Tiedonlouhinta 2013
Kotikoe (palautuspäivä 24.6.2012)

Tee tehtäviä kaikista osioista. (Kannattaa ehdottomasti tehdä kaikki tehtävät! Silloin harjoitustyökin sujuu helpommin.) Tehtävät arvostellaan ja mikäli saat vähintään 15 pistettä (60% maksimipisteistä), saat harjoitustyöaiheen. (Lisäksi tehtäväpisteet vaikuttavat myös arvosanaan.) Huom! Osasta tehtäviä voi saada 2 pistettä, mikäli ne on tehnyt kattavasti ja oikein.

Tehtävissä tarvittavat datajoukot ja työkalut löytyvät sivulta `http://cs.joensuu.fi/pages/whamalai/DM13/kotikoe.html`.

I Data ja sen esiprosessointi

1. (**1 piste**) Joukkoa tutkittuja rottia on luonnehdittu seuraavilla muuttujilla. Mitä ovat muuttujien tyypit? (numeerinen vai kategorinen, diskreetti vai jatkuva, ordinaalinen vai nominaalinen, onko binäärinen)

   - vuodenaika jolloin pyydetty: 1=talvi, 2=kevät, 3=kesä, 4=syksy
   - pyyntipaikka: Konneveden kaatopaikka, Keskimmäisen kaatopaikka, turkistarha, laboratorio
   - sukupuoli (1=naaras, 2=uros)
   - oliko laboratoriorotta (1=tosi, 0=epätosi)
   - vartalon pituus senttimetreinä (yhden desimaalin tarkkuudella)
   - maksan paino grammoina (0 desimaalin tarkkuudella)
   - lisämunuaisten paino milligrammoina (0 desimaalin tarkkuudella)
   - vatsahaavan aste (arvosana 1,2,3,4,5, missä 1 tarkoittaa ettei vatsahaavaa ja 5 pahalaatuista vatsahaavaa)

2. (**1 piste**) Tutki datajoukosta naaraatvalikoitu.csv seuraavien muuttujien histogrammeja: vaika, paino, thymus-ratio, hantaratio. Valitse luontevat diskretointivälit visuaalisen tarkastelun perusteella: vaika 3 osaväliä (mahdollisimman tasakokoisia, mutta siten että talvi säilyisi yhtenäisenä), paino 3 osaväliä (pienet, keskikokoiset ja suuret), thymusratio 2 osaväliä (normaali tai suuri, jälkimmäiset todennäköisesti poikasia, joten tällä voi erottaa ne) ja häntäratio 3 osaväliä (lyhyt-, normaali- ja pitkähäntäiset). Valitse diskretointirajat siten, että jokaiselle osavälille tulee vähintään 30 datariviä. (Lisäkysymys: miksi tämä on tärkeää?)

3. (**1 piste**) Tutustu seuraaviin automaattisiin muuttujien diskretointimenetelemiin. Anna jokaisesta esimerkkejä, milloin menetelmä on hyvä ja milloin huono.

   - Tasavälidiskretointi (equal width)
   - Tasafrekvenssidiskretointi (equal frequency)

- Informaatioteoreettiset menetelmät (kuten maximum entropy)
- Klusteroimalla diskretointi.

4. (**1 piste**) Tarkastellaan lehmien ruokinta-automaattien lokidataa. Jokaisesta syönnistä on lokitietue, joka sisältää seuraavat tiedot: lehmän tunniste, automaatin tunniste, aloitus- ja lopetusaika, rehun tyyppi sekä syöty määrä. Millaisia uusia muuttujia kannattaisi muodostaa seuraaviin ongelmiin?

   - Haluttaisiin tunnistaa napostelijat ja suursyömärit eli ne jotka syövät vähän mutta usein ja ne jotka syövät harvoin mutta paljon kerralla.
   - Eri rehutyypit eivät ole vertailukelposia, koska niiden paino ja ravitsevuus vaihtelevat. Haluttaisiin silti tunnistaa syönnit, joissa määrä on ollut poikkeuksellisen suuri tai pieni.
   - Haluttaisiin tunnistaa yksilöt jotka syövät muita merkittävästi nopeammin tai hitaammin sekä kerrat, jolloin eläin on syönyt omaksi itsekseen poikkeuksellisen nopeasti tai hitaasti.
   - Taustatiedoissa on kerrottu eläimen paino ja poikimakertojen lukumäärä, jotka heijastavat eläimen asemaa lauman hierarkiassa. Haluttaisiin tutkia, onko näillä vaikutusta siihen, mitkä yksilöt syövät ruuhka-aikoina ja mitkä niiden ulkopuolella.

II Riippuvuusanalyysi

1. (**2 pistettä**) Etsi worldstatcleaned.csv-datasta kaikki (Pearsonin) korrelaatiot ja mutual informationit eli MI:t (Mutuali-laskimella). Tutki kaikki riippuvuudet joiden korrelaation/MI:n itseisarvo on vähintään 0.40. Mitkä löytyvät molemmilla mitoilla ja mitkä vain toisella? Plottaa kuvat kiinnostavimpien muuttujaparien suhteen (kuvia saat myös Mutuali-ohjelmasta). (Vinkki: Mutualissa on annettava diskretointivälien lukumäärä. Binääri- ja ordinaalimuuttujille annetaan arvojen lukumäärä ja numeerisille muuttujille riittävän iso lukumäärä, esim. 100.)

2. (**1 piste**) Tiedostossa marjasaantoja.txt on soiden kasvillisuusdatasta löytyviä riippuvuussääntöjä. Tutustu sääntöjen tulkintaan (notaatiot `http://www.cs.joensuu.fi/~whamalai/kfstatistics.pdf`) ja vastaa seuraaviin kysymyksiin.

   - Mitä kasveja kasvavilta soilta kannattaa etsiä puolukkaa? Entä mustikkaa?
   - Mitä vinkkejä saat variksenmarjan etsintään? Miten suurella todennäköisyydellä löydät variksenmarjaa, jos noudatat löytämiäsi vinkkejä?
   - Mitä selviää karpaloiden kasvupaikoista? Voiko löydettyihin riippuvuuksiin luottaa?
   - Säännöt 5 ja 8 ovat esimerkki riippuvuudesta ja sen yleistyksestä. Mitä näiden sääntöjen vertailu kertoo?

3. (**2 pistettä**) Etsi naaraatvalikoitudiskr.csv-datasta riippuvuussääntöjä Kingfisher-ohjelmalla. Muunna ensin data transaktiomuotoon (poista header, korvaa pilkut välilyönneillä ja koodaa attribuutit numeerisiksi namescodes-ohjelmalla). Ohjeet löytyvät myös mikroharjoitusten tehtävistä. Etsi ainakin 100 parasta sääntöä $\ln(p)$-mitalla (oletusarvo). Käännä sääntöjen koodit nimiksi namescodes-ohjelmalla (liitteeksi) ja anlysoi huolella. Mitä säännöt kertovat vatsahaavan esiintymisestä? Entä kaatopaikka- ja turkistarharottien eroista? Entä pitkähäntäisyyden vaikutuksesta hyvinvointiin?

4. (**2 pistettä**) Seuraavassa taulukossa on vertailtu worldstatcleaned.csv-datasta muodostettuja lineaariregressiomalleja eliniän (E) ennustamiseen, kun selittävät muuttujat ovat S=syntyvyys, L=lapsikuolleisuus, M=maaseutuväestön osuus ja T=terveysmenot (asukasta kohden bruttokansantuotteesta). Mittaluvut ovat r^2=moninkertaisen korrelaatiokertoimen neliö, p_F=regressiomallin tilastollinen merkitsevyys, SS=virheiden neliösumma.

malli	r^2	p_F	SSE
$S, L, M, T \to E$	0.88	6.2e-77	1899.1
$S, T, L \to E$	0.88	2.1e-77	1944.7
$L, M, T \to E$	0.87	2.4e-75	2057.1
$S, M, T \to E$	0.77	1.2e-53	3720.0
$S, L, M \to E$	0.87	5.0e-76	2019.6
$S, T \to E$	0.76	1.5e-53	3861.1
$S, L \to E$	0.87	2.4e-75	2139.4
$S, M \to E$	0.76	2.3e-53	3879.7
$L, T \to E$	0.87	2.8e-75	2142.4
$M, T \to E$	0.44	6.4e-22	9093.3
$L, M \to E$	0.86	5.3e-74	2218.2
$S \to E$	0.74	3.9e-52	4171.5
$T \to E$	0.28	4.8e-14	11572.4
$L \to E$	0.85	4.1e-72	2437.2
$M \to E$	0.37	4.9e-19	10131.6

Lisäksi tunnetaan selittävien muuttujien väliset korrelaatiot $r(S,T) = -0.48$, $r(S,L) = 0.86$, $r(S,M) = 0.58$, $r(T,L) = -0.45$, $r(T,M) = -0.52$, $r(L,M) = 0.56$.

   (a) Selvitä mitä mittaluvut merkitsevät! Ovatko kaikki mallit tilastollisesti merkitseviä? Mikä on ennustuksen kannalta tärkein muuttuja? Entä vähiten tärkeä? Kärsivätkö mallit multikollineaarisuudesta? Mitkä olisivat siinä mielessä parhaat mallit? Minkä mallin valitsisit näiden hyvyysmittojen perusteella? (Muista varoa ylisovittumista!)

   (b) Laske joka tasolta (3 muuttujan, 2 muuttujan ja 1 muuttujan regressiomallit) SSE:n perusteella parhaalle mallille Mallowin C ja valitse sen perusteella paras malli. (Datan koko on $n = 173$ ja MSE_k saadaan ensimmäisen eli täydellisen mallin SSE:stä.)

5. (**1 piste**) Valitse datajoukosta naaraatvalikoitu.csv muuttujat paino, lisamunratio, maksaratio, pernaratio, sydanratio, thymusratio, umpisratio, gonratio,

uusrasvaratio, batratio, hantaratio, painoind ja etsi mahdollisimman hyvä lineaariregressiomalli hantaration ennustamiseen. Voit kokeilla eri muuttujayhdistelmiä taulukkolaskentaohjelmassa tai antaa esim. wekan valita parhaan muuttujajoukon (wekassa classify → linearregression). Miten hyvän mallin löysit? (Raportoi ainakin korrelaatiokerroin ja neliövirhe) Kärsiikö mallisi multikollineaarisuudesta?

III Luokittelu

1. (**2 pistettä**) Vertaa eri luokittelumenetelmiä entisten siirtomaiden (muuttuja exColony) tunnistukseen datasta worldstatcleaned.csv. Kokeile ainakin seuraavia menetelmiä: Päätöspuu (J48 wekassa), NaiviBayes, K-lähintä naapuria (kokeile eri naapurien lukumääriä tai anna wekan määrittää optimimäärä) ja yksinkertainen neuroverkko (1 piilotaso). Raportoi kaikista sekaannusmatriisit ja luokitteluvirhe 10-kertaisella ristiinvalidoinnilla (wekassa oletusarvo). Millä oli pienin virhe? Entä mikä oli paras malli jos haluttiin tunnistaa oikein mahdollisimman moni ex-siirtomaa?

2. (**1 piste**) Määritä edellisen tehtävän datasta luokitteluvirheen odotusarvo, jos käytetään satunnaisarvausta. Entä mikä on odotusarvo parhaan mahdollisen luokittelijan virheelle? Tämän saat arvioitua laskemalla epäkonsistenttien rivien lukumäärän (osuuden) datasta. Olivatko edellisessä tehtävässä löytämäsi luokittelijat tämän perusteella hyviä?

IV Klusterointi

1. (**2 pistettä**) Klusteroi data naaraatvalikoitu.csv eri menetelmällä 5 klusteriin seuraavien muuttujien suhteen: lisamunratio ja gonratio (molemmat hyvinvointi-indikaattoreita). Kokeile ainakin seuraavia: K-means, hierarkinen klusterointi eri linkitysmetriikoilla sekä EM-menetelmällä. Liitä vastaukseen klusterointien kuvat (visualisoinnit). Millä tuli uskottavimman näköisiä klustereita? (Tämä tehtävä kannattaa tehdä vasta, kun olet lukenut seuraavan tehtävänannon – tehtävät on nopeinta tehdä samassa weka-istunnossa.)

2. (**1 piste**) Klusteroi edellisen tehtävän data 5 klusteriin samojen muuttujien suhteen, mutta säästä muuttujat vaika, vhaava, rasitus ja paikka. Käytä edellisessä tehtävässä parhaana pitämääsi menetelmää (tai vaikka hierarkista klusterointia average-link-metriikalla). Talleta klusteroinnit ja analysoi niiden sisältöä muuttujien vaika, vhaava, rasitus ja paikka suhteen. Löytyykö näiden suhteen kiinnostavia klustereita? (esim. talvella pyydettyjen, vatsahaavallisten tai imettävien rottien klustereita)? (Huom! Klusteritietojen kanssa talletettu data kannattaa ensin järjestää klusterin perusteella, jotta saman klusterin jäsenet saa peräkkäin.)

V Temporaalinen data ja sen visualisointi

1. (**1 piste**) Tiedostossa ulkolampokesa.txt on ulkolämpötilamittauksia tunnin välein kesäkuun lopun – heinäkuun alun ajalta. Plottaa alkuperäinen data. Laske sen jälkeen jokaisen vuorokauden tunnin keskiarvot ja plottaa niiden kuvaaja. Mihin kellonaikaan lämpötila on matalin? Entä korkein? Kuvaajassa pitäisi näkyä outo piikki. Tutki dataa ja selvitä onko kyse yksittäisestä virheestä vai säännönmukaisuudesta, jolle pitäisi keksiä syy!

2. (**1 piste**) Laske datan autokorrelaatiot 1, 2, ..., 30 tunnin viiveillä alkuperäisestä datasta ja piirrä/plottaa vastaava korrelogrammi. Poista sen jälkeen ensimmäiset 5 vuorokauden (120h) havainnot datan alusta ja muodosta tästä autokorrelaatiot ja korrelogrammi. Vertaa kuvia ja tulkitse, mistä erot kertovat! (Vinkki: Voit halutessasi käyttää kurssisivun työkalulistasta löytyvää autokorrelaatiot-pakettia (C-ohjelma + gnuplot-skripti) autokorrelaatioiden laskemiseen ja kuvien plottaukseen. Voit myös tehdä omat skriptit tai etsiä valmiita työkaluja, esim. R:lle.)

3. (**1 piste**) Tarkastellaan seuraavaa toimintojen tunnistusongelmaa: Viidelle lehmälle on asennettu kiihtyvyysanturit jotka mittaavat lehmän kiihtyvyyksiä jatkuvalla seurannalla. Lisäksi lehmiä on videoitu, jotta voidaan määrittää, mitä lehmä on tehnyt kullakin ajanhetkellä (vaihtoehdot: makaa, seisoo tai kävelee). Tehtävänä olisi muodostaa datajoukko, josta voidaan oppia luokittelija lehmän toiminnon tunnistamiseksi. Opetusjoukon alkioiksi halutaan 5s-mittaisia pätkiä, joiden kiihtyvyysmittauksista on laskettu erilaisia piirteitä ja joiden aikainen toiminto on tarkistettu videolta. Ongelmana on, että liian pieni datajoukko altistaa ylisovittumiselle, mutta toisaalta videoiden katselu on niin aikaavievää, että sen määrä haluttaisiin minimoida. Voisiko ylisovittumista välttää seuraavilla aikaasäästävillä datan generointistrategioilla? Perustele huolella!

   (i) Etsitään jokaiselta lehmältä 2 makaamis-, 2 seisomis- ja 2 kävelyhavaintoa, joista kukin on vähintään 5min pituinen. Jaetaan nämä 5s-mittaisiin pätkiin, jolloin opetusjoukkoon saadaan kustakin 5min toiminnosta 60 alkiota. Yhteensä datajoukon kooksi tulee siis 1800 alkiota. Videoita täytyy tarkistaa 150min ajalta (mahdollisesti pikakelauksella) + toimintojen etsintä.

   (ii) Etsitään jokaiselta lehmältä kymmenen 5-sekunnin mittaista makuu-, seisomis- ja kävelyhavaintoa. Yhteensä saadaan siis 150 data-alkiota, joiden tarkistus vaatii vain 12–13 minuuttia videoiden katselua + toimintojen etsintä. Koska data on liian pieni, kopioidaan jokainen alkio 10 kertaa, jolloin datajoukon kooksi tulee 1500 alkiota.

VI Menetelmien valinta ja vertailu

1. (**2 pistettä**) Tarkastele seuraavia luokittelumenetelmiä: Päätöspuu, Naiivi-Bayes, yleinen Bayes-luokittelija, neuroverkko, SVM, K lähintä naapuria. Mikä

tai mitkä menetelmät sopivat parhaiten seuraavanlaisten datojen luokitteluun? Entä mitä menetelmää ei ainakaan kannata käyttää? Perustele!

(a) Data sisältää tiedot 120:stä Tietorakenteet-kurssille ilmoittautuneesta opiskelijasta. Halutaan muodostaa luokittelija keskeyttäjien (20%) ja hylättyjen (15%) erottamiseen hyväksytyistä (loput 65%). Ennuste halutaan laatia jo ennen kurssin alkua, jotta riskitapauksille voidaan järjestää erityisohjausta. Muuttujat ovat opiskelijan sukupuoli ja ikä, montako vuotta opiskellut, suoritettujen opintopisteiden kokonaismäärä, Ohjelmointikurssin arvosana, matematiikan opinnot (on/ei ole) ja tieto käykö opintojen ohella työssä (kyllä/ei). Data on melko epäkonsistenttia, mikä on tyypillistä tällaiselle ihmisistä kerätylle datalle.

(b) Edellisen kohdan dataan on lisätty tiedot opiskelijan suorittamista harjoitustehtävistä (montako pistettä) viikottain. Haluattaisiin luokittelija, jonka ennusteita voisi päivittää joka viikko, kun saadaan selville uusia muuttujan arvoja. (Oletuksena että samat tehtävät toistuvat joka vuosikurssilla samoilla kurssiviikoilla.)

(c) Data koostuu 100 hiiren geenitiedoista. Hiiristä 50 sairastaa diabetesta ja loput 50 ovat kontrollieläimiä. Muuttujina on 2000 geenin eskpressioarvot (jotka kertovat miten aktiivinen ko. geeni on). Haluttaisiin muodostaa luokittelija joka ennustaa diabeteksen esiintymisen. Muuttujissa esiintyy kohinaa ja joissain jopa virheellisiä arvoja.

(d) Data koostuu 2000 alkiosta, joista kukin kuvaa 15s aikaikkunassa jonkin toiminnon tiedot. Mahdollisia toimintoja ovat makaa, istuu, seisoo, kävelee, juoksee. Aikaikkunassa lasketut muuttujat ovat kokonaiskiihtyvyyden minimi, maksimi, keskiarvo ja keskihajonta, sydämen syke ja keskimääräinen verenpaine. Toimintojen suorittamisessa on runsaasti yksilöllisiä eroja (esim. käveleekö hitaasti vai nopeasti, makaako lattialla vai löhööäkö sohvalla). Tavoitteena on luokittelija, jolla voisi tunnistaa henkilön toiminnot reaaliajassa.

2. (**1 piste**) Mitkä klusterointimenetelmät ja etäisyysmitat soveltuvat parhaiten ja mitkä huonoiten, mikäli datassa on

   (a) hyvin paljon rivejä (esim. 30 000), mutta vain muutamia muuttujia?

   (b) paljon ulkopuolisia pisteitä (outliers)?

   (c) yli 20 muuttujaa?

   (d) erikokoisia ja erimuotoisia klustereita?

   (e) päällekkäisiä klustereita?

3. (**1 piste**) Tiedonlouhintaprosessissa on kaikkein tärkeintä muistaa aina kysyä itseltään, mikä on järkevää. Ts. ovatko data ja tulokset järkevännäköisiä, onko jokin esiprosessointi- tai mallinnusmenetelmä järkevä ko. tilanteessa, onko mallien/hahmojen validointi järkevää. Keksi esimerkkejä, mitä järjetöntä voisi huolimaton tiedonlouhija tehdä seuraavissa tilanteissa! Millä tavalla keksimäsi vaarat voisi välttää?

- Datassa on sekaisin oikeita numeerisia muuttujia ja nominaalimuuttujia, joille on annettu numerokoodit.
- Datassa ei ole mitään selkeitä klustereita, vaan se on jakautunut tasaisen sattumanvaraisesti.
- Osa datasta on tullut vahingossa kahteen kertaan (eli datassa on paljon duplikaatteja).
- Datan selittävät muuttujat ovat keskenään vahvasti korreloivia, mutta luokittelun kannalta hyvin huonoja (ts. erottelevat huonosti luokkia).
- Dataa on kerätty vain parilta koehenkilöltä tai -eläimeltä ja näiltäkin vain lyhyeltä ajanjaksolta.