Tiedonlouhinta 2013
Harjoitus 4

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

Tiedostossa kyselysanat.csv on lueteltu eräässä esitietokyselyssä esiintyneet sanat (lukuunottamatta joitain merkityksettömiä "stopwordeja"). Sanojen oikeinkirjoitusta ei ole tarkoituksella korjattu.

Tiedostossa hiirirottaerot.csv on tiedot 11466 geenistä. Muuttujat ovat geeni=geenin nimi, hikrom=sijaintikromosomi hiirellä, rokrom= sama rotalla, hiero=geenin aktiivisuusero sairailla mutanttihiirillä verrattuna terveisiin, roero1=vastaava ero vastasyntyneillä rotilla verrattuna nuoriin aikusiin ja roero2=vastaava ero vanhoilla rotilla verrattuna nuoriin aikuisiin.

Tiedostossa syontiakt.txt on kuvattu navetan ruokailuaktiivisuutta noin kuukauden ajalta. Jokaiselta tunnilta (kenttä h) on seuraavat arvot: ruokailun kokonaisfrekvenssi (kaikki syönnit automaateilla), kokonaiskulutus (kiloina), Mansikin ruokailufrekvenssi ja kulutus sekä Mustikin ruokailufrekvenssi ja kulutus.

Huom! Weka on hieman hidas klusteroimaan suuria datajoukkoja (voi kestää jopa minuutin). Voit halutessasi käyttää muitakin työkaluja.

1. (**2 pistettä**) Klusteroi hiirien ja rottien geenierodataa (hiirirottaerot.csv) K-means-algoritmilla (vaihtoehtoisesti voit käyttää malliperustaista klusterointia, jos Sinulla on tehokas työkalu/kone; Wekassa siis nimenä EM). Wekassa täytyy kertoa, että klusteroinnissa käytetään vain kolmea numeerista muuttujaa. Jos et halua kuitenkaan hukata muiden muuttujien arvoja, käytä apuna FilteredClusterer:ia. Jos valitset lisäksi MakeDensityBasedClusterer, saat tietää klusterien keskiarvot ja keskihajonnat. (Kummankin kanssa voi kutsua haluamaansa klusterointialgoritmia.) Kokeile ainakin paria K:n arvoa ja aja klusterointi useaan kertaa, jotta näet erot. Vaihtelevatko klusterien sisällöt merkittävästi eri kerroilla? (Vertaa ainakin klusterien kokoja ja centroideja). Mitä ulkopuolisia (outliers) löydät? Vaihtuvatko ulkopoliset eri kierroksilla? Kokeile klusteroida sekä kaikkien 3:n muuttujan suhteen että valita vain jotkin 2, jolloin voit visualisoida tuloksia. (Wekassa klikkaamalla tulosta oikealla napilla ja valitsemalla Visualize Cluster Assignments.)

   (**2p**, jos kokeilit sekä 2:n että 3:n muuttujan suhteen ja tarkistit muuttuivatko outlierit.)

2. (**2 pistettä**) Testaa hierarkisia klusterointimenetelmiä kyselysanat.csv-datalla. Sanat kannattaa klusteroida erittäin moneen klusteriin (esim. 500-1000:een) ja etäisyysmittana pitää käyttää editointietäisyyttä (edit distance). Huomaa, että Wekassa muuttujan tyyppi on ensin muutettava merkkijonoksi filtterillä NominalToString.

   a) Vertaile eri linkitysmetriikkojen (linktype) tuottamia klusterointeja. (Voit

visualisoida hierarkiaa klikkaamalla tulosta oikealla napilla ja valitsemalla Visualize tree.) Kokeile muuttuvatko tulokset, jos sotket datan järjestystä. Mikä metriikka vaikuttaisi vakaimmalta? (Tässä kohdassa voit halutessasi käyttää vain satunnaista otosta datasta, jos sen analysointi on helpompaa.)

b) Tutki miten järkeviä/yhdenmukaisia sanaklustereita menetelmät tuottavat. (Klusterin sisällön tutkimisesta Wekassa oli ohjeita mikroharjoituksien tehtävissä.) Vertaa ainakin paria K:n arvoa ja paria metriikkaa. Voisiko menetelmää käyttää saman sanan eri muotojen tunnistukseen tesktidatan analysoinnissa? Tai esiintyykö klustereissa muuten samantyyppisiä (esim. saman sanaluokan) sanoja?

3. (**2 pistettä**) Temporaalisen datan visualisointi ja tulkintatehtävä.

   a) Visualisoi syontiakt.txt-dataa siten, että näet muuttuja-arvojen kehityksen ajan funktiona. Voit halutessasi ottaa vain pienemmän palan tiedostosta tarkasteluun (kuitenkin useita vuorokausia). Tutki sekä koko navetan että Mansikin ja Mustikin syöntiaktiivisuutta (määriä ja frekvenssejä). Mitä voit päätellä kuvista? Varaudu esittelemään kuvia harjoituksissa.

   b) Laske jokaisen vuorokauden 24:lle tunnille keskiarvot (ja mielellään myös keskihajonnat tai keskivirheet) kaikista muuttujista. Esitä tulokset graafisesti. Vertaa Mansikin ja Mustikin keskimääräistä syöntiaktiivisuutta toisiinsa ja navetan keskiarvoon. Mitä saat selville? (Jos laskit myös keskihajonnat/-virheet, näet miten paljon vaihtelua vuorokausien välillä on.)

4. (**1–2 pistettä**) Ideointitehtävä. Eläinten käyttäytymisdatan koostuu perusmuodossaan sarjasta kellonaikoja ja eläimen senhetkisiä toimintoja. Vartalon ja suun toiminnot voidaan kirjata erikseen. Sarja voisi olla esim. (1,seisoo,marehtii),(15,seisoo,ei mitaan),(140,laskeutuu, ei mitaan),(520,makaa,ei mitaan),(600,makaa,märehtii), missä kellonaika on sekunteina alkuhetkestä.

   Keksi temporaalisia mallinnusmenetelmiä tällaiselle datalle! Menetelmän tulisi siis jotenkin ottaa huomioon aikadimensio tai toiminnon ajallinen konteksti (esim. toistuvat toimintoketjut) ja paljastaa jotain uutta (ainakin potentiaalisesti) kiinnostavaa tietoa. Menetelmäsi voi perustua vain uusien piirteiden muodostukseen, joita sitten mallinnetaan jollain olemassaolevalla menetelmällä (esim. klusterointi, riippuvuussäännöt, autokorrelaatio, Markovketju) tai voit keksiä aivan uuden menetelmän. Koeta arvioida kuinka paljon dataa (esim. havaintovuorokausia) menetelmäsi vaatii. Vaihtoehtoisesti voit myös keksiä aivan uuden havainnollisen tavan visualisoida tällaista dataa. Havainnollista ideaasi jollain esimerkillä! Kuvat ovat aina hyviä. (**Voit itse päättää onko suunnitelmasi 1 vai 2 pisteen arvoinen** riippuen esim. ajankytöstä tai miten toteuttamisvalmis ideasi on.)

5. (**1 piste**) Lupaudu täyttämään kurssikysely! Kurssikysely aukeaa vasta kurssin lopussa ja saat vielä muistutuksen siitä.