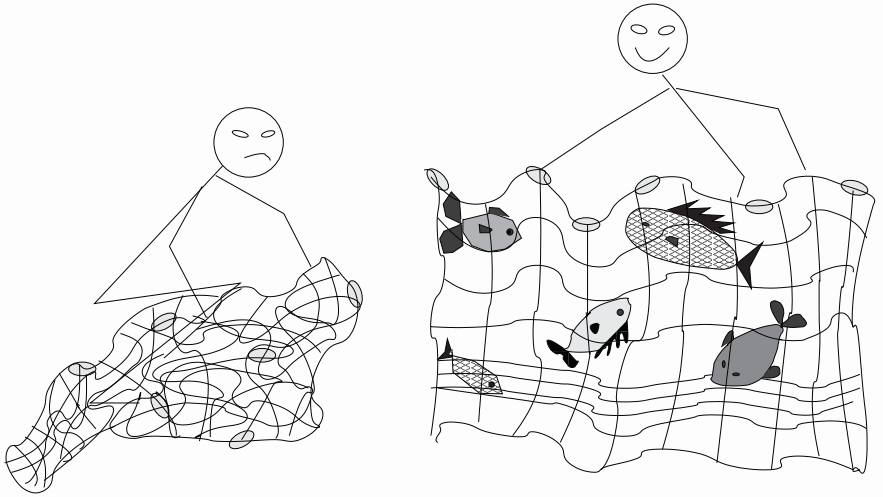
Tiedonlouhinta (kl 2013)



Kurssin kuvaus

- joko kandidaatti- tai maisteritason valinnainen kurssi
- Toteutus pääasiassa ongelmalähtöisesti (tiedonlouhintaprosessin harjoittelu oikealla datalla)
- Yleiskuva menetelmistä ja niiden valinnasta luennoilla (vain 4 vk)
- Perustaitojen harjoittelu laskuharjoituksissa (4 vk)
- Ei koetta, mutta edellyttää aktiivista työskentelyä kurssin aikana! (4op $\approx$ 108h)

Esitiedot

- TKT-opiskelijat: Ohjelmointi-kurssit, TRA1 ja perus-linux-taidot
- Muiden alojen ihmiset: perustaidot tietokoneen käytöstä ja sopiva **datajoukko** harjoitustyöaiheeksi
- Täytä esitietokysely heti! (kurssisivulla)
- Luentoja saavat tulla kuuntelemaan muutkin (mutta harjoitusryhmiin ei ylimääräisiä)

Kurssin arvostelu

- Harjoitustyöt 75% pisteistä (sis. esityksen arvion)
- Harjoituspisteet 25% pisteistä (mikroharjoituksista 3 läsnäolopistettä)
- Esitykset arvioidaan yhdessä
- Ryhmä saa ehdottaa, miten harjoitustyöpisteet jaetaan ryhmäläisten kesken

Opettajat

Wilhelmiina Hämäläinen (luennoija, Kuopion harjoitukset ja ohjaus) whamalai@cs.uef.fi
Kurssin ajan vastaanottoaika ti klo 15-16 (Technopolis F007) tai sopimuksen mukaan

Mikko Malinen (Joensuun harjoitukset ja ohjaus) mmali@cs.uef.fi

Alustava ohjelma

Viikko 15 Luento ma 8.4. Luento ti 9.4. Muu	 Kurssin esittely, johdatus tiedonlouhintaan Datan tyypit ja esiprosessointi. Harjoitustöiden jako ja ohjeistus. Mikroharjoitukset. Tapaamiset ohjaajan kanssa
Viikko 16 Luento ma 15.4. Luento ti 16.4. Muu	 Tilastollisten riippuvuuksien haku Bayesläinen riippuvuuksien mallinnus Harjoitukset. Loput tapaamiset ohjaajan kanssa

Viikko 17 Luento ma 22.4. Luento ti 23.4. Muu	 Luokittelusäännöistä ja luokittelijoista Klusterointi Harjoitukset
Viikko 18 Luento ma 29.4. Luento to 2.5. Muu	 Erikoisempien hahmojen haku (verkkodata, aikasarjadata) Oikean menetelmän valinta. Tulosten järkevyyden ja merkitsevyyden arviointi. Harjoitukset

Viikko 19 Muu	Omatoimista työskentelyä. Tapaamiset ohjaajan kanssa
Viikko 20	Omatoimista työskentelyä.
Viikko 21 Luento ma 22.5. Luento ti 23.5. Muu	 Loppuesitykset Loppuesitykset Loppuraporttien palautus

Mitä on Tiedonlouhinta?

- Data mining (DM) (“data fishing”, “data snooping”, “information harvesting”)
- Knowledge discovery (from databases) (KDD)

Intuitiivisesti

- Laskennallisesti epätriviaalia data-analyysia (uuden informaation etsimistä datasta)

Määritelmiä

“A non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.” (Fayyad et al. 1996)

“Automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies.” (wikipedia)

“Analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.” (Hand et al. luku 1)

Data-analyysi

- raaka data=prosessoimattomia ja tulkitsemattomia faktoja, esim. mittausten tuloksia
- datalla itsessään ei mitään merkitystä
- informaatio=tietoa jolla on merkitys, “tulkittua dataa”
- informaation voi esittää myös rakenteellisena mallina
- suhteellisia termejä: yhden prosessin tuloksena saatu informaatio voi olla toiselle prosessille dataa

Esimerkki: Data vs. informaatio

Data: maanpinnan korkeus eri mittauspisteissä ympäri Suomea eri vuosina

Informaatiota:

- Suomen korkein kohta on Haltilla
- suurin osa maasta on alankoa (alle 200m merenpinnasta)
- Suomen keskikorkeus on n. 154m
- Oulun–Kemin alueella maa kohoaa n. 9 mm vuodessa
- maanpinnan korkeusmalli

Data-analyysia harjoittavia tieteenaloja

Tiedonlouhinnan lisäksi

- tilastotiede (erityisesti deskriptiivinen tilastotiede, eksploratiivinen data-analyysi)
- koneoppiminen (machine learning) (sis. hahmon tunnistuksen, pattern recognition)
- informaation eristäminen tekstistä (information extraction from text) (luonnollisen kielen prosessoinnin, tiedonhaun (information retrieval) ja DM:n välimaastossa)

Tiedonlouhinta vs. tilastotiede?

- Tiedonlouhinnan juuret eksploratiivisessa data-analyysissa (exploratory data analysis): tavoite etsiä uusia piirteitä datasta
- $\rightarrow$ laskennallisesti hankalaa! Tarve algoritmeille
- Tiedonlouhinta mahdollisti monimutkaisempien mallien etsinnän
- Keskeinen ero **laskennallinen elementti** – tiedonlouhinta edellyttää epätriviaalia laskentaa eli kekseliäitä algoritmeja
- Käytännön eroja:
  - tiedonlouhinnassa paljon suuremmat datajoukot
  - tilastotieteessä tilastollisen merkitsevyyden arviointi keskeistä

Tiedonlouhinta vs. koneoppiminen?

Rajanveto vaikeaa!
Joskus molemmat luetaan kattokäsitteen Knowledge discovery alle

Koneoppimisen osa-alueet

- Ohjattu koneoppiminen (supervised ML): ennustava (prediktiivinen) mallinnus (esim. luokittelu)
- Ohjaamaton koneoppiminen (unsupervised ML): kuvaava (deskriptiivinen) mallinnus (esim. klusterointi)

$\leftrightarrow$ Tiedonlouhinnassa lähinnä deskriptiivistä mallinnusta

Filosofinen ero

Koneoppiminen

- Uskomus että data on jonkin mallin tuottamaa.
- Tavoite löytää datan parhaiten selittävä/kuvaava malli.

Tiedonlouhinta:

- Data on ensisijaista. Sillä ei välttämättä ole mitään mallia tai vain lokaaleja hahmoja (osittaismalleja).
- Tavoite löytää kaikki kiinnostavat hahmot.

Malli vs. hahmo (Hand et al luku 6)

Malli (model)

- kattava, abstrakti esitys datasta
- globaali eli koskee kaikkia datapisteitä
- esim. Kaikki opiskelijat voidaan luokitella laiskoihin ja ahkeriin laskuharjoituspisteiden ja luentoaktiivisuuden perusteella.

Hahmo (pattern)

- jotain datan osaa kuvaava piirre (paikallinen malli)
- lokaali eli koskee vain tiettyjä datapisteitä
- esim. Korkeat harjoituspisteet indikoivat hyvää arvosanaa.

DM vs. ML: Käytännön eroja (yleistys)

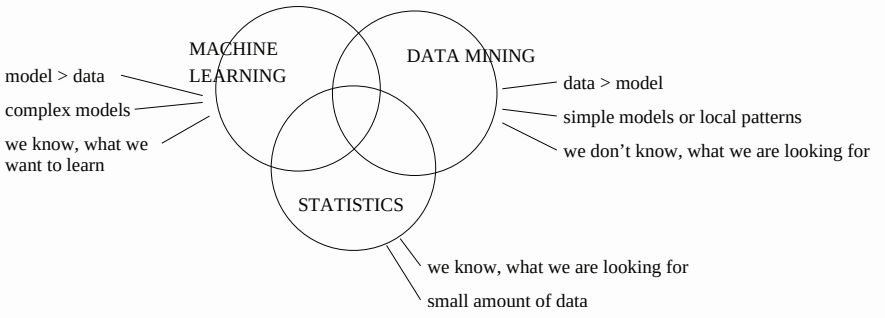
Koneoppimisessa

- monimutkaiset mallit $\rightarrow$ huonommin skaalautuvat algoritmit
- pienemmät datajoukot (satoja–tuhansia rivejä, kymmeniä muuttujia)
- data ja muuttujat valittu huolella mallinnustarkoitukseen

Tiedonlouhinnassa

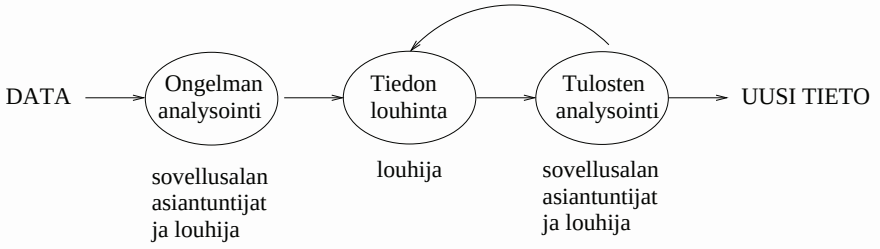
- suhteellisen yksinkertaiset hahmot $\rightarrow$ tehokkaammat algoritmit
- suuret datajoukot (jopa miljoonia rivejä, satoja–tuhansia muuttujia)
- data saatu usein jonkin toisen projektin sivutuotteena

Yhteenveto: Tieteenalojen suhteet

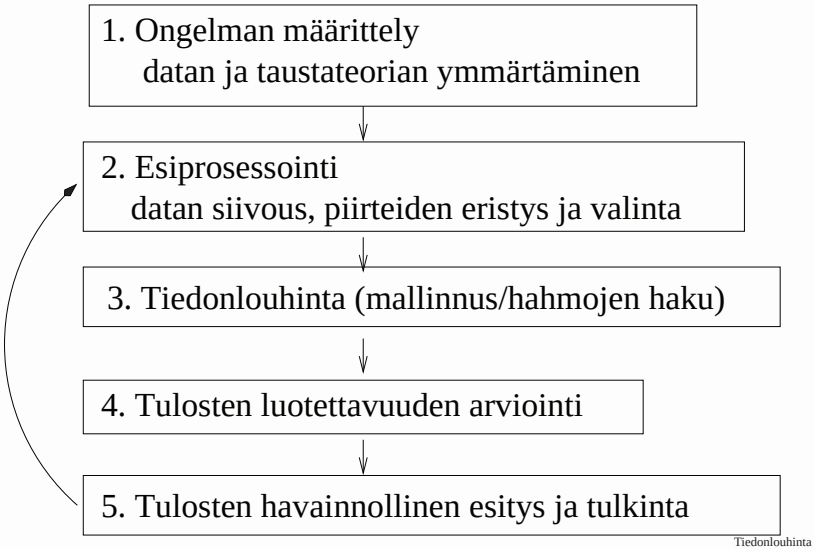


Tiedonlouhintaprosessi

Edellyttää yhteistyötä eri alojen ihmisten välillä!



Prosessin tyypilliset vaiheet



1. Ongelman määrittely

- Datan ymmärtäminen: mitä muuttujat mittaavat/kuvaavat?
- Paljonko dataa on (suhteessa muuttujien määrään)?
- Millaisia hahmoja olisi kiintoisaa/hyödyllistä löytää?
- Mitä tiedetään ennestään?
- Taustateoriaan kannattaa louhijankin perehtyä!
- Vaikeus: Eri alojen ihmisten löydettävä yhteinen kieli (koeta puhua siten, että “maallikkokin” ymmärtää)

2. Esiprosessointi

- Datan yhdistys (eri lähteistä) (voi vaatia normalisointia tms.)
- Alustava analysointi: muuttujien keskiarvot, keskihajonnat, jakaumat, korrelaatiot (tilastollisilla työkaluilla)
- Datan siivous: puuttuvien arvojen käsittely, virheiden tarkistus ja korjaus
- Uusien potentiaalisesti hyvien muuttujien johtaminen (esim. muuttujien yhdistys, diskretointi, normalisointi)

2. Esiprosessointi (jatk.)

- Mallinnuksen kannalta hyvien muuttujien valinta
  - Jos mahdollista, aluksi ei kannata karsia mitään, ennen kuin tiedetään, mitkä muuttujat tarpeettomia
  - Jos laskennallisesti mahdotonta, luodaan vaihtoehtoisia datajoukkoja (samaa asiaa kuvaavat vaihtoehtoiset muuttujat eri joukkoihin)
  - Muuttujien valintamenetelmät eivät koskaan takaa optimaalista tulosta! (*NP*-kova ongelma)

3. Tiedonlouhinta

- Aina kannattaa aloittaa riippuvuusanalyysistä! $\rightarrow$ muuttujien ja mallinnusmenetelmien valinta
- Yleensä deskriptiivinen mallinnus auttaa prediktiivisten mallien muodostuksessa
- esim. Deskr: Etsitään 100 merkitsevintä muuttujaan “Diabetes” liittyvää riippuvuussääntöä
- Pred: Etsitään (valikoidusta datasta) paras diabetesta ennustava malli
- Tyypillisiä mallinnusmenetelmiä riippuvuusanalyysi, klusterointi, luokittelu

Tiedonlouhinta: Riippuvuusanalyysi

Annettu joukko kategorisia muuttujia $R = A_1, \ldots, A_k$, etsi merkitsevimmät tilastolliset riippuvuussäännöt muotoa $X \rightarrow C$ ($X \subsetneq R$, $C \in R$)

- esim. $A_1, A_4 \wedge A_9 \rightarrow A_2$ (A_1 ja A_4 lisäävät A_2:n todennäköisyyttä)
- esim. geeni–elintapa–sairaus -datasta voi löytyä *ApoE-e4, alkoholiakohtuudella* $\rightarrow$ *Alzheimer*

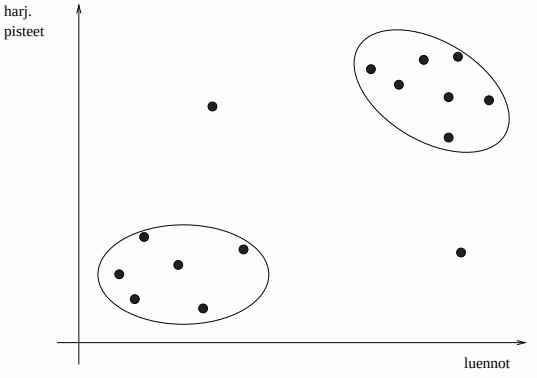
Annettu joukko numeerisia muuttujia $R = A_1, \ldots, A_k$ sekä C, muodosta lineaarinen malli joka ennustaa parhaiten C:tä

- Malli muotoa $C = \beta_1 B_1 + \beta_2 B_2 + \ldots \beta_m B_m + \alpha$, missä $B_i \in R$.

Tiedonlouhinta: Klusterointi

Jaa datapisteet samankaltaisten/lähekkäisten pisteiden muodostamiin ryhmiin eli klustereihin

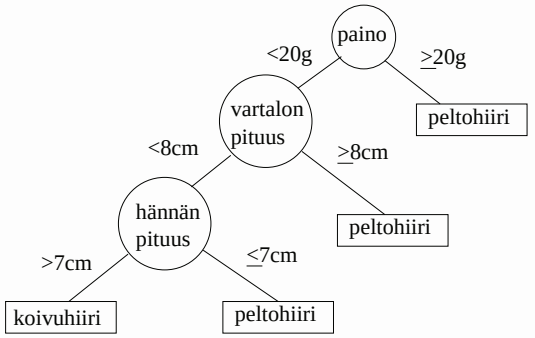
- Esim. opiskelijoiden klusterointi luennoille osallistumisen (kerrat) ja harjoituspisteiden perusteella



Tiedonlouhinta: Luokittelu

Annettuna datapisteiden luokittelu (jako luokkiin), muodosta malli, joka määrittää pisteen luokan muiden muuttujien perusteella.

- esim. päätöspuu pelto- ja koivuhiiren erotukseen painon, vartalon ja hännän pituuden perusteella



4. Tulosten luotettavuuden arviointi

- Ovatko löydetyt hahmot tai mallit järkeviä?? (Muista **aina** tarkistaa!)
  - voi olla, ettei datasta oikeasti löydy mitään mallia – mallinnusmenetelmä palauttaa silloin jonkin satunnaismallin
- Prediktiivisten mallien validointi helppoa:
  - määritetään hyvyysmitta (tai virhemitta, esim. väärin luokiteltujen alkioiden osuus)
  - malli opitaan opetusjoukosta (training set) ja sen virhettä tutkitaan testijoukossa
  - opetus- ja testijoukkojen oltava edustavia!
  - vaihtoehto ristiinvalidointi

4. Tulosten luotettavuuden arviointi (jatk.)

- Deskriptiivisten mallien luotettavuuden arviointi vaikeampaa
- Tavoite: halutaan takuita, että löydetty hahmo ei johdu sattumasta
- Mielellään takuita myös siitä, että samanlainen hahmo pätee jatkossa, samasta–samantapaisesta populaatiosta kerättävässä datassa
- välineenä hahmon tilastollisen merkitsevyyden arviointi

5. Tulosten esitys ja tulkinta

- Tulokset kannattaa esittää mahdollisimman havainnollisesti (esim. graafisesti), jolloin oleellinen erottuu
- Alan asiantuntija keskeisessä roolissa!
- Löytyikö jotain uutta? Voiko tulosten pohjalta muodostaa uutta hypoteesia? Miten sitä voisi tutkia lisää?
- Johtaa useimmiten uuteen louhintakierrokseen: kokeillaan uusia muuttujia ja mahdollisesti uusia menetelmiä
- Kierrokset lopetetaan vasta, kun ollaan tyytyväisiä tai mitään uutta ei näytä enää löytyvän

- “Täydellinen” data-analyysi silti jokseenkin mahdotonta – Miksi?

Tarvitaanko tiedonlouhintaan tietojenkäsittelytieteilijää?

- Soveltava tiedonlouhinta: käytetään yleensä valmiita työkaluja
- Periaatteessa kuka tahansa oppii käyttämään (pitää vain tuntea perusteet – mitä kannattaa tehdä ja miten)
- Valmiita helppokäyttöisiä työkaluja on, mutta ei uusimmista algoritmeista (ohjelmoija osaa toteuttaa algoritmin)
- Joskus ongelma vaatii ihan uutta menetelmää (uudenlaisen hahmon hakua) $\rightarrow$ tarvitaan tiedonlouhinta-algoritmien suunnittelijaa ja ohjelmoijaa

Tarvitaanko tietojenkäsittelytieteilijää?

- Esiprosessointivaiheessa tarvitaan yleensä ohjelmointitaitoja (uusien hyvien muuttujien muodostus on taidelaji!)
- Mutta: jos data on tosi siistiä, muuttujat hyviä ja työkalu kattava, voi louhinnasta selvitä osaamatta koodata riviäkään.
- Työelämässä tiedonlouhijoilla ja data-analyytikoilla on paljon kysyntää (ja kysyntä kasvaa)!