**Tiedonlouhinta 2013**
**Harjoitus 3**

Osasta tehtäviä saa 2 pistettä, mikäli sen on tehnyt kattavasti. Jos olet ratkaissut tehtävän vain puolittain/huolimattomasti, merkitse itsellesi vain 1p tehtävästä.

Tehtävissä 1 ja 2 käytetään datajoukkoa, jossa on tietoja 269 villirottanaaraasta. Datajoukot löytyvät osoitteista `http://cs.joensuu.fi/pages/whamalai/DM13/naaraatvalikoitu.csv` ja `http://cs.joensuu.fi/pages/whamalai/DM13/naaraatvalikoitudiskr.csv`.

Datajoukko naaraatvalikoitu.csv on pääosin numeerinen–ordinaalinen ja sen muuttujat ovat vaika=vuodenaika (3=loppusyksy–talvi, 2=kevät–alkukesä, 1=loppukesä–syksy), paino, lisamunratio=lisämunuaistenpainon ja painon suhde (+skaalaus), maksaratio=sama maksalle, pernaratio=sama pernalle, sydanratio=sama sydämelle ,thymusratio=sama thymukselle (kateenkorvalle, joka suurin poikasilla), umpisratio=sama umpisuolelle, neitsyt (2=neitsyt, 1=ei ole), rasitus=lisääntymisrasitus (1=ei kantava eikä imetä, 2=kantava tai imettävä, 3=kantava ja imettävä), gonratio=gonadirasvan ja painon suhde, uusrasvaratio=sama uusrasvalle, batratio=sama bat-rasvalle, vhaava=vatsahaavan aste (1=lievä,4=pahin), hanta=hännän pituus, hantaratio=hännän pituus/vartalon pituus, vart=vartalon pituus, painoind=paino/vartalon pituus, paikka=pyyntipaikan koodi (1=ensimmäinen kaatopaikka, 2=turkistarha, 3=toinen kaatopaikka).

Datajoukko naaraatvalikoitudiskr.csv on valmiiksi diskretoitu versio edellisestä.

Tehtävässä 4 käytetään lehmien kiihtyvyyksistä johdettua datajoukkoa, joka löytyy osoitteesta `http://cs.joensuu.fi/pages/whamalai/DM13/lehmakiiht.csv`. Datan luokkamuuttuja $C$ kertoo aktiivisuustason (nouseminen, laskeutuminen sekä rajummat päänliikkeet luokkaa 3, kävely luokkaa 2 ja makaaminen/seisominen luokkaa 1). Muut muuttujat ovat kiihtyvyydestä johdettuja piirteitä: a6=jerk-aallon keskimääräinen amplitudi, a16=$yz$-jerk-aallon amplitudin keskihajonta, a28=$xy$-jerk-aallon positiivisten piikkien keskimääräinen korkeus–leveys-suhde, a49=$y$-kiihtyvyyden keskihajonta, a52=$xz$-jerk-signaalin keskimääräinen magnitudi.

1. (**2 pistettä**) Analysoi datajoukon naaraatvalikoitu.csv kahden muuttujan välisiä korrelaatioita/riippuvuuksia. Voit käyttää haluamaasi työkalua (mieluiten sellaista joka laskee myös $p$-arvot eli korrelaation merkitsevyyden) tai hätätapauksessa osoitteesta `http://cs.joensuu.fi/pages/whamalai/DM13/korrelaatiot.tar.gz` löytyvää C-ohjelmaa (ei laske $p$-arvoja).

   a) Etsi datasta kaikki kahden muuttujan korrelaatiot Pearsonin, Kendallin ja Spearmanin korrelaatiokertoimilla. Tutki vain niitä, joissa jokin kolmesta on itseisarvoltaan vähintään 0.40. Analysoi sellaiset riippuvuudet, joissa yksi korrelaatiokerroin tuottaa muista selvästi poikkeavan tuloksen. Ovatko riippuvuudet uskottavia? Huomioi $p$-arvot, mikäli ohjelmasi kertoo ne. Plottaa ainakin pari esimerkkiä poikkeuksellisista riippuvuuksista. (Varaudu esittelemään kuvia harjoituksissa kalvokopiona tai koneelta.)

b) Milloin kaikki kolme tuottavat jokseenkin saman tuloksen? Entä milloin jokin poikkeaa? Koeta keksiä jotain systemaattisuutta, esim. riippuuko tulos muuttujien tyypistä tai ulkopuolisista? Vertaa havaintojasi siihen, mitä kirjallisuudessa kerrotaan korrelaatiokertoimien eroista.

2. (**2 pistettä**) Testaa seuraavia luokittelijoita naarasrottien vatsahaavan ennustamisessa (vain 2 luokkaa: on tai ei ole). Kaikki löytyvät Wekasta. Wekassa täytyy numeerisen datan kohdalla tehdä seuraavat esiprosessoinnit: aseta nominaalimuuttujat (ainakin vhaava ja paikka) (filtteri NumerictoNominal) ja yhdistä kaikki vatsahaavalliset luokat (asteet 2–4) yhteen (toista kahdesti MergeTwoValues-filtteri). (Voit toki yhdistää ei-vastahaavaluokat muutenkin ennen Wekaa.) Mitä johtopäätöksiä voit tehdä vertailun perusteella?

   a) Numeerinen data: J48-päätöspuu.
   b) Numeerinen data: neuroverkko (Multilayer Perceptron). Kokeile ainakin seuraavia: oletusarvoparametrit, vain yksi piilotaso ja nolla piilotasoa.
   c) Numeerinen data: $K$-nearest neighbours (Wekassa IBk). Anna ohjelman määrätä itse $K$:n arvo. (Voit kokeilla myös asettaa sen itse.)
   d) Numeerinen data: Bayes-luokittelija (BayesNet).
   e) Diskretoitu data: J48.
   f) Diskretoitu data: Bayes-luokittelija. Kokeile paraneeko malli, jos valikoit muuttujia.

3. (**1 piste**) Testaa mahdollisimman kattavasti Samin tekemää MutualInformation-korrelaation laskinta. Ensisijaisesti kannattaa varmaan testata oman harjoitustyönne datajoukkoa, mutta jos siinä ei ole sopivia numeerisia muuttujia tai oletat kaikkien riippuvuuksien olevan nätisti lineaarisia (vahva Pearsonin korrelaatio), voit kokeilla edellisen tehtävän naarasrottadatalla. Etsi esimerkkejä riippuvuuksista jotka ovat MutualInformationin perusteella vahvoja mutta eivät lineaarisia tai päinvastoin. Osaatko antaa niille selityksen? Plottaa ainakin pari kiinnostavaa esimerkkiä (ja varaudu esittelemään niitä harjoituksissa).Testaa myös annetun $K$-arvon (diskretointivälien lkm) vaikutusta tuloksiin.

   Ohjelma löytyy osoitteesta `http://cs.joensuu.fi/pages/whamalai/DM13/mutuali.zip`

4. (**1 piste**) Muodosta monen muuttujan lineaariregressiomalli lehmakiiht.csv-datasta lehmän aktiivisuustason ennustamiseksi. (Tämän voi tehdä mm. Wekalla.) Miten hyvän ennustavan mallin saat? (Esim. keskimääräinen virhe ristiinvalidoinnilla.) Voiko mallia käyttää deskriptiiviseen mallinnukseen? Mitä se silloin kertoo?

5. **Lisätehtävä** (**1 lisäpiste**) Koeta parantaa edellisen tehtävän mallia joko PCA:lla tai ICA:lla tai kokeilemalla muita regressiomenetelmiä (SVM, neuroverkko, tms.).