Yleistä asiaa

1. Haluaisitteko työkalujen demosession?
2. Haluaisitteko tehdä itse alustavan arvostelun harjoitustyöllenne?
   Kohdittain pisteet ja töiden jakautuminen ryhmässä, esim.
   - Datan ymmärrys (+perusanalyysi muuttujien arvoista)
   - Esiprosessointi
   - Mallinnus (voi jakaa osiin, jos useita menetelmiä)
   - Tulosten tulkinta
   - Raportointi

Luento 7: ≈ Aputekniikoita harjoitustöihin

1. Tekstidatan kaivelu
   - yleiskuva menetelmistä (DM:n lähialoista)
   - tekstin klusterointi sanojen perusteella
   - HITS-algoritmi (alkujaan internetin kaiveluun)
2. Temporaalisen datan mallinnus
   - erityisesti aikasarjadata!
   - trendien ja periodisuuden tunnistus
   - piirteiden eristys kiihtyvyysdatasta

Tekstidatan kaivelu

Tällä kurssilla: Miten tekstidatasta saisi piirteitä tai ryhmittelyjä helposti? VS.

- Tiedon eristäminen tekstistä (information extraction from text, IET): Halutaan eristää tekstistä automaattisesti vastaukset kysymyksiin tai tiettyjen muuttujien arvot.
- Luonnollisen kielen prosessointi: halutaan ymmärtää tekstin sisältö (merkitys)
- “Tiedon haku” (Information retrieval): halutaan löytää suuresta tietokannasta (esim. internet) kyselyn kannalta relevantit dokumentit

[EXTRA] Esimerkki (IET): Tiivistä tiedot uutisartikkelista

19 March – a bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT)

[EXTRA] Esimerkki (IET): Tiivistä tiedot

- Missä? (El Salvador)
- Milloin? (19 Mar)
- Miten? (bombing)
- Ketkä? (urban querilla commandos)
- Kohde? (power tower)
- Uhrit? (no injury or death)

[EXTRA] Monia vaiheita, ongelmaspesifejä menetelmiä

- Erota virkkeet (välimerkkien perusteella)
- Erota sanat (+mahd. luvut)
- Esitä sanan vartalo (stemming) → saman sanan esiintymät
- Tunnista substantiivilausekkeet (noun phrase) "Tämä virke sisältää kaksi substantiivilauseketta."
- (Lisähankaluudet: yhdyssanat, erisnimet, pronominien viittaussuhteet)
- Tunnista semanttiset suhteet (yleensä vain tietyn tyyppiset relaatiot) "SVM:ien keksijä Vladimir Vapnik" eli Keksijä(SVM,Vapnik)

→ **Mitä saisi selville pelkästään sanoista?**

Tiedon kaivuu annettuna vain sanat

- dokumettien klusterointi sanojen perusteella
- samankaltaisten dokumenttien haku (implisiittinen klusterointi: samanlaiset vs. erilaiset)
- sanojen väliset riippuvuussuhteet →
  - käsitekartat
  - moniosaisten ilmaisujen tunnistus

Sanavartaloiden erottelu

- Mitä sanoja/sanaryhmiä (englannin "yhdyssanat") tekstissä esiintyy?
- suomen kielen erityisongelma: taivutus – Mikä on sanan vartalo?
  - sanastosta tarkistus (jos saatavilla)
  - heuristiset "stemmerit" (esim. Wekaan saa Snowball-paketissa, mutta suomenkielinen huono)
- Idea: Löytyisivätkö vartalot sanoja klusteroimalla?
  - etäisyysmittana minimum edit distance
  - esim. d(internetissä, internetin)=3 vs. d(kissa,viisas)=3 → toimii paremmin pitkillä sanoilla

Dokumenttien klusterointi sanojen perusteella

Ongelma: Klusteroi dokumentit $D = \{d_1, \ldots, d_n\}$, joissa esiintyvät sanat (sanavartalot) $w_1, \ldots, w_k$.

- Karsitaan pois usein toistuvat epäinformatiiviset sanat (stopwords), kuten apunsanat (ja, tai, sekä, hän, se)
- Karsitaan liian harvinaiset sanat (parhaita tosin sellaiset, jotka esiintyvät usein mutta vain osassa dokumentteja)
- Dokumentti esitetään vektorina $(fr(w_1), \ldots fr(w_k))$, missä $fr(w_i)$=sanan w_i frekvenssi kyseisessä dokumentissa

Dokumenttien klusterointi (jatk.)

- Tai $(fr(w_1) \cdot idf(w_i), \ldots, fr(w_k) \cdot idf(w_k))$, missä idf (inverse document frequency) painottaa sanan yleisyydellä

$$idf(w_i) = \log \frac{|D|}{|\{d \in D | w_i \in d\}|}$$

- Lopuksi kannattaa normalisoida vektorit jakamalla vektorin pituudella (jolloin saadaan $||d_i|| = 1$)

Dokumenttien klusterointi (jatk.)

- Yleisin etäisyysmitta kosinimitta

$$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{||d_1||\,||d_2||}$$

- dokumenttiklusterin centroidi on vektori $c = (c_1, \ldots, c_k)$, missä c_i on keskiarvo dokumenttivektorien vastaavista painoista
- centroidien etäisyys tai centroidin ja dokumentin etäisyys lasketaan samalla kosinimitalla
- Klusteroidaan halutulla menetelmällä (usein K-means, koska nopea)
- Ongelma: Jos pieniä dokumentteja, vähän yhteisiä sanoja ja toimii huonosti!

Pienien tekstipätkien klusterointi

- Dokumentin esityksessä riittää k:n bitin vektori: esiintyykö sana dokumentissa vai ei?
- Vaihtoehtoinen mitta Jackard:

$$Jackard(d_1, d_2) = 1 - \frac{\#(d_1\text{:ssä ja } d_2\text{:ssa esiintyvät sanat})}{\#(d_1\text{:ssä tai } d_2\text{:ssa esiintyvät sanat})} =$$

$$1 - \frac{|\{w_i | w_i \in d_1 \wedge w_i \in d_2\}|}{|\{w_i | w_i \in d_1 \vee w_i \in d_2\}|}$$

Pienien tekstipätkien klusterointi

Samankaltaisuusverkon ositusmenetelmät (spektraaliklusterointi sukulaisineen) saattavat tuottaa parempia tuloksia!

- **Affinity propagation** – uusi kehuttu menetelmä!
- samankaltaisuusmitaksi dokumenttien välinen Mutual Information
- HITS-algoritmi – helppo ja monikäyttöinen

HITS-algoritmi

Hyperlink-Induced Topic Search (=Hubs and Authorities)

- Alkuperäinen idea web-dokumentien hakemiseen:
  - Hub=sivu joka viittaa moniin autoratiivisiin sivuihin
  - Authority=autoratiivinen sivu johon monet hubit viittaavat
- pävitetään sivujen hub- ja authority-arvoja (h- ja a-vektorit), kunnes systeemi kovergoi
- samalla lasketaan (huomaamatta) vastaavien matriisien prinsipaaliominaisvektorit

[EXTRA] Hits-algoritmi

Olk. $G = (V, E)$ verkko, jonka jokaiseen solmuun v liittyy kaksi painoa $h(v)$ ja $a(v)$.

1. Alusta h ja a tasaisesti s.e. $\sum_i h(v_i)^2 = \sum_i a(v_i)^2 = 1$
2. Kunnes h- ja a-vektorit konvergoivat, päivitä kaikilla v_i:
   - $h(v_i) = \sum_{(v_i,w)\in E} a(w)$ (v_i:n osoittamien auktoriteettien arvojen summa)
   - $a(v_i) = \sum_{(w,v_i)\in E} h(w)$ (v_i:hin osoittavien hubien arvojen summa)
   - Normalisoi painot: $h(v_i) = \frac{h(v_i)}{H}$ ja $a(v_i) = \frac{a(v_i)}{A}$, missä $H = \sqrt{\sum_i h(v_i)^2}$ ja $A = \sqrt{\sum_i a(v_i)^2}$
3. Lopussa parhailla auktoriteeteilla on suurimmat a-arvot ja parhailla hubeilla suurimmat h-arvot

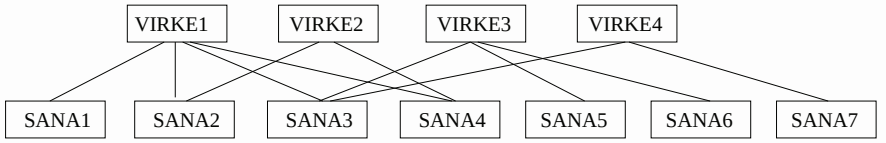
Samankaltaisten sivujen haku

Ongelma: Annettu sivu p, etsi samankaltaisia sivuja.

- Muodosta verkko seuraavista sivuista: kaikki sivut joihin p osoittaa ja osa (tai kaikki), jotka osoittavat p:hen
- Alusta sivujen h ja a-arvot, esim. $h(p) = |(p, q)|$ ja $a(p) = |(q, p)|$ + normalisointi
- Suorita $HITS$
- Lopussa sivu jolla suurin a-arvo on samankaltaisin

Sovellus sanoille ja virkkeille

Annettuna avainsana tai -sanoja, muodostetaan niihin liittyvä 2-jakoinen verkko:



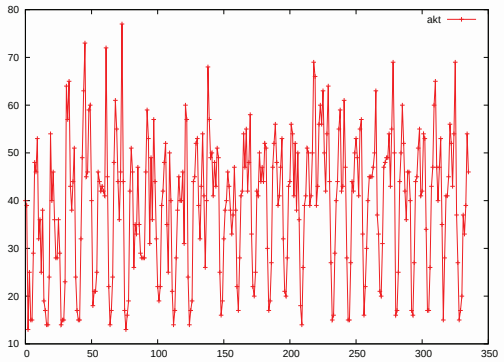
- Virkkeille lasketaan V-painoja ja sanoille S-painoja
- Alustetaan tasaisesti s.e. neliöiden summa 1
- Päivityssääntö: $V(v) = \sum_{w_i \in s} S(w_i)$ ja $S(w) = \sum_{w \in s_i} V(s_i)$ + normalisointi
- Lopussa suurimman S-painon saaneet sanat samankaltaisimpia (ja suurimman V-painon saaneet virkkeet)

Tuloksia

- Testattu suomenkielisillä kissa-aforismeilla ja -tarinoilla
- ei mitään vartalon etsintää tai stopwordin poistoa (olisi kannattanut)
- sanan “kissa” kanssa samankaltaisia mm. koira, ihminen (molemmat eri taivutusmuodoissa), runous, mukava, kotieläin
- sanan “rakastavat” kanssa samanlaisia mm. ruokkivat, pitävät, huolta, täytyy, jumalia
- homonyymeillä voi kiusata ohjelmaa (“Lassie palaa kotiin” vs. “Tuli palaa takassa”)
- lisäksi paljon roskaa! (erit. stopwordeja) → toimisi varmasti paremmin tekstin esiprosessoinnin jälkeen

Aikasarjadatan analyysiä

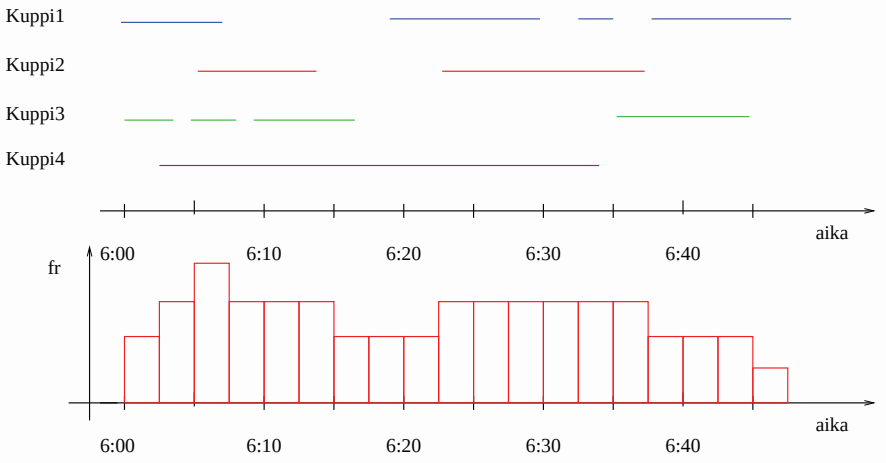
Aikasarja (time series): Jonkin muuuttujan X arvot eri ajanhetkillä. (Kuvassa lehmän aktiivisuus)



Aikasarjadata

- Yleensä X mitattu tasavälein jollain tiheydellä= ”resoluutiolla” (esim. 10 kertaa/s = 10Hz, tai kerran tunnissa)
  - resoluutio=erotustarkkuus
  - tyypillisesti mittaus on ko. mittausvälin keskiarvo
- Lokidatan tapahtumissa mittauksia vain kun jotain tapahtui
- Voi silti muodostaa tasavälisen aikasarjadatan, kun esim. summaa X:n arvot tietyllä aikaikkunalla (esim. minuuteittain, tunneittain)

Esimerkki: syöntifrekvenssi automaateittain



Aikasarjadataa koskevia ongelmia

- Onko nousevaa tai laskevaa **trendiä**?
- Löytyykö säännönmukaisuutta?
  - **periodisuus** (jaksollisuus tai syklisyys): sama hahmo toistuu tasaisin aikavälein uudestaan
  - esim. ulkolämpötilan vaihtelee vuoden sykleissä
  - elollisten käyttäytyminen vuorokauden sykleissä
  - tasaisen kävelyn kiihtyvyyksissä esim. sekunnin sykli
- Millaiset piirteet luonnehtivat luokan C aikasarjoja? → **piirteiden eristys**
- Löytyykö aikasarjasta s lyhyempi aikasarja v? → **aikasarjan sovitus** (subsequence matching)

Ensin kannattaa suodattaa kohinaa

- Liukuva keskiarvo muuntaa x_i:n x'_i:ksi:

$$x'_i = \sum_{j=-m}^{M} w_j x_{i+j}$$

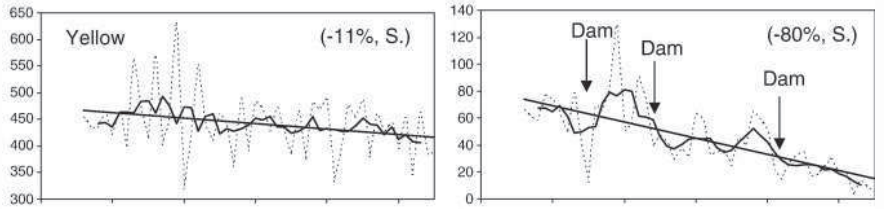
  eli painotettu keskiarvo pisteistä $x_{i-m}, x_{i-m+1}, \ldots, x_i, \ldots, x_{i+M}$
- voi olla m=M tai $M = 0$ (vain edelliset pisteet)
- w_j voi olla tasainen (jos pieni ympäristö) tai painottaa kaukaisia naapureita vähemmän
- mediaanisuodatin palauttaa tutkitun välin mediaanin → suodattaa häiriöpiikkejä (voivat olla kiintoisia!)
- joskus vaikea erottaa kohinaa “oikeasta” signaalista!

Trendi

- pitkän aikavälin muutos keskiarvossa tai mediaanissa
- viittaa yleensä lineaariseen muutokseen, mutta voi olla esim. polynomi (joskus jopa termi “syklinen trendi”)
- lineaarinen trendi löytyy helposti yhden muuttujan lineaariregressiolla!
- samalla idealla (pienimmän neliösumman minimointi) voi sovittaa muitakin funktioita
- trendi kannattaa poistaa ennen periodisuuden tutkimista!

Trendiesimerkki

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

Periodisuuden tunnistus

Ongelma: Onko periodeja? Millä aikavälillä toistuvat?

- Helppo tekniikka: **analysoidaan autokorrelaatiot**
  - tavallisen Pearsonin korrelaation sovellus aikasarjalle
- Hieman hankalampaa: **spektraalianalyysi**
  - analysoidaan signaalin “spektriä” eli mistä taajuuskomponenteista se koostuu
  - esim. Fourier- ja Wavelet-analyysi
  - ei toimi aina (herkkä kohinalle tai ulkopuolisille?)

Autokorrelaatio ja korrelogrammit

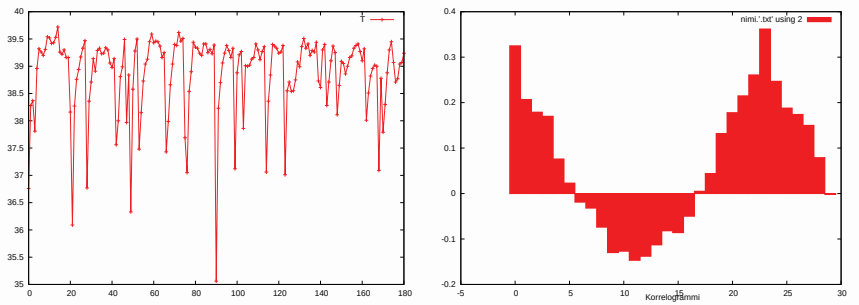
Olkoon aika sarja $x_1, \ldots, x_n$. Yksinkertaistus: Ol. että X:n keskiarvo ja varianssi suurinpiirtein vakiot. Estimaatit: $\mu_X = \frac{1}{n}\sum_i x_i$ ja $\sigma_X^2 = \frac{1}{n}\sum_i (x_i - \mu_X)^2$

- Idea: Tutkitaan Δ:n päässä toisistaan olevien arvojen korrelaatiota

$$R(\Delta) = \frac{\sum_{i=1}^{n-\Delta}(x_i - \mu_X)(x_{i+\Delta-\mu_X})}{\sigma_X^2}$$

- kannattaa tutkia systemaattisesti eri Δ:n arvoilla (jos periodia ei arvata)
- tulokset voi esittää havainnollisesti **korrelogrammina**

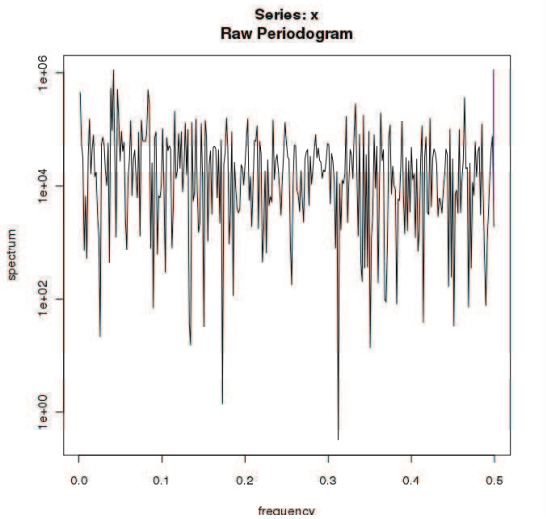
Korrelogrammi: autokorrelaatiot Δ:n funktiona



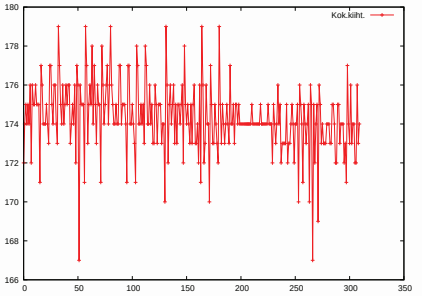
Vasemmalla pötsin lämpötila tunneittain ja oikealla korrelogrammi.

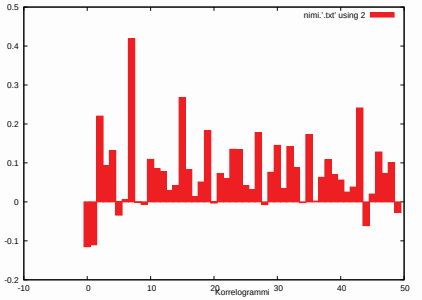
[Ekstra] Miten FT?

Ei selkeitä piikkejä – liian epäsäännöllinen data?

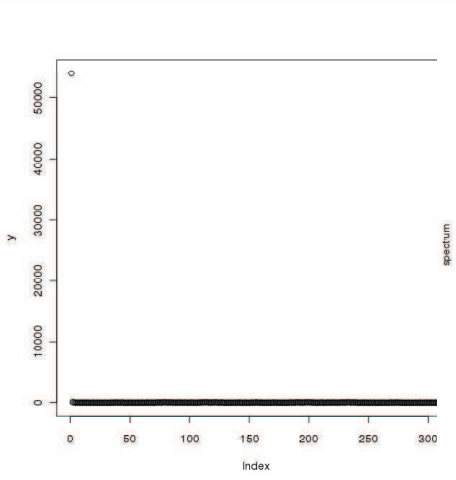


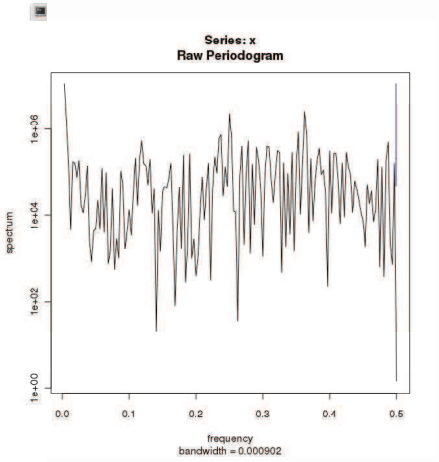
[Extra] Esim.2: Sama märehtimiselle (kiihtyvyysdata)





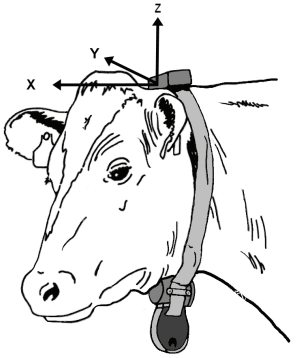
[Extra] Esim. 2: Miten FT?





Piirteiden eristys: esimerkkinä kiihtyvyys

Ongelma: Datasta on kerättyä näytepaloja, esim. 10s pätkiä, joilta on kiihtyvyydet ja joiden aikainen toiminto (luokka) tiedetään. Eristä piirteitä joilla eri toimintoihin liittyvät sarjat voisi erottaa toisistaan!



Tavallisia piirteitä

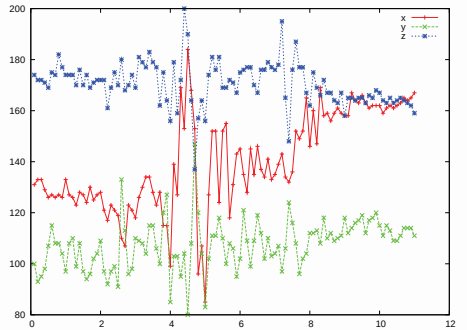
Kukin “data-alkio” on siis osasarja $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$. Ajanhetkellä i kokonaiskiihtyvyys (sen magnitudi)

$$kok = \sqrt{x_i^2 + y_i^2 + z_i^2}.$$

- x:n, y:n ja z:n sekä kokonaiskiihtyvyyden keskiarvo, **keskihajonta**, minimi, maksimi, “energia” (neliösumma)
- joskus myös vinouma (skew) ja kurtoosi (“huipukkuus”, kurtosis)
- jos pätkässä tarpeeksi pisteitä, myös xy-, xz- ja yz-korrelaatiot voivat olla hyviä (mutta herkkiä ulkopuolisille)
- Fourier-analyysistä ensimmäiset komponentit (jos tasaisen periodinen liike)

Heuristisia piirteitä: piikit

- Piikkien lukumäärä, keskimääräinen tai maksimaalinen korkeus, korkeus/leveys-suhde, jne.
- Hyviä kun raju epäsäännöllinen liike!
- Voi tutkia myös muunnetusta signaalista

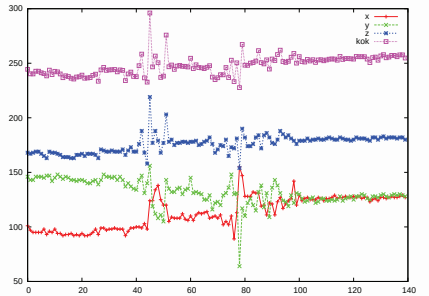


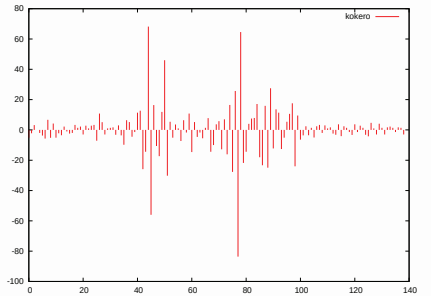
Lehmä nousee maakuulta

Muunnettu signaali: kiihtyvyyden muutokset

- Usein informatiivisempaa tutkia kiihtyvyyden magnitudin ja suunnan muutoksia!
- Kokonaiskiihtyvyyden magnitudierot mittarin orientaatiosta riippumattomia!
- kulmaerot vaativat origon tietämistä (arvio riittää)
- kannattaa tutkia myös xy-, xz- ja yz-tasoissa
- muunnetun signaalin “aaltojen” amplitudi ja frekvenssi sekä piikit hyviä piirteitä

Esimerkki: laskeutumisen kokonaiskiihtyvyyden magnitudierot





Tapahtumadatan mallinnusmenetelmiä

Aikasarja muotoa $(t_1, A_1), \ldots, (t_n, A_n)$, missä A_i:t tapahtumia (kategorisia muuttujia)

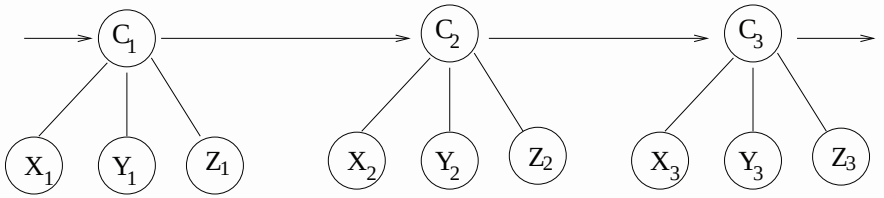
- Markov-ketjut
  - $\approx$ probabilistisia tilakoneita
  - esim. Jos Mansikki juo nyt, millä todennäköisyydellä Mansikki alkaa nuolla seuraavaksi?
  - Ongelma: riippuu siitä, monestiko on jo toistanut juo–nuolee-sykliään

Tapahtumadatan mallinnusmenetelmiä

- Episodit
  - Usein toistuvia tapahtumahahmoja, esim. tapahtumat A ja B esiintyvät peräkkäin tai “rinnakkain”
  - Ongelma: perustuu vain frekvensseihin – onko mitään tilastollista relevanssia?
  - Haku edellyttää että hahmon templaatti on valmiiksi määritelty

HMM:t ja dynaamiset Bayes-verkot (ennustukseen)

Kuvassa dynaaminen naiivi-Bayes-verkko



esim. C= toiminto ja X, Y ja Z kiihtyvyysdatasta eristettyjä piirteitä. Voidaan hyödyntää tietoa edellisestä toiminnosta!

Vinkkejä

- Piirrä (plottaa) paljon kuvia! (usein “puolet” mallinnusta)
- Periodisten hahmojen löytäminen vaatii paljon dataa!
  - esim. eläimillä periodi yleensä vuorokausi eli oltava paljon vuorokausia joka yksilöltä
- Tarjolla monia tutkimushaasteita, esim.
  - tilastotieteellisesti soundi (ja yleisempi) versio episodeista
  - temporaalisen datan louhinta riippuvuussäännöillä (mukaan ajallista kontekstia kuvaavat muuttujat)