Tiedonlouhinta 2013
Harjoitus 1 (mikroharjoitukset)

Harjoitusten oheismateriaali löytyy sivulta `http://cs.joensuu.fi/pages/whamalai/DM13/harjoitus1.html`.

Lataa koneellesi datatiedostot worldstat.csv (numeerista dataa), worldmaitta.names (nominaalidataa transaktiomuodossa) ja worldmaitta.codes (sama data, mutta nominaaliattribuutit korvattu numeerokoodeilla). Voit tallettaa myös viimeeksi mainittuihin liittyvän kooditaulun (koooditaulu.txt), joka kertoo mitä muuttujaa mikin koodi vastaa.

Datan jokainen rivi vastaa jotain maailman valtiota. Näitä kuvaavat muuttujat on kerrottu tiedoston worldstat.csv alussa. Nominaalidatassa samat muuttujat on binarisoitu ja lisäksi mukana on joitain uusia attribuutteja. (Ks. kooditaulu.txt. Attribuuttien sanalliset nimet pyrkivät olemaan mahdollisimman havainnollisia.)

Harjoituksia wekalla

Käynnistä weka-ohjelma ja avaa tiedosto worldstat.csv.

Huom! Aina jos teet virheen, voit perua sen "Undo"-painikkeesta! Sillä voi perua myös viimeiset muutokset.

1. Valitse Preprocess (on oletusarvo alussa).
2. Datan katselu: Katso miltä muuttujien jakaumat näyttävät (visualize all).
3. Puuttuvien arvojen (datassa "?") muunto: Valitse filtteri klikkaamalla "Choose". →Unsupervised→Attribute→ReplaceMissingValues. Klikkaa filtterin nimeä "Choose"-napin vierestä, jolloin aukeaa ikkuna. "More" kertoo lisää kyseisestä filtteristä. Aseta "IgnoreClass" todeksi ja paina "ok". Valitse sitten "Apply".
4. Klusterointi: Valitse "Cluster". Merkitse "Country"-attribuutti turhaksi (IgnoreAttributes). Valitse haluamasi klusterointimenetelmä (klikkaa "Choose"). Klikkaa parametrien asetusikkuna vanhaan tapaan, jolloin voit lukea lisätietoja, mitä menetelmä tekee. Asetusikkunassa voit asettaa mm. klusterien lukumäärän (jotkin menetelmät osaavat määrittää sen automaattisesti) sekä etäisyysfunktion. Kokeile ainakin jotain menetelmää. Osaatko tulkita tuloksia?
5. Klusterin sisällön tutkiminen. Klusterien jäsenten luettelu on wekassa hieman hankalaa, mutta voit tehdä sen seuraavasti: Klusteroituasi datan, klikkaa tuloslistasta haluttua klusterointia oikealla napilla. Valitse valikosta "Visualize cluster assignments". Valitse sieltä "Save", joka tallettaa datan klusterin jäsentiedoilla arff-tiedostoon (kuin csv, mutta alussa extrakuvauksia). Avaa

tiedosto editorissa ja poista alkukuvaukset. Järjestä rivit klusteri-kentän perusteella. Nyt saman klusterin jäsenet on lueteltu peräkkäin.

Varoituksen sana Wekan aiempien versioiden klusterointialgoritmit eivät aina toimineet oikein, joten kannattaa aina tarkistaa saadun klusteroinnin järkevyys. (Erityisesti että tulostiedostossa alkio liitetään oikeaan klusteriin. Itse klusterien kuvaukset ovat tiettävästi oikein.)

6. Attribuuttien muunto: Wekan filttereillä voi muuntaa attribuutteja, mutta useimmat toiminnot ovat aika rajoittuneita (usein on helpompi tehdä oma pikkuskripti haluttuun muuntoon). Muuntoon käytettävät filtterit löytyvät Preprocess-ikkunasta kun klikkaat "Filter"-sanan alta "Choose". Tutki hieman millaisia toimintoja filttereistä löytyy. Kun olet valinnut jonkin filtterin, klikkaa filtterin nimeä "Choose"-painikkeen vierestä, jolloin aukeaa säätöikkuna. Siinä voit säätää parametreja. Lisätietoja saat klikkaamalla "More". Kun olet asettanut parametrit mieleisiksesi, klikkaa "Apply". Voit perua muutokset painikkeesta "Undo".

   Kokeile ainakin seuraavaa:

   (a) Muunna attribuutti Ex-colony nominaaliseksi luokkamuuttujaksi filttereillä Unsupervised→Attribute→ NumericToNominal. Varmista ettei muita attribuutteja muunneta nominaalisiksi!

   (b) Valitse sitten uusi filtteri, Supervised→Attribute→Discretize. Tämä pyrkii diskretoimaan attribuutit siten, että diskretointi olisi luokittelun kannalta mahdollisimman hyvä.

7. Ennen luokittelua poista attribuutti "Country".

8. Luokittelu: Valitse "Classify". Klikkaa "Choose" ja valitse luokittelija. Kokeile ainakin seuraavaa: Trees→J48. Sitten klikkaa "Choose"-painikkeen vierestä sanaa J48, jolloin aukeaa parametrien asetusikkuna. Tarkista, että "Ex-colony" on luokka ja paina "Start". Tutki päätöspuuta. Mikä on tärkein attribuutti (lähinnä puun juurta)? Kuinka hyvä luokittelija on ristiinvalidoinnin perusteella? Osaatko tulkita, mitä "Confusion matrix" kertoo? Huom! Mallin voi esittää myös visuaalisesti, mikäli se ei ole liian monimutkainen - klikkaa "Result list" ja valitse "Visualize tree".

9. Assosiaatiosäännöt: Wekan Associate etsii perinteisiä "assosiaatiosääntöjä", jotka voivat olla varsin järjettömiä (yleensä ei mitään tekemistä tilastollisten 'assosiaatioiden kanssa). Tätä toimintoa ei siis kannata käyttää! Oikeiden tilastollisten riippuvuussääntöjen etsintään löytyy muita ohjelmia.

10. Attribuuttien valinta: Wekan "Select attributes"-toiminto pyrkii valitsemaan luokkamuuttujaa parhaiten ennustavat attribuutit. Tarjolla on monia menetelmiä. "Attribute Evaluator" luettelee menetelmiä (lisätietoja vanhaan tapaan "More":lla) ja Search method" hakualgoritmeja. (Joihinkin menetelmiin on tarjolla useita hakualgoritmeja, kun taas toiset vaativat tiettyä hakualgoritmia.) Jos datan koko vain sallii, "Exhaustive Search" takaa globaalisti optimaaliset tulokset. Näitä toimintoja voit tutkia, jos aikaa jää.

Lisätietoja ja -ohjeita `http://www.cs.waikato.ac.nz/ml/weka/documentation.html` (katso esim. Miscellaneous information).

Riippuvuussääntöjen haku Kingfisherillä

Seuraavaksi harjoitellaan komentorivipohjaisen ohjelman asennusta ja käyttöä Linuxissa. Tämän tyyppisiä tehokkaita linux-työkaluja löytyy runsaasti eri tarkoituksiin. Asennus tapahtuu samalla tavalla ja kun osaa käyttää yhtä, osaa yleensä käyttää kaikkia muitakin.

Käynnistä ensin linux (jos et tiedä, miten se tapahtuu mikroluokassasi, kysy vierustoverilta tai opettajalta).

1. Talleta kingfisher1_1.tar.bz2-ohjelmapaketti koneellesi. Pura paketti komennolla "tar -xvjf kingfisher1_1.tar.bz2". README-tiedostossa on kerrottu perusohjeet. Käännä ohjelma kirjoittamalla komento "make". Nyt ohjelma on valmis käyttöön. Saat listauksen komentoriviparametreista kirjoittamalla vain "kingfisher".

2. Datatiedostona käytetään nyt tiedostoa worldmaitta.codes. Voit siirtää sen samaan hakemistoon, johon purit kingfisherin (muuten Sinun pitää antaa tiedoston koko polku komentoriviparametrina).

3. Kokeile komentoa "kingfisher -i worldmaitta.codes -k41 -M-40". Parametrit kertovat datatiedoston lisäksi, että datassa esiintyvät attribuutit numero 0..41 ja oletusarvoisen mittafunktion (joka on Fisherin p:n logaritmi) raja-arvo on -40. (Mitä pienempi raja-arvo, sitä merkitsevämpiä sääntöjen täytyy olla.) Ohjelma tulostaa oletusarvoisesti 100 parasta sääntöä sekä niihin liittyviä mittalukuja. Katso kooditaulusta, mitä paras sääntö kertoo. Huomaa, että mato $\sim$ merkitsee attribuutin negaatiota (esim. $X \rightarrow\sim C$ kertoo, että X:n ja $\neg C$:n välillä on positiivinen riippuvuus eli X:n ja C:n välillä negatiivinen riippuvuus).

4. Koeta nyt itse toteuttaa seuraavat haut katsomalla komentoriviparametrien selostukset:

   - Etsi vain positiiviset riippuvuudet.
   - Etsi vain 10 parasta sääntöä.
   - Etsi korkeintaa 3 attribuuttia sisältävät säännöt. (Attribuuttien monimutkaisuuden rajaaminen keventää hakua ja voi siten olla tarpeen todella raskaiden datajoukkojen kohdalla. Jos ohjelma joskus "juuttuu" laskemaan, pysäytä se ja kokeile tätä.)
   - Karsi säännöt joiden frekvenssi on alle 0.2. (Minimifrekvenssillä voi keventää hakua, mutta se voi hukata merkitseviä riippuvuuksia. Käytä tätä vain hätätapauksessa!)

- Vaihda mittafunktio χ^2:ksi (chi2) ja keksi sille sopiva raja-arvo, siten että löydät ainakin 100 sääntöä. (Jos raja-arvosi on liian tiukka, et löydä niin montaa. Raskailla datoilla tarpeeksi tiukka raja-arvo taas keventää hakua. Huomaa, että raja-arvo tehostaa vain haun alkua. Kun ohjelma on löytänyt 100 raja-arvoa parempaa sääntöä, se tiukentaa raja-arvoa automaattisesti.)

Huom! Jos ohjelma joskus juuttuu todella raskailla datajoukoilla, voit keskeyttää sen näppäinyhdistelmällä ctrl-c. (Sama toimii kaikkien linux/unix-ohjelmien kohdalla.)

Lisätehtävä (jos jää aikaa)

1. Asenna ja käännä apuohjelma namescodes. Sillä voi muuntaa merkkijonoja numerokoodeiksi ja päinvastoin.
2. Etsi riippuvuussääntöjä kingfisherillä kuten edellä. Talleta 100 parasta riippuvuussääntöä tiedostoon kingfisherin formaatissa 2 (optiot -p ja -o).
3. Muunna sääntöjen numerokoodit nimiksi ohjelmalla namescodes (tämä vaatii kooditaulun).
4. Tutki muunnettua tulostiedostoa. Näyttävätkö löydetyt riippuvuudet mielestäsi uskottavilta?