Yleistä asiaa 1. Työkalujen demosessio ma 6.5. klo 16.15→

Nyyttikesti-periaatteella! Ohjelmassa ainakin:

- SPSS (Salla ja Sari)
- R (Jonne)
- Cytoscape (verkkojen visualisointi, Antti)
- SOM:it VisualDatalla (Nina)
- Fourier-muunnos ja sen tulkinta matlabilla (Markus)
- joku yleinen visualisointityökalu, ehkä gnuplot

tarpeen mukaan 5–10(–15) min demous, ekstrapisteitä oman valinnan mukaan (esim. 1–3p)

Yleistä 2. Harjoitustyöt

- tästä eteenpäin Wilhelmiina ohjaa kaikkia töitä (Käyttäkää hyväksi! tapaamisia, sposteja, välitulosten tarkastelua, jne.)
  päivystys ti 15-16 (huone F007 tai työpuh. 050-4422472)
- sovitaan lopputapaamiset viikoille 19–20 (kun jo mallinnustuloksia)
- varaudu intensiiviseen työskentelyyn!
- loppuesitykset ja -raportit viikolla 21 (raporttien deadline ke 29.5.)
- arviointi pääosin itsearviointina

Luento 8: Mallien arviointi ja vertailu sekä menetelmien valinta

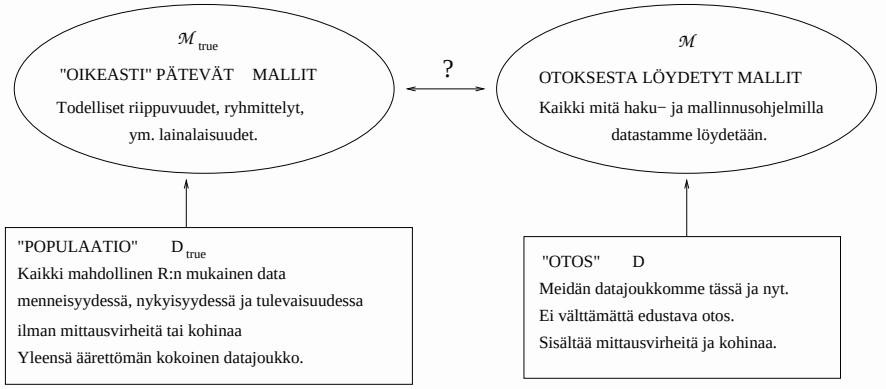
1. Kun löydetty jokin hahmo tai malli
   - Onko se “todellinen” vai vain sattuman tuotetta?
   - Miten hyvä se on?
   - Miten testata sen pätemistä?
2. Annettuna data ja ongelma
   - Mitä luokittelu- tai klusterointimenetelmää kannattaa kokeilla?
   - Miten parametrit kannattaa valita?
   - “Voisiko Sami tehdä ohjelman, joka valitsisi parhaan luokittelijan parametreineen automaattisesti?”

Luento 8: Sisältö

1. Mallien arvioinnin perusongelma
2. Riippuvuuksien systemaattinen haku
   - merkitsevyys ja moninkertaisen testauksen ongelma
3. Luokittelijat
   - Miten arvioida luokittelijan hyvyyttä?
   - validointi ja testaus
   - luokittelumenetelmän valinta
4. Klusterointi
   - Onko löydetty klusterointi “aito”?
   - klusterointimenetelmän valinta

1. Mallien arvioinnin perusongelma

Meillä on vain D, joten voimme muodostaa vain $\mathcal{M}$:n. Sisältyykö se tai edes osa siitä $\mathcal{M}_{true}$:een?



Kommervenkkejä

Huom! Satunnaisotos=otos jonka alkiot valittu satunnaisesti populaatiosta → todennäköisesti edustava otos

1. **D ei ole satunnaisotos (edustava)**, vaan jotenkin valikoitunut
   - Yleistä DM:ssä! Data usein sivutuotetta, ei kyseiseen mallinnukseen kerätty.
   - löydetään vain kyseistä osapopulaatiota kuvaavia malleja (tai havaitaan, ettei sille löydy malleja)
   - esim. Suurin osa turkistarhan rotista pyydetty talviaikaan. Onko talven takia enemmän vatsahaavaa vai onko turkistarha stressaavampi asuinpaikka?

Kommervenkkejä (jatk.)

2. **D on satunnaisotos, mutta sisältää hirveästi kohinaa** (epätarkat–virheelliset muuttuja-arvot)
   - voi olla oikeita malleja (eli $\mathcal{M}_{true} \neq \emptyset$), mutta ei pystytä löytämään
   - tai löydetään kohinan tuottamia malleja vaikka $\mathcal{M}_{true} = \emptyset$
   - esim. yhden päivän kiihtyvyysdatasta opitaan luokittelija, joka tunnistaa kaikki lehmien toiminnot lähes 100% tarkkuudella, mutta ei kykene tunnistamaan mitään toisen päivän datasta.
   - kohina + ylisovittunut malli paha yhdistelmä!

Kommervenkkejä (jatk.)

3. **D on hyvin pieni**
   - ei voi olla edustava (ellei hyvin triviaali data)
   - ei takeita että mallit (tai niiden puute) pätisivät muille D_{true}:n otoksille
   - esim. Poikivien lehmien toimintodataa vain muutamalta lehmältä ja kultakin vain 1–2 vajaata vuorokautta. → Voi tehdä hypoteeseja, mutta ei takeita, että havainnot pätisivät yleisemmin. (=kvalitatiivinen tutkimus, tavoitteena ymmärrys)
   - Huom! Voidaan kyllä **falsifioida** aiemman tutkimuksen väitteitä (jos löytyy vastaesimerkkejä). (Esim. lehmien pesänrakennus)

Ratkaisuja

Deskriptiiviset mallit/hahmot

- Tehdään johtopäätöksiä vain siitä populaatiosta, jolle datamme olisi edustava otos
- Tutkitaan löytöjen tilastollista merkitsevyyttä: millä todennäköisyydellä johtuisivat sattumasta?

Prediktiiviset mallit

- Testataan opetusjoukon ulkopuolisella (siitä riippumattomalla) datalla
- Arvioidaan yleistysvirhettä (=odotettua virhettä ulkopuolisella datalla)

2. Riippuvuksien systemaattinen haku

Ongelma: Etsitään datasta K parasta riippuvuussääntöä $X \rightarrow A$ tai lineaariregressiota $A = \beta_1 A_1 + \ldots \beta_k A_k + \alpha$. Millä todennäköisyydellä hahmot johtuvat sattumasta?

- Arvioidaan yksittäisten hahmojen tilastolliset merkitsevyydet (p-arvot)
- Fisheriläiset vain raportoivat hahmot ja niiden p-arvot
- Neyman-Pearsonilaiset vertaavat jokaista p:tä johonkin valitsemaansa raja-arvoon p_{max} ja kertovat mitkä hahmot olivat merkitseviä merkitsevyystasolla p_{max} (tyypillisesti $p_{max} = 0.05$ tai $p_{max} = 0.01$)
  - Mielivaltaisia raja-arvoja, mutta perinne jatkuu vahvana!

Moninkertaisen testauksen ongelma (multiple testing problem)

- Olk. jokin testisuure T s.e. hahmo johtuu sattumasta todennäköisyydellä $p \leq 0.05$ jos $T \geq T_{0.05}$.
- Hahmo "läpäisee testin" tasolla 0.05, jos sen T-arvo oli $T \geq T_{0.05}$. Muuten se hylätään sattuman tuotteena.
- Jos testataan 1000 hahmoa, jotka kaikki läpäisevät testin, on odotettavissa, että 50 näistä onkin sattuman tuotetta.
- ei ole olemassa hyvää ratkaisua!
- klassinen ratkaisu kiristää tasovaatimuksia, kun testien lukumäärä kasvaa
  esim. Bonferroni-korjaus: Jos testataan m hahmoa, vaaditaankin että $p_{max} = \frac{0.05}{m}$ (voi olla kova vaatimus ja hukata oikeasti merkitseviä hahmoja)

Ongelma: Mittasuureiden esioletukset

Mittasuureet T (esim. χ^2, F, z) vastaavat p-arvoja vain tiettyjen esioletusten pätiessä. Usein epärealistisia!

Esim. χ^2-testi riippuvuussäännöille

- vaatii, että kaikki kombinaatiot $fr(XA)$, $fr(X\neg A), fr(\neg XA)$ ja $fr(\neg X \neg A)$ riittävän suuria
- lisäksi allaoleva hypergeometrinen jakauma ei saa olla liian vino (skewed, käytännössä $fr(A) << 0.5$ tai $fr(A) >> 0.5$)
- Ratkaisu: laske suoraan p_F hypergeometrisesta jakaumasta ("Fisherin eksakti testi") tai käytä Mutual Information-mittaa (robusti, antaa suurinpiirtein samat tulokset)

Esim.2 lineaariregression hyvyys

Monia mittalukuja ja testejä!

Tarkastellaan esimerkkiä: M=rotan paino, V=vartalonpituus ja H= hännän pituus. Halutaan tutkia miten V ja/tai H selittävät M:ää. $n = 50$.

Malli	r^2	p_F	MSE
$\overline{M} = 32.37H - 268.40$	0.67	1.9e-13	1324
$\overline{M} = 24.24V - 271.52$	0.75	1.4e-16	989
$\overline{M} = 14.56H + 16.30V - 332.89$	0.81	4.9e-18	782

Onko "Häntä, Vartalo → Paino" paras malli?

Em. mittalukujen perusteella olisi. Mutta onko ylisovittunut? → lasketaan Mallow:in C-mitta:

$$C = \frac{SSE}{MSE_k} - n + 2p,$$

missä p on mallin muuttujien lkm, SSE on mallin virheiden neliösumma ja MSE_k on keskimääräinen virheiden neliösumma kun kaikki k muuttujaa mukana.

Pieni C hyvä. Suositus: Jos mahdollista, valitse malli, jossa $C \approx p$.

Onko "Häntä, Vartalo → ylisovittunut?

Nyt C(vartalo→ paino)=14.8>2.0= C(häntä, vartalo→ paino).

⇒ valitaan "Häntä, Vartalo → Paino"

(Kannattaisi vielä **ristiinvalidoida**!)

Ks. esim. Milton ja Arnold: Introduction to probability and statistics: Principles and applications for engineering and computing sciences.

3. Luokittelijat

Ongelma: Datasta opittu hieno luokittelija, joka luokittelee yli 81% rotista oikein vatsahaavallisiin tai -haavattomiin, pelkästään vuodenajan, paikan ja ruumiin mittojen perusteella. Lisäksi oikeellisuus on määritetty **ristiinvalidoinnilla**, joten sen pitäisi kuvata mallin hyvyyttä opetusjoukon ulkopuolella.

Eikö Liisan pitäisi julkaista heti merkittävä tutkimustulos?? "Rottien vatsahaavan salat selvitetty"

Katsotaan sekaannusmatriisia

		Luokittelijan mukaan		
		Vatsa-haava	Ei vatsa-haavaa	Σ
Oikeasti	Vatsa-haava	TP=0	FP=50	TP+FP=50
	Ei vatsa-haavaa	FN=0	TN=217	FN+TN=217
	Σ	TP+FN=0	FP+TN=267	n=267

Tähän ei tarvitsisi mitään luokittelijaa! (Oikeasti oli 81.3% vatsahaavattomia)

$\kappa = 0$ olisi paljastanut että luokittelijan hyvyys sattumaa.

Luokittelun hyvyyden mittoja: *accuracy* ja κ

Olk. $M(x_i)$ luokittelijan määräämä luokka pisteelle x_i ja c_i sen oikea luokka. (Kaavoissa ol. binääriluokittelu)

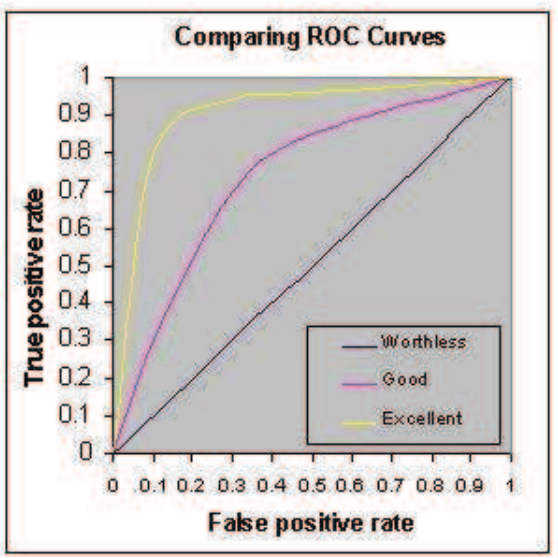
$$accuracy = \frac{\sum_{i=1}^{n} I(M(x_i) = c_i)}{n} = \frac{TP + TN}{n}$$

(I=indentiteettifunktio eli $I(M(x_i) = c_i) = 1$ kun $M(x_i) = c_i$ ja $I(M(x_i) = c_i) = 0$ muuten). $err = 1 - accuracy$

$$\kappa = \frac{n(TP + TN) - fr(C)(TP + FP) - fr(\neg C)(TN + FN)}{n^2 - fr(C)(TP + FP) - fr(\neg C)(TN + FN)}$$

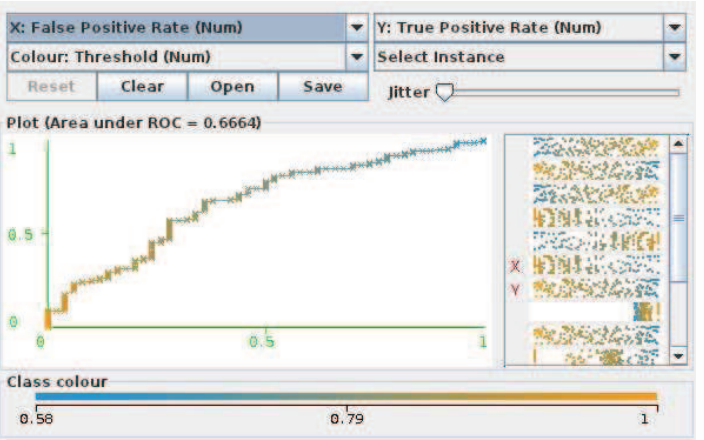
(tutkii monastiko luokittelija ja todellisuus samaa mieltä, ilman että se olisi sattumaa) Suositus: oltava $\kappa > 0.4$.

Luokittelun hyvyyden mittoja: ROC-area (AUC)



AUC Wekassa

Klikkaa tulosta ja valitse Visualize threshold curve. Katso että x-akselilla false positive rate ja y-akselilla true positive rate.



Validointi ja testaus

Termejä käytetään sekaisin!

Perusmenetelmät:

1. Valitaan satunnaisesti osa (esim. 1/3) datasta testaukseen ja loput mallin oppimiseen.
   - Ongelma: Oppimiseen jää vähemmän dataa ja voi tulla huonompi malli.
   - Testidatassa ei saisi esiintyä samoja alkioita kuin opetusdatassa! (duplikaatit ja melkein duplikaatit ongelma) **Miksi?**

Validointi ja testaus (jatk.)

2. m-kertainen ristiinvalidointi (cross validation):
   - Jaa data satunnaisesti m:ään suurinpiirtein yhtä suureen palaan. Olkoon palat $d_1, \ldots, d_m$.
   - tyypillisesti $m = 10$, mutta pienessä datassa voi olla $m = n$ (Leave-one-out cross validation, LOOCV)
   - Toista mallin muodostus ja testaus m kertaa: Kerralla i valitaan testijoukoksi d_i ja lopuista opitaan malli.
   - Laske virheiden (tms. mittalukujen) keskiarvot.

Validointi ja testaus (jatk.)

- Joskus luokittelijan oppimisalgoritmi käyttää itse erillistä validointijoukkoa esim. muuttujien valintaan
- Tätä joukkoa ei saa käyttää testaukseen! (on jo ollut mukana mallin muodostuksessa)
- Testijoukon pitäisi arvioida miten hyvin luokittelija toimii muodostusdatan ulkopuolisella datalla (off-sample data)
- Suositus: Luokittelijan valintaan (monesta vaihtoehdosta) ja testaukseen eri datat!

Vinkkejä

- Jos yleistysvirhe (testidatalla tai ristiinvalidoinnilla) hirmuhyvä, kannattaa epäillä! (bugi tai valikoitunut data)
- Jos lasket datan epäkonsistenttien rivien osuuden, saat alarajan parhaalle mahdolliselle virheelle (olettaen että sama epäkonsistenssi oikeassa populaatiossa)
- Tutki aina mitä olisi luokitteluvirhe puhtaalla arvauksella! (pitää saada sitä parempi)
- luokkien epätasainen jakauma voi aiheuttaa ongelmia $\rightarrow$ tasoita

Luokittelumenetelmän valinta

“Voisiko Sami tehdä ohjelman, joka valitsisi parhaan luokittelijan parametreineen automaattisesti?”

- Universaalisti parasta luokittelumenetelmää ei ole! (Ei edes satunnaisarvausta parempaa – sanoo **“No free lunch” -teoreema**!)
- Sopivimman menetelmän valinta heuristista ja riippuu esitietämyksestä (data, ongelma-alue, tavoitteet)
- jopa parhaat luokittelijoiden oppimisalgoritmit tuottavat huonoja tuloksia joissain ongelmissa!

$\Rightarrow$ **Tiedonlouhijan hallittava valikoima erilaisia menetelmiä**

[Extra] No free lunch -teoreema

Olk. F opittava kohdefunktio joka liittää johonkin muuttujajoukkoon X luokan C.

Olk. $E(err|F, d, A_i)$= virheen odotusarvo algoritmilla A_i ja opetusdatalla d, kun odotusarvo yli kaikkien mahd. funktioiden F.

NFL: Jos kaikki F:t yhtä todennäköisiä, niin kaikille algoritmeilla A_1 ja A_2 pätee $E(err|F, d, A_1) = E(err|F, d, A_2)$

- toinen algoritmeista voisi olla vaikka satunnaisarvaaja!

[Extra] No free lunch -teoreema

- Jos F:ien todennäköisyydet vaihtelevat, lasketaan odotusarvo yli kaikkien $P(F)$:ien, jolloin pätee $E(err|d, A_1) = E(err|d, A_2)$
- Samat tulokset pätevät, kun ehdollistetaan datan koolla n itse datan d sijasta.
- Yllätys: Vaikka tuntisimme d:n (mutta emme F:ää), ei mikään algoritmi ole parempi toista.

“Paha ei ole kenkään algoritminen, vaan toinen on heikompi toista.”

Opas luokittelumenetelmän valintaan

Ominaisuus	Päätöspuu	Naiivi-Bayes	Yl. Bayes-luokittelija	Neuro-verkko	SVM	K lähintä naapuria
Datan tyyppi numeerinen						
kategorinen						
sekadata						
Datan laatu sis. ulkopuolisia						
kohinaa						
puuttuvia arvoja						

Opas luokittelumenetelmän valintaan

Ominaisuus	Päätöspuu	Naiivi-Bayes	Yl. Bayes-luokittelija	Neuro-verkko	SVM	K lähintä naapuria
Datan määrä vähän						
Tehokkuus luokittelijan oppiminen						
luokittelu						
luokittelijan päivitys						

Opas luokittelumenetelmän valintaan

Ominaisuus	Päätöspuu	Naiivi-Bayes	Yl. Bayes-luokittelija	Neuro-verkko	SVM	K lähintä naapuria
Ilmaisuvoimainen (epälin. lk-rajat)						
Havainnollinen						

4. Klusteroinnin arviointi

Ongelma: Onko datassa klustereita? Montako? Johtuvatko sattumasta? Miten klusteroinnin tilastollista merkitsevyyttä pitäisi arvioida?

- Ei kunnollista teoriaa!
- Eräs tapa iteratiivinen testaus (eri K:n arvoille)
  - Olk. $M(K)$ klusteroinnin hyvyysmitan paras arvo, kun K klusteria
  - Ol. että koko data yhtä normaalijakautunutta klusteria
  - Lasketaan $M(2)$:n **jakauma** tällä oletuksella ja verrataan saavutettua M-arvoa siihen → saadaan todennäköisyys, että näin hyvä $M(2)$-arvo johtuisi sattumasta

4. Klusteroinnin arviointi (jatk.)

- Mikäli 2 klusteria oli uskottavaa, tehdään siitä uusi nollahypoteesi (oletus)
- Lasketaan $M(3)$:n jakauma tällä oletuksella jne.
- Ongelma: $M(K)$:n jakaumaa ei voi määrittää tarkasti, vain karkeita arvioita! (Ks. Duda et al. s. 558)
- Malliperustainen klusterointi määrittää datan uskottavuuden annettuna klusterointi!
- lisäksi monia heuristisia klusteroinnin hyvyysmittoja → voi käyttää optimaalisen K:n määritykseen (mahd. käännekohdasta K)

Vinkkejä

- Jos muuttujat eri skaaloissa (tai eri suureita), skaalaa samalle välille
- Kokeile monia eri menetelmiä ja kutakin useita kertoja (välillä data sotketaan, jos järjestyksestä riippuvainen menetelmä)
- Jos vaihtelu vähäistä, voi ainakin klusterien ydinalueita pitää uskottavina
- Klusterointia voi käyttää piirteiden valintaan luokittelua varten (valitaan piirteet jotka tuottavat homogeenisia klustereita) (toimii erityisesti jos $K - NN$-luokittelija)

Opas klusterointimenetelmän valintaan

Ominaisuus	K-means	K-medians	Single-link	Complete-	Average-link	Centroid-link	Minimum variance	Model-based (+Norm. jak.)
Klusterit muoto vapaa								
voi olla erikokoisia								
päällekäisiä								

Opas klusterointimenetelmän valintaan

Ominaisuus	K-means	K-medians	Single-link	Complete-link	Average-link	Centroid-link	Minimum variance	Model-based (+Norm. jak.)
Datan laatu sietää kohinaa								
ulkopuolisia								
epärelevantteja muuttujia								

Opas klusterointimenetelmän valintaan

Ominaisuus	K-means	K-medians	Single-link	Complete-link	Average-link	Centroid-link	Minimum variance	Model-based (+Norm. jak.)
Ei riipu datan järj.								
Tehokas (kun n iso)								