Luku 2. Datasta tietoon: mitä dataa? mitä tietoa?

T-61.2010 Datasta tietoon, syksy 2011

professori Erkki Oja

Tietojenkäsittelytieteen laitos, Aalto-yliopisto

31.10.2011

Tämän luennon sisältö

Data-analyysin ongelma

- Tulevien vuosien valtava haaste on digitaalisessa muodossa talletetun datan kasvava määrä
- Arvioita:
  - Yhdysvaltojen kongressin kirjasto Washingtonissa: 32 miljoonaa kirjaa ja lehteä, 3 miljoonaa äänitettä, 14.7 miljoonaa valokuvaa, 5.3 miljoonaa karttaa, 61 miljoonaa käsikirjoitusta. Kerätty *200 vuoden aikana*. Nyt sama datamäärä kertyy levyille *joka 15. minuutti* (noin 100 kertaa vuorokaudessa).
  - Tämä on 5 exatavua vuodessa. (Kertaus: Exatavu = 2^{60} tavua = 1,152,921,504,606,846,976 tavua $\approx 1.15 \times 10^{18}$ (triljoona) tavua).
  - Sama määrä tulisi, jos *kaikki ihmispuhe kaikkina aikoina* (n. 100.000 vuotta) koodattaisiin sanoiksi ja digitoitaisiin (R. Williams, CalTech).

Data-analyysin ongelma (2)

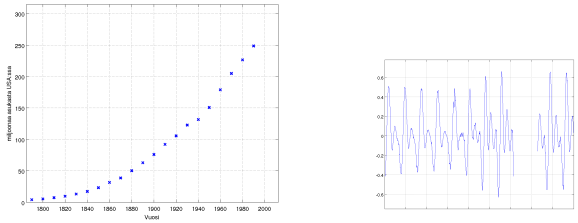
- Aiemmin talletettu data oli lähinnä tekstiä ja numerodataa (taloudellishallinnollinen IT), mutta nyt yhä enemmän ns. reaalimaailman dataa (digitaaliset kuvat, videot, äänet, puhe, mittaustiedot, bioinformatiikan tietopankit jne.)
- Lopulta mikä tahansa tieto josta voi olla hyötyä saattaa tulla digitaalisesti haettavaksi, esim. Webin kautta
- Tämä asettaa suuria haasteita tallennus- ja tietokantatekniikoille
- Eräs keskeinen kysymys: *kuinka haluttu tieto (information, knowledge) löytyy?* Tarvitaan jonkinlaisia “älykkäitä” datan analyysi-, louhinta- ja hakumenetelmiä.

Mallit ja oppiminen

- Peruslähtökohta data-analyysille on datan *mallitus*
- Malli tarjoaa tiivistetyn *esitystavan (representaation)* datalle
- Mallin perusteella on paljon helpompi *tehdä päätelmiä* kuin raakadatasta

Mallit ja oppiminen (2)

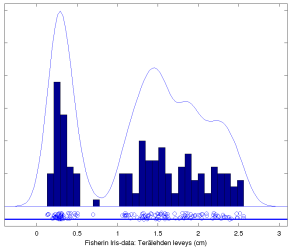
- Esimerkki: aikasarjan ennustaminen



Kuva: (a) Yhdysvaltain asukasmäärä 1790-1990. Onko kasvu jatkunut? (b) Äänisignaalin aaltomuoto. Miten täytetään puuttuva kohta?

Mallit ja oppiminen (3)

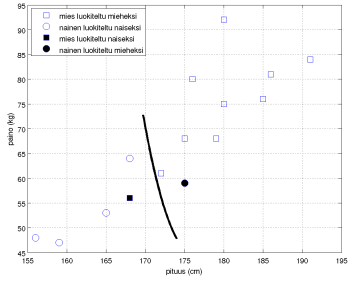
- Toinen esimerkki: datan todennäköisyysjakauma



Kuva: Datapisteet, histogrammi ja estimoitu jakauma. Fisherin kuuluisa datajoukko kolmesta liljalajista. Terälehden leveys mitattu 50 yksilöstä kolmesta lajista. Yksi laji erottuu, kaksi menee “sekaisin”. HUOM! näytteet ripoteltu y-akselin suunnassa.

Mallit ja oppiminen (4)

- Kolmas esimerkki: luokitus

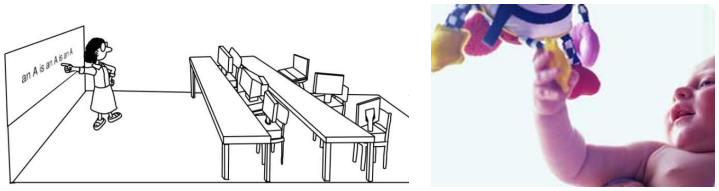


Kuva: Pituus- ja painohavainnot 15 ihmisestä. Laskettu luokitin, joka jakaa tason kahteen osaan (vain osa rajasta piirretty). Kun tunnetaan luokat, tiedetään, että luokitin tuottaa kahdessa tapauksessa virheellisen luokituksen.

Mallit ja oppiminen (5)

- Mistä malli sitten löytyy?
- Joskus voidaan käyttää olemassaolevaa tietoa (fysikaalisia luonnonlakeja, inhimillistä kokemusta, tms)
- Usein kuitenkin joudutaan käyttämään *tilastollisia malleja*, jotka muodostetaan suoraan datan perusteella (kuten edellisissä esimerkeissä)
- Kurssi “datasta tietoon” (ainakin alkuosa) käsittelee tilastollisia malleja ja niiden johtamista datajoukoista.
- Usein mallin automaattista muodostamista datajoukosta kutsutaan *koneoppimiseksi*

Mallit ja oppiminen (6)



Kuva: Koneoppimista vs ihmisen oppiminen.

- Sana tulee ihmisen oppimisesta, joka myös pohjimmiltaan on mallien oppimista
- Datasta oppimisen menetelmät jakaantuvat kahteen pääluokkaan: *ohjattu oppiminen* ja *ohjaamaton oppiminen*

Mallit ja oppiminen (7)

- Ohjatussa (kone)oppimisessa annetaan joukko data-alkioita ja niitä vastaavia nimikkeitä joihin ne halutaan liittää: esimerkiksi äänipätkä ihmisen puhetta a-äänteen osalta ja kirjainsymboli “a”
- Tehtävä: muodosta malli joka liittää toisiinsa data-alkiot ja nimikkeet (automaattista puheentunnistusta varten)
- Vastaa ihmisellä opettajan johdolla tapahtuvaa oppimista
- Ohjaamattomassa (kone)oppimisessa annetaan vain joukko data-alkioita mutta ei mitään muuta; esimerkiksi iso määrä kaupan asiakkaistaan keräämiä tietoja
- Tehtävä: muodosta malli, joka yhdistää toisiinsa samoista tuotteista kiinnostuneet asiakkaat (täsmämainontaa varten)
- Vastaa ihmisellä itsekseen tapahtuvaa oppimista.

Esimerkkejä

- Kieliteknologia
  - Konekääntäminen luonnollisten kielten välillä
  - Puheen tunnistaminen (audiotietokannan automaattinen muuntaminen tekstiksi; TV-lähetysten on-line tekstitys)
- Käyttöliittymät
  - Puheentunnistus
  - Käsinkirjoitettujen merkkien tunnistus
  - Eleiden, ilmeiden, katseen suunnan tunnistus
  - Käyttäjän profilointi
- Web-haku
  - Googlen laajennukset; semanttinen verkko
  - Oppiva semantiikka; ontologioiden tms. tiedon rakenteiden automaattinen muodostus
- Teknistieteellinen data

Esimerkkejä (2)

- Tietoliikenne (verkon kuormituksen ennustaminen; ympäristöään tarkkaileva kännykkä)
- Neuroinformatiikka (biolääketieteen mittaukset kuten EEG, MEG, fMRI); ihmisen ja koneen väliset kehittyneet käyttöliittymät
- Bioinformatiikka: geenisekvenssit, DNA-sirudata
- Ympäristön tila, ilmasto
- Sensoriverkot

- Taloudellinen data
  - Aikasarjojen (pörssikurssit, valuuttakurssit) ennustaminen
  - Yritysten tilinpäätöstietojen analyysi
  - Luottokorttien käytön seuranta
  - Asiakkaiden ryhmittely ja käyttäytymisen ennustaminen (ostoskorianalyysi)
- .... ja paljon paljon muuta!

Case study: WEBSOM

- WEBSOM (Web Self Organizing Map) on informaatiotekniikan laboratoriossa prof. Teuvo Kohosen johdolla kehitetty dokumenttien selaus- ja hakujärjestelmä
- Se perustuu SOM (Self Organizing Map) -neuroverkkoon
- Suurimmassa sovelluksessa tehtiin kartta 7 miljoonalle dokumentille, jotka ovat elektronisessa muodossa olevia patenttien tiivistelmiä
- WEBSOMin visuaalisen käyttöliitymän avulla voi helposti selailla tietyn alan patentteja.

▸ Demo: `http://websom.hut.fi/websom/stt/doc/fin/`