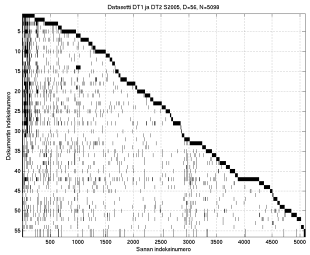
Suuret 0-1 datajoukot

- Paljon rivejä (muuttujia), paljon sarakkeita (havaintoja)
- 0-1 dataa: esiintymädataa. Esiintyykö jokin tietty ilmiö tietyssä havainnossa?
- Matriisin (taulun) D alkio $D(m, h)$ kertoo, esiintyykö muuttujan m kuvaama ilmiö havainnossa h eli $D(m, h) = 1$ tai $D(m, h) = 0$

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 1 & 0 \end{bmatrix}$$

0-1-datan esimerkki
Sanojen esiintyminen dokumenteissa

- Havainto: dokumentti
- Muuttuja: (perusmuodossa oleva) sana
- Data $D(s, d) = 1$: esiintyikö sana s ainakin kerran dokumentissa d?



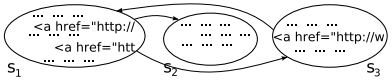
0-1-datan esimerkki (2)
Sanojen esiintyminen dokumenteissa

Kuva: Sana-dokumentti-matriisi tietokoneharjoituskierrokselta 5. Musta juova tarkoittaa, että sana s on dokumentissa d eli $D(s, d) = 1$. Mukana syksyllä 2005 palautettujen harjoitustöiden (kaksi aihetta) tekstiaineisto: 5098 eri sanaa ja 56 dokumenttia, joista kaksi tahallisesti generoituja (“kopioita”).

0-1-datan esimerkki
Webbisivut

- Havainto: webbisivu
- Muuttuja: webbisivu
- $D(s, p) = 1$ jos ja vain jos sivulta s on linkki sivulle p

(s, p)	s_1	s_2	s_3
s_1	0	1	1
s_2	0	0	0
s_3	1	0	0



0-1-datan esimerkki
Nisäkäslajien esiintyminen luonnossa

- Havainto: 50x50 km ruutu r
- Muuttuja: Nisäkäslaji l
- $D(l, r) = 1$ jos ja vain jos laji l esiintyi ruudussa r

0-1-datan esimerkki
Osamolekyylien esiintyminen molekyyleissä

- Havainto: Molekyyli m (tyypillisesti suuri)
- Muuttuja: Molekyyli p (pienempi)
- $D(p, m) = 1$ jos ja vain jos molekyyli p on osa molekyyliä m

NSF 0-1-dokumenttiesimerkki
Lähdeaineisto

- Abstraktitietokanta: 128000 abstraktia jotka kuvaavat NSF:n rahoittamia perustutkimushankkeita
  - `http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html`
- Esimerkki abstraktin tietokentistä
- Esimerkki abstraktista

NSF 0-1-dokumenttiesimerkki
Datan esitys

- Tekstidokumentit (abstraktit) muutettuna 0-1-taulukkomuotoon, jossa 1 = esiintyy ainakin kerran
- Sarakkeet vastaavat dokumentteja
- Rivit vastaavat sanoja
- Esimerkkisanoja indekseineen: . . ., 73:abscissa, 74:absence, 75:absent, 76:absolute, 77:absolutely, 78:absorb, 79:absorbed, . . .

	#1	#2	#3	. . .
327:additional	1	0	0	. . .
⋮	⋮	⋮	⋮	⋮
11339:genetic	1	1	0	. . .
⋮	⋮	⋮	⋮	⋮
26457:studies	1	1	1	. . .
⋮	⋮	⋮	⋮	⋮

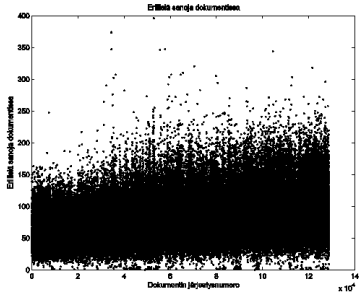
NSF 0-1-dokumenttiesimerkki
Miksi tällainen esitysmuoto?

- Tiedonhaun kannalta dokumentin sisältävien sanojen luettelo on monesti riittävä
- Määrämuotoista dataa on helpompi käsitellä kuin vaihtelevan mittaisia tekstijonoja

NSF 0-1-dokumenttiesimerkki
Perustilastoja

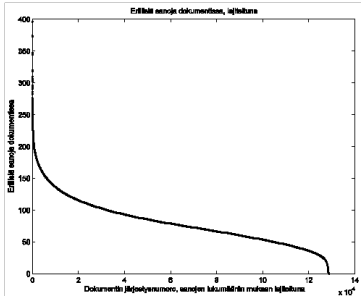
- Dokumentteja: 128000, sanoja: 30800; ei mitenkään erityisen suuri dokumenttijoukko
- Taulussa $30800 \times 128000 = 3942400000 \approx 4 \cdot 10^9$ alkiota
- Ykkösiä datassa 10449902 eli $0.002650 = 0.265$ % kaikista alkioista; ei kannata esittää nollia
- Noin 81 erilaista sanaa / dokumentti (1 = sana esiintyy ainakin kerran)
- Kukin sana esiintyy keskimäärin 340 dokumentissa
- Jakauma vino: jotkin sanat esiintyvät usein, toiset hyvin harvoin
- Joitakin useimmin esiintyviä sanoja on jätetty pois (“blacklist”)

NSF 0-1-dokumenttiesimerkki
Kuva #1: Erilaisten sanojen lukumäärä dokumenteittain



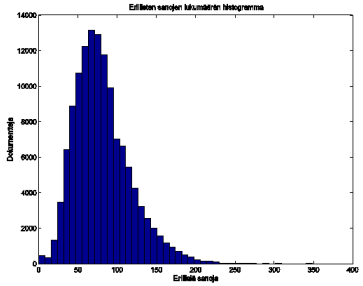
Kuva: *x-akseli*: Dokumentin järjestysnumero, yhteensä $12.8 \cdot 10^4$ dokumenttia. *y-akseli*: Erilaisten sanojen lukumäärä dokumentissa.

NSF 0-1-dokumenttiesimerkki
Kuva #2: Erilaisten sanojen lukumäärä dokumenteittain



Kuva: *x-akseli*: Dokumentin järjestysnumero sanojen lukumäärän mukaan lajiteltuna. *y-akseli*: Erilaisten sanojen lukumäärä dokumentissa

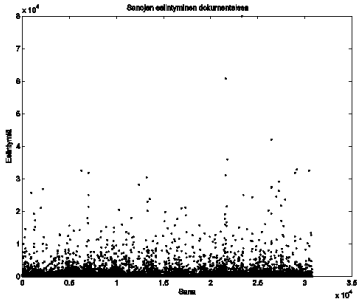
NSF 0-1-dokumenttiesimerkki
Kuva #3: Erilaisten sanojen lukumäärä dokumenteittain



Kuva: Erilaisten sanojen lukumäärä dokumentissa histogrammina. Dokumentin pituus vaihtelee pääosin noin 30 – 180 (eri) sanan välillä.

NSF 0-1-dokumenttiesimerkki

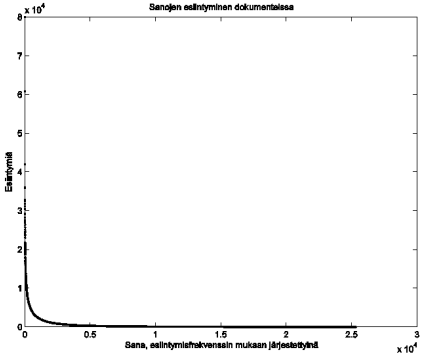
Kuva #4: Kuinka monessa dokumentissa kukin sana esiintyy?



Kuva: *x-akseli*: Sanan indeksinumero $(1, \ldots, 30800)$. *y-akseli*: Esiintymiä $\times 10^4$ $(1, \ldots, 80000)$

NSF 0-1-dokumenttiesimerkki

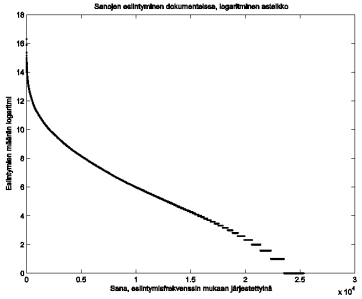
Kuva #5: Kuinka monessa dokumentissa kukin sana esiintyy?



Kuva: Sama kuin edellä, mutta nyt järjestettynä esiintymisfrekvenssin mukaisesti

NSF 0-1-dokumenttiesimerkki

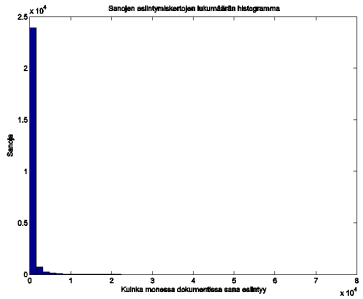
Kuva #6: Kuinka monessa dokumentissa kukin sana esiintyy?



Kuva: Sama kuin edellä, mutta nyt esiintymisfrekvenssi (y-akseli) 2-kantaisena logaritmina. Miksi kuvaaja on osalta $(\sim 4000, \ldots, 16000)$ lähes suora?

NSF 0-1-dokumenttiesimerkki

Kuva #7: Kuinka monessa dokumentissa kukin sana esiintyy?

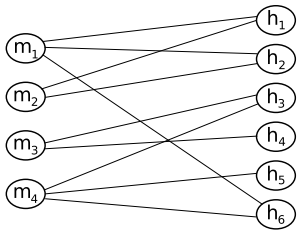


Kuva: Sanojen esiintymiskertojen lukumäärän histogrammi

0-1-datajoukko verkkona

- Verkko: joukko solmuja, joista osa on yhdistetty toisiinsa kaarilla
- 0-1-datajoukosta on helppo tehdä kaksijakoinen (“bipartite”) verkko: muuttujat ja havainnot ovat solmuja, ja muuttujan ja havainnon välillä on kaari jos ja vain jos vastaava datan arvo on 1

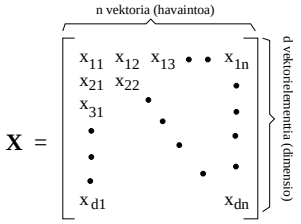
	h_1	h_2	h_3	h_4	h_5	h_6
m_1	1	0	1	0	0	1
m_2	1	1	0	0	0	0
m_3	0	0	1	1	0	0
m_4	0	0	1	0	1	1



Mitä tällaisesta datasta voisi etsiä?

- Mitä datassa on?
- Minkälaisia ryhmittymiä datan muuttujilla ja havainnoilla on?
- Miten muuttujat riippuvat toisistaan?
- Jne.
- Muuttujien ja havaintojen astelukujen analyysi (kuinka monta kaarta muuttujasta lähtee, eli kuinka monessa havainnossa muuttuja on mukana): power laws – hyvin aktiivinen tutkimusalue
- Jatkossa: *miten etsitään pieniä kiinnostavia muuttujajoukkoja*

Usein esiintyvien muuttujakombinaatioiden etsintä
Merkintöjä



- d-ulotteinen havaintovektori $\mathbf{x}$; sen alkiot $x_1, x_2, ..., x_d$
- Useat havaintovektorit $\mathbf{x}(1), \ldots, \mathbf{x}(n)$
- Havaintovektorin $\mathbf{x}(j)$ muuttujan i arvoa merkitään $x(i, j)$ (kuvassa x_{ij})

Kattavat joukot
Määritelmä

- Jos muuttujia on esim. 30000, kaikkien parittaisten korrelaatioiden etsintä ei ole mahdollista
- Usein yhdessä esiintyvät muuttujat?
- Etsi kaikki muuttujaparit (a, b) siten että on olemassa ainakin N havaintoa $\mathbf{x}(i)$, jolla $x(a, i) = 1$ ja $x(b, i) = 1$
- *Yleisemmin*: etsi kaikki muuttujajoukot $\{a_1, a_2, \ldots, a_k\}$ siten että on olemassa ainakin N havaintoa $\mathbf{x}(i)$, jolla $x(a_1, i) = 1$ ja $x(a_2, i) = 1$ ja $\ldots$ ja $x(a_k, i) = 1$
- Kutsutaan tällaista muuttujajoukkoa $\{a_1, a_2, \ldots, a_k\}$ *kattavaksi* (“frequent set”)

Kattavat joukot
Esimerkki kattavasta joukosta

- Olkoon tutkittavana neljä kauppakassia, joissa on tai ei ole appelsiineja (a), banaaneja (b) ja omenia (c).
- Kassi #1: 5 appelsiinia, 2 banaania. Kassi #2: 4 banaania. Kassi #3: 1 appelsiini, 2 banaania, 1 omena. Kassi #4: 8 appelsiinia, 2 omenaa.
- Koodataan esiintyminen kassissa 1:llä ja poissaolo 0:lla.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

- Kynnysarvolla $N = 2$ kattavia joukkoja ovat $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$ ja $\{a, c\}$. Esimerkiksi appelsiinit (a) ja omenat (c) esiintyvät yhdessä ($\{a, c\}$) vähintään kahdessa kassissa.

Kattavat joukot
Miten paljon kattavia joukkoja on?

- Jos datassa on pelkkiä ykkösiä, niin kaikki muuttujaparit tai muuttujaosajoukot täyttävät ehdon
- Ongelman mielekkyys edellyttää, että data on harvaa. (Tämä onkin tyypillistä. Edellisessä esimerkissä sekä kauppakasseja että tuotteita käytännössä tuhansittain, mutta yksi asiakas ostaa kerrallaan vain pienen määrän tuotteita. Harvassa matriisissa on siis nollia huomattavasti enemmän kuin ykkösiä.)
- Koneoppimisterminologialla: etsitään usein esiintyviä positiivisia konjunktioita ($A \wedge B$)

Kattavat joukot
Miten kattavia muuttujajoukkoja etsitään?

- Triviaali lähestymistapa: käydään läpi kaikki muuttujien osajoukot
- d muuttujaa $\Rightarrow 2^d$ muuttujaosajoukkoa
- Jos muuttujia on esim. 100, niin osajoukkoja on liikaa
- Täytyy siis toimia jotenkin nokkelammin
- *Perushavainto: jos muuttujajoukko $\{a_1, a_2, \ldots, a_k\}$ on kattava, niin kaikki sen osajoukot ovat kattavia.*
- *Kääntäen: $\{a_1, a_2, \ldots, a_k\}$ voi olla kattava vain jos jokainen sen osajoukko on kattava*. Osajoukkojen kattavuus on siis välttämätön muttei riittävä ehto.

Kattavat joukot
Esimerkki kattavien muuttujajoukkojen etsinnästä

- Oletetaan, että muuttujajoukot $\{a\}, \{b\}, \{c\}, \{e\}, \{f\}, \{g\}$ (joukon koko 1) esiintyvät tarpeeksi usein eli niistä on vähintään N havaintoa
- Silloin pari (i, j) on *mahdollisesti* kattava joukko, kun $i, j \in \{a, b, c, e, f, g\}$. Ehdokaspareja on $\binom{6}{2} = \frac{6!}{4! \cdot 2!} = 15$
- Oletetaan, että ehdokkaista vain parit $\{a, b\}$, $\{a, c\}$, $\{a, e\}$, $\{a, f\}$, $\{b, c\}$, $\{b, e\}$, $\{c, g\}$ ovat kattavia (koko 2) eli ne esiintyvät vähintään N kertaa yhdessä
- Luodaan kahden kokoisista kattavista kolmen kokoiset joukot. Pudotetaan pois ne joukot, joissa on mukana kahden kokoisia ei-kattavia joukkoja. Näin saadaan 3-kokoiset ehdokasjoukot.

Kattavat joukot (2)
Esimerkki kattavien muuttujajoukkojen etsinnästä

- Esimerkiksi kattavista muuttujapareista $\{a, b\}$ ja $\{a, f\}$ saadaan $\{a, b, f\}$, jonka alijoukko $\{b, f\}$ ei ole kattava, jolloin joukkoa $\{a, b, f\}$ ei hyväksytä ehdokkaaksi.

- Silloin $\{a, b, c\}$ ja $\{a, b, e\}$ ovat *ainoat mahdolliset* kolmen kokoiset kattavat joukot, koska niiden *kaikki* alijoukot ovat kattavia.
- Oletetaan, että $\{a, b, c\}$ kattava (koko 3).
- Lisäksi jo nyt tiedämme, että mitään neljän (tai isomman) kokoista kattavaa joukkoa *ei* voi olla, koska $\{a, b, c, e\}$:n alijoukoista mm. $\{a, c, e\}$ ei ole kattava. Toisin päin: ei-kattavaa joukkoa $\{a, c, e\}$ ei voi saada kattavaksi lisäämällä siihen alkioita.
- Kuten edellä jo pääteltiin, neljän kokoisia kandidaatteja ei synny, joten etsintä loppuu.

Kattavat joukot (3)
Esimerkki kattavien muuttujajoukkojen etsinnästä

- Luetellaan lopuksi kaikki kattavat joukot $\{\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$: $\{\{\}, \{a\}, \{b\}, \{c\}, \{e\}, \{f\}, \{g\}, \{a, b\}, \{a, c\}, \{a, e\}, \{a, f\}, \{b, c\}, \{b, e\}, \{c, g\}, \{a, b, c\}\}$
- Yllä “oletetaan” tarkoittaa siis, että “datasta on laskettu”. Tässä esimerkissä varsinaista dataa (datamatriisi **X**) ei ole annettu.

▸ Algoritmi pseudokoodina

Tasoittainen algoritmi
Periaate

- Yksinkertainen mutta tehokas algoritmin kattavien joukkojen etsintään on *tasoittainen algoritmi* (“levelwise algorithm”, “apriori”)
- Etsitään ensin yhden kokoiset kattavat joukot, sitten kahden jne.
- Hyödynnetään edellä tehtyä perushavaintoa: muuttujajoukko voi olla kattava vain jos sen kaikki osajoukot ovat kattavia.

Tasoittainen algoritmi
Joukon alkioiden lkm=1

- Kattavat joukot, jossa on yksi alkio: muuttuja a, jolla on vähintään N arvoa 1 datassa, ts. vähintään N havaintoa (saraketta) i, jolla $x(a, i) = 1$. Nämä on helppo etsiä lukemalla data kertaalleen läpi ja laskemalla kunkin muuttujan esiintymisfrekvenssi.

Tasoittainen algoritmi
Joukon alkioiden lkm=2

- Kun yhden kokoiset kattavat joukot tunnetaan, niin muodostetaan kahden kokoiset *ehdokasjoukot* (“candidate”)
- Muuttujajoukko $\{a, b\}$ on ehdokasjoukko, jos sekä $\{a\}$ että $\{b\}$ ovat kattavia
- Käydään data uudestaan läpi ja lasketaan kullekin ehdokasjoukolle kuinka monessa havainnossa on sekä a- että b-muuttujan kohdalla 1.
- Näin löydetään kahden kokoiset kattavat joukot.

Tasoittainen algoritmi
Joukon alkioiden lkm=k

- Oletetaan, että tunnetaan kaikki k:n kokoiset kattavat joukot; olkoon tämä kokoelma $\mathcal{C}_k$
- Muodostetaan kokoa $k+1$ olevat *ehdokasjoukot*: joukot, jotka saattaisivat olla kattavia
  - Ehdokasjoukko on sellainen muuttujajoukko, jossa on $k+1$ alkiota ja jonka kaikki osajoukot ovat jo kattavia. Riittää tarkastaa, että kaikki k alkiota sisältävät osajoukot ovat kattavia.
  - Otetaan kaikki joukkoparit $A = \{a_1, \ldots, a_k\}$ ja $B = \{b_1, \ldots, b_k\}$ kokoelmasta $\mathcal{C}_k$, lasketaan niiden yhdiste $A \cup B$, tarkistetaan, että yhdisteessä on $k+1$ alkiota ja että sen kaikki k:n alkion osajoukot ovat kattavia (ts. kuuluvat kokoelmaan $\mathcal{C}_k$).

Tasoittainen algoritmi (2)
Joukon alkioiden lkm=k

- Kun $k+1$:n kokoiset ehdokasjoukot on laskettu, niin käydään data läpi ja lasketaan kullakin ehdokasjoukolla, monessako sarakkeessa ehdokasjoukon kaikki muuttujat ovat =1.
- Näin löydetään kokoa $k+1$ olevat kattavat joukot.
- Algoritmia jatketaan kunnes uusia ehdokasjoukkoja ei enää löydy.

▸ Algoritmi pseudokoodina

NSF-esimerkki tasoittaisen algoritmin tuloksesta
Algoritmeja saatavilla

- Jatketaan saman NSF-datan kanssa
- Monia algoritmeja löytyy valmiina, esim. ▸ http://fimi.ua.ac.be/ tai ▸ http://adrem.ua.ac.be/~goethals/software/
- Tietokoneharjoituskierroksella T5 käytetään Bart Goethalsin `apriori`-ohjelmaa kassakuittien yleisimpien tuotteiden ja kurssin harjoitustyödokumenttien yleisten sanojen etsintään
- Laskuharjoituksessa H5/2 esitellään algoritmin laskennan kompleksisuutta

NSF-esimerkki tasoittaisen algoritmin tuloksesta

Kattavien joukkojen etsintä datajoukosta kynnysarvolla $N = 10000$

- Kynnysarvolla $N = 10000$ haku tuottaa yhteensä 250 kattavaa joukkoa aineiston 30800:sta muuttujasta (sanasta), jotka peräisin 128000 dokumentista
- 9 erilaista kolmen sanan joukkoa löytyy ainakin 10000 dokumentista

koko	ehdokkaita	kattavia joukkoja
1	30800	134
2	8911	107
3	79	9
4	0	0

NSF-esimerkki tasoittaisen algoritmin tuloksesta

Kattavien joukkojen etsintä datajoukosta kynnysarvolla $N = 2000$

- Pienemmällä kynnysarvolla $N = 2000$ kattavia joukkoja tulee yhteensä 16040 ja ehdokasjoukkoja vielä runsaasti enemmän
- Löytyy yksi kuuden sanan joukko, joka löytyy ainakin 2000 dokumentista

koko	ehdokkaita	kattavia joukkoja
1	30800	1171
2	685035	7862
3	105146	6098
4	5813	889
5	92	19
6	1	1

NSF-esimerkki tasoittaisen algoritmin tuloksesta

Kattavien joukkojen lukumäärä kynnysarvon N funktiona

- Kynnysarvo N: ykkönen (ykköset) täytyy löytyä vähintään N:stä sarakkeesta

kynnysarvo N	kattavia joukkoja
10000	250
5000	1539
2000	16040
1000	96223
⋮	⋮

NSF-esimerkki tasoittaisen algoritmin tuloksesta

Kuuden kokoinen kattava joukko kynnysarvolla $N = 2000$

- `egrep '21582|21614|23313|24416|26454|29137' words.txt`

```
21582 program
21614 project
23313 research
24416 science
26454 students
29137 university
```

Kattavat joukot
Mitä muuttujajoukon kattavuus merkitsee?

- Jos muuttujajoukko esiintyy usein, voiko se johtua sattumasta?
- Olkoot a ja b muuttujia, ja olkoon datassa n_a havaintoa, joissa a esiintyy ja n_b havaintoa, joissa b esiintyy
- Kaikkiaan n havaintoa
- $p(a = 1)$: todennäköisyys sille, että satunnaisesti valitussa havainnossa on $a = 1$
- $p(a = 1) = n_a/n$ ja $p(b = 1) = n_b/n$

Kattavat joukot
Riippumattomuus

- Kuinka monessa havainnossa a ja b esiintyvät yhtäaikaa, jos niiden oletetaan olevan riippumattomia?
- $p(a = 1 \wedge b = 1)$: todennäköisyys sille, että havainnossa on sekä $a = 1$ että $b = 1$ ($\wedge$ = ja)
- Jos havainnot riippumattomia niin:

$$\begin{aligned} p(a = 1 \wedge b = 1) &= p(a = 1)p(b = 1) \\ n_{ab}/n &= (n_a/n)(n_b/n) = n_a n_b/n^2 \end{aligned}$$

NSF-esimerkki riippumattomuudesta
Esimerkki riippumattomuudesta

- a = 'computational': $n_a = 7803$
- b = 'algorithms': $n_b = 6812$
- Riippumattomuus: 'computational' ja 'algorithms' esiintyvät satunnaisessa sarakkeessa todennäköisyydellä

$$\frac{n_a}{n}\frac{n_b}{n} = \frac{7803}{128000}\frac{6812}{128000} = .003244$$

- Odotusarvo tällaisten sarakkeiden lukumäärälle on

$$n\frac{n_a}{n}\frac{n_b}{n} = \frac{n_a n_b}{n} = 415$$

- Datassa tällaisia sarakkeita on $n_{ab} = 1974$ kappaletta.
- Merkitseekö tämä mitään? (Palataan tähän pian.)

NSF-esimerkki riippumattomuudesta
Ylimäärin esiintyvien parien etsintä awk-skriptinä

- Syötteenä NSF-datan sanojen ja sanaparien frekvenssit
- Käyttö: `cat bag-of-words.txt | ./this.awk`

```
#!/usr/bin/awk -f
BEGIN {N=128000}
{
if (NF==2) {
        f[$1]=$2;}
if (NF==3) {
        p[$1, $2] = $3;}
}
END {
        for (u in f) {
                for (v in f) {
                   if (p[u, v]>0) {
                        print u, v, f[u], f[v], p[u,v], p[u,v]/(f[u]*f[v]/N)
                   }
                }
    }
}
```

NSF-esimerkki riippumattomuudesta

Karkea analyysi millä pareilla $p(a=1\&b=1)/p(a=1)p(b=1)$ on suuri?

- Lasketaan suhde sanaparien todellisen esiintymisen ja sattumanvaraisen esiintymisen välillä

$$\frac{p(a=1\&b=1)}{p(a=1)p(b=1)} = \frac{n_{ab}/n}{(n_a/n)(n_b/n)} = \frac{n \cdot n_{ab}}{n_a \cdot n_b}$$

- Sarakkeet: (1) sanapari (a,b), (2) n_a, (3) n_b, (4) n_{ab}, (5) $\frac{p(a=1\&b=1)}{p(a=1)p(b=1)}$
- Näissä esimerkeissä suhde $\frac{p(a=1\&b=1)}{p(a=1)p(b=1)}$ on suuri:

```
fellowship postdoctoral 1458 3021 1028 29.8741
film thin               2275 3122 1431 25.789
dissertation doctoral   2272 2818 1135 22.6912
business innovation     4417 3514 2689 22.1754
polymer polymers        2768 2540 1208 21.9926
microscopy scanning     2925 2163 1079 21.8298
carolina north          1407 4687 1024 19.8756
poorly understood       2065 4100 1224 18.5049
...
```


NSF-esimerkki riippumattomuudesta

Muuttujapareja löytyy...

- Muutamia muuttujapareja esimerkkinä
- Pari nro 100

```
held workshop 5417 4890 1644 7.94409
```

- Pari nro 500

```
gene identify 4897 8450 1230 3.80477
```

- Pari nro 1000

```
engineering manufacturing 15201 3991 1334
2.81457
```


Kattavien joukkojen merkitsevyyden tutkiminen

- Ajatellaan datan generointia *toistokokeena* (kts. binomijakauma)
- Tehdään n havaintoa (NSF-data: $n = 128000$)
- Kullakin havainnon kohdalla "onnistutaan" (tuotetaan havainto, jossa on sekä $a = 1$ että $b = 1$ todennäköisyydellä $p = n_a n_b / n^2$)
- Onnistumisten lukumäärä on binomijakautunut $Bin(n,p)$ parametrein n ja $n_a n_b / n^2$
- Kuinka todennäköistä on, että saadaan vähintään n_{ab} onnistumista?
- Kuinka pieni on todennäköisyys, että saisimme odotusarvosta niin paljon poikkeavan tuloksen?
- Jos tämä todennäköisyys on pieni, niin löydetty hahmo kertoo jotakin kiinnostavaa.

Kattavien joukkojen merkitsevyyden tutkiminen

Binomijakauman hännän todennäköisyyden arviointi

- Binomijakauman hännän todennäköisyyden arviointi:

$$\sum_{i=n_{ab}+1}^{n} B(n,i,p),$$

missä $B(n,i,p)$ on todennäköisyys, että n:n toiston kokeessa saadaan tasan i onnistumista, kun onnistumisen todennäköisyys on p:

$$B(n,i,p) = \binom{n}{i} p^i (1-p)^{n-i}.$$

- Perinteiset approksimaatiot (normaaliapproksimaatio, Poisson-approksimaatio) eivät välttämättä suoraan sovi tähän (todennäköisyys p on pieni, mutta onnistumisten määrä on suuri)

Kattavien joukkojen merkitsevyyden tutkiminen
Chernoffin raja

- Olkoon X onnistumisten lukumäärä toistokokeessa, jossa n koetta ja todennäköisyys onnistumiselle on p. Tällöin onnistumisen lukumäärän odotusarvo on np.
- Todennäköisyys $Pr(X > (1+\delta)np)$ on rajoitettu (Chernoffin raja):

$$Pr(X > (1+\delta)np) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np} < 2^{-\delta np}.$$

- (Katso lisää paperilaskareista H5/2)

NSF-esimerkki kattavan sanaparin merkitsevyydestä
Simulointitestaus

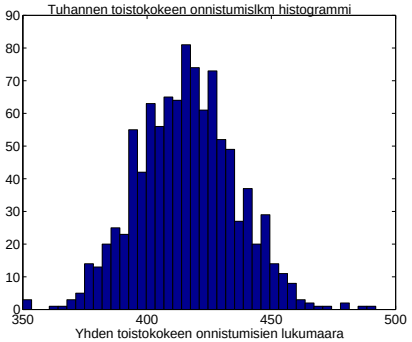
- NSF-data:
  - a = 'computational': $n_a = 7803$
  - b = 'algorithms': $n_b = 6812$
  - yhteiset havainnot $n_{ab} = 1974$
  - riippumattomuusoletuksella laskettu $p(a=1)p(b=1) = n_a n_b / n^2 = 0.003244$
- Satunnaistus: kokeillaan miten käy kun heitetään rahaa $n = 128000$ kertaa
- Onnistumistodennäköisyys $p = 0.003244$
- Lasketaan onnistumisten määrä

NSF-esimerkki kattavan sanaparin merkitsevyydestä
Simulointitestaus Matlab-ohjelmana

```
p = 0.003244;
count=zeros(10,1);     % alustus
for a=1:10
   for i=1:128000      % hidas tapa
      if rand(1,1)<p
         count(a)=count(a)+1;
      end
   end
end

count=zeros(1000,1);  % alustus
for a=1:1000          % nopeampi tapa Matlabissa
   count(a) = sum(rand(128000,1)<p);
end
hist(count,40);
```


NSF-esimerkki kattavan sanaparin merkitsevyydestä
Simuloinnin tulos



Todennäköisyys saada 1974 onnistumista lienee siis pieni; mutta kuinka pieni?

NSF-esimerkki kattavan sanaparin merkitsevyydestä
Verrataan Chernoffin rajaan

- Toistokoe $X \sim Bin(n, p)$, jossa onnistumisen lukumäärän odotusarvo on np

$$Pr(X > (1+\delta)np) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np} < 2^{-\delta np}.$$

- Nyt $p = .003244$, odotusarvo $np = 415$, havaittu arvo $X = 1974$.
- $X > (1+\delta)np$ eli $\delta = (X - np)/np = 3.757$
- Todennäköisyys on varsin pieni:

$$\begin{aligned} Pr(X > 1974) &= Pr(X > (1+3.757)415) \\ &< (\frac{42.8}{1667})^{415} \approx 0.0257^{415} < 10^{-660} \end{aligned}$$

- $2^{-\delta np} \approx 2^{-3.757*415} \approx 10^{-469}$

Moninkertaisen testauksen ongelma

- Oletetaan, että löydetään 100 kiinnostavaa muuttujaparia (a, b), ja tutkitaan kaikilla pareilla onko n_{ab} merkittävästi suurempi (tai pienempi) kuin riippumattomuusoletuksen antaman arvio $n_a n_b / n$.
- Saadaan tulokseksi esim., että parilla (a, b) näin suuri ero esiintyy sattumalta todennäköisyydellä 0.01.
- Todennäköisyys, että 100 parista jollakin tulee tällainen tulos on

$$1 - (1 - 0.01)^{100} \approx 0.63$$

(miksi?)

- Tutkittaessa usean hahmon esiintymäfrekvenssien satunnaisuutta pitää ottaa huomioon se, että tarkastelemme useita hahmoja.

Moninkertaisen testauksen ongelma (2)

- Bonferroni-korjaus: jos tutkitaan M:n hahmon esiintymistodennäköisyyksiä, niin sattumalta esiintymisen todennäköisyydet tulee kertoa M:llä.
- Edellä (NSF-esimerkki) yksittäisen hahmon todennäköisyys oli niin pieni, että korjaus ei muuta asiaa.
- Suurin osa löydetyistä muuttujapareista esiintyy niin usein yhdessä, että tämän tapahtuman esiintyminen sattumalta on erittäin epätodennäköistä.

Entä suuremmat kattavat joukot?

- Oletetaan, että muuttujajoukko $\{a, b, c\}$ esiintyy datassa n_{abc} kertaa (ts. datassa on n_{abc} saraketta, joilla muuttujat a, b ja c ovat kaikki 1.
- Onko tämä enemmän kuin satunnaisessa tapauksessa voisi olettaa?
- Mikä on satunnainen tapaus tässä tilanteessa?
- Esim. $n = 1000$ ja $n_a = n_b = n_c = 500$. Jos muuttujat ovat riippumattomia, niin n_{abc} on luultavasti noin 125.
- Jos $n_{abc} = 250$, niin $\{a, b, c\}$ on jotenkin epätavallinen kokoelma.
- Tämä voi johtua siitä, että esim. a ja b ovat vahvasti toisistaan riippuvia. (Jos $n_{ab} = 500$, niin sen jälkeen $n_{abc} = 250$ on odotettavaa.)

Entä suuremmat kattavat joukot? (2)

- Mikä on sovelias nollahypoteesi? Esim. ns. maksimientropiahypoteesi.

Miten vaikeaa kattavien joukkojen etsintä on?

- Annettuna 0-1 data ja kynnysarvo N
- Etsittävä kaikki kattavat joukot
- Vastaus voi olla hyvin suuri
- Käytännössä tasoittainen algoritmi on riittävän tehokas

Muunlaisten usein esiintyvien hahmojen etsintä

- Samoja periaatteita voidaan käyttää myös muihin tehtäviin
  - Usein esiintyvien episodien etsintä
  - Usein esiintyvien aliverkkojen etsintä
  - Usein esiintyvien alipuiden etsintä
- Sama tasoittaisen algoritmin perusidea toimii

Yhteenveto

Tässä luvussa esiteltiin erityyppisen diskreetin aineston muuttamista 0-1-dataksi. Tätä voidaan käsitellä ja analysoida tehokkaasti mm. tasoittaisella algoritmilla.

NSF-data
Esimerkki abstraktin kentistä

Takaisin kalvoihin

```
Title       : Mathematical Sciences: Structure and Rigidity of Graphs with Applications to
               Network Models of Materials
Type        : Award
NSF Org     : DMS
Latest
Amendment
Date        : July 5,  1994
File        : a9412017

Award Number: 9412017
Award Instr.: Standard Grant
Prgm Manager: Michael H. Steuerwalt
              DMS  DIVISION OF MATHEMATICAL SCIENCES
              MPS  DIRECT FOR MATHEMATICAL & PHYSICAL SCIEN
Start Date  : July 1,  1994
Expires     : June 30,  1998       (Estimated)
Expected
Total Amt.  : $67877             (Estimated)
Investigator: Deborah S. Franzblau   (Principal Investigator current)
Sponsor     : Rutgers Univ New Brunswick
              ASB III, 3 Rutgers Plaza
              New Brunswick, NJ  08901    732/932-0150

NSF Program : 1271      COMPUTATIONAL MATHEMATICS
Fld Applictn: 0000099   Other Applications NEC
              21        Mathematics
Program Ref : 9161,9162,9263,AMPP,
```


NSF-data
Esimerkki abstraktista

Takaisin kalvoihin

```
Abstract    :
   9412017  Franzblau       In this project, the principal investigator aims to
   obtain  the following results: (1) an implementation of a practical
   combinatorial algorithm to compute bounds on degrees of freedom  of a network
   (in 3 or more dimensions), (2) new combinatorial  conditions for rigidity or
   simple formulas for degrees of freedom  in families of graphs, (3) new, easily
   computed measures of  medium-range order in network models of glasses.  The
   methods  employed include those of combinatorial optimization, graph  theory,
   and discrete algorithm design.  Problems on network  models are addressed
   largely through computer experiments, and  make use of algorithms developed and
   implemented by the  investigator.       This project has two related aims, and
   is intended to  contribute new results both to mathematics and materials
   science.  The first aim is to address key open problems in the mathematical
   theory of rigidity, including computing the degrees of freedom of  a network
   (also called a graph).  This theory has a long history,  which includes the
   work of Maxwell (1864) on determining whether  a "scaffold" made of rigid bars
   and movable joints is itself  rigid. The second aim is to address basic issues
   on network  models of solids; such models (also called ball-and-stick  models),
   in which points represent atoms and connections between  points represent
   chemical bonds, are often studied to better  understand the properties of both
   crystalline solids and glassy  materials.  One focus is to characterize the
   relationship between  the structure of a network model and its rigidity, and
   the other  is to find useful measures which capture this network structure.
   The work therefore leads to a better understanding of materials  properties.
```


Tasoittainen algoritmi

1: $k \leftarrow 1$ ▷ Init level $k = |X|$
2: $C_k \leftarrow \{ \{ a \} \mid a \in \text{variables} \}$ ▷ Init candidate sets C_1
3: **while** $C_k \neq \emptyset$ **do**
4: counter[X] $\leftarrow$ 0 for all X ▷ Init counters for each candidate
5: **for** observation in data **do** ▷ Count frequencies of candidates
6: **for** X in C_k **do** ▷ Check if all variables in X are present
7: good $\leftarrow$ True
8: **for** var in X **do**
9: **if** observation[var] = 0 **then**
10: good $\leftarrow$ False
11: **if** good **then**
12: counter[X] $\leftarrow$ counter[X] + 1
13: $F_k \leftarrow \emptyset$ ▷ Init frequent sets F_k at level k
14: **for** X in C_k **do** ▷ Select frequent candidates
15: **if** counter[X] $\geq N$ **then** ▷ Threshold N
16: $F_k \leftarrow F_k \cup \{ X \}$

Tasoittainen algoritmi (2)

17: $C_{k+1} \leftarrow \emptyset$ ▷ Init candidate sets C_{k+1} at level $k+1$
18: **for** A in F_k **do** ▷ Generate next candidates
19: **for** B in F_k **do**
20: $X \leftarrow A \cup B$
21: **if** $|X| = k + 1$ **then**
22: good $\leftarrow$ True
23: **for** var in X **do** ▷ Ignore sets with nonfreq. subsets
24: **if** $X \setminus \{$ var $\}$ not in F_k **then**
25: good $\leftarrow$ False
26: **if** good **then**
27: $C_{k+1} \leftarrow C_{k+1} \cup \{ X \}$
28: $k \leftarrow k + 1$ ▷ Return to row 3

koodi: Jouni Seppänen.

Takaisin kalvoihin: esimerkki

Takaisin kalvoihin: algoritmin kuvaus