Harj.2 Teht.1

Yhteistä

Standardoimalla (normalisoimalla) muodostettavat piirteet: Mikäli datassa on useisiin eri viiteryhmiin kuuluvia alkioita, esim. naaras- ja urosrotat, tyypin 1 ja tyypin 2 rehua syövät lehmät, kannattaa yksilölliset erot huomioida. Meitä kiinnostavat erityisesti sellaiset arvot, joissa eläin tai havainto poikkeaa viiteryhmästään. Siksi alkuperäisestä muuttujasta X (esim. paino, hännänpituus, aktiivisuus vuorokaudessa) muodostetaan uusi standardoitu muuttuja X'. Jos vanhan muuttujan arvo on x_i, on uuden muuttujan arvo

$$x'_i = \frac{x_i - avg_R(X)}{stdev_R(X)},$$

missä $avg_R(X)$ on X:n keskiarvo joukossa (viiteryhmässä) R ja $stdev_R$ vastaava keskihajonta. (Kyseessä on siis tilastotieteen z-transformaatio.)

Uusi muuttuja X' noudattaa yleensä suurinpiirtein standardinormaalijakaumaa eli sen keskiarvo on 0 ja keskihajonta 1. Jos $X' > 2$, voi arvoa pitää normaalia merkitsevästi korkeampana ja jos $X' < -2$, normaalia merkitsevästi alhaisempana. (Huom! Jos alkuperäisen X:n minimiarvo oli 0, voi uuden muuttujan X jakauma katketa vasemmalla kuin seinään.)

Esim. rotille kannattaa laskea tällä tavalla standardoidut arvot naaraiden ja urosten joukossa. Samoin jos datassa (esim. lokidata) on suhteellisen vähän eläimiä, esim. 20 lehmää, ja kustakin paljon havaintoja, kannattaa tutkia, milloin eläin käyttäytyy omaksi itsekseen poikkeavalla tavalla. Nyt viiteryhmänä ovat siis kaikki saman yksilön havainnot.

Jos dataa on todella paljon, voi jopa tutkia, poikkeaako eläimen käytös sen normaalikäytöksestä ko. vuorokauden aikana. Eli nyt viiteryhmä koostuisi saman yksilön tietyn vuorokauden ajan mittauksista.

Käyttö: Diskretoi, luo sopivia binäärimuuttujia, ja etsi riippuvuussääntöjä Kingfisherin uudella versiolla (siinä voi kieltää eheysehdoilla triviaalit säännöt, tyyliä $(X = x_1) \rightarrow (X' = x'_1)$). Näin kaikki X:stä muodostetut muuttujat voivat olla yhtaikaa mukana ja niistä löytää parhaat (joilla kiintoisimpia riippuvuuksia).

Voit myös tutkia mikä X:stä johdettu piirre on paras numeerisen datan mallissa, esim. lineaariregressiossa. Nyt mukana voi kuitenkin olla kerrallaan vain yksi X:stä johdettu muuttuja (tai voit saada triviaaleja riippuvuuksia).

Huom Voit toki standardoida muullakin tavalla. Pääasia on, että voit vertailla eri yksilöitä ryhmäänsä ja mahdollisesti myös itseensä.

a-kohta

Koska rotan ikää ei tiedetä, kannattaa laskea suhteellisia painoja: esim. rotan elimen paino/koko paino. Samoin kannattaa muodostaa painoindeksi, paino/vartalon pituus, jotta ali- ja ylipainoiset erottuvat (pituuteensa nähden painavat eivät välttämättä ylipainoisia, sillä kantavat rotat painavia). Pitkähäntäisyys on kiinnostava tekijä, joten kannattaa laskea myös häntäindeksi, esim. hännän pituus/vartalon pituus.

b-kohta

Koska aktiivisuus on mitattu vain tunneittain, kannattaa muutkin muuttujat laskea tunneittain. Esim. kesto-muuttujat (kauanko aikaa käyttänyt makaamiseen, seisomiseen, tms.) ja frekvenssimuuttujat (montako kertaa on noussut ylös tai lähtenyt liikkeelle). Makuulla vs. seisaalla märehtiminen on kiintoisaa (terve lehmä märehtii makuulla), joten kummallekin toiminnolle kannattaa luoda omat muuttujat.

c-kohta

Standardoiduista muuttujista oli jo edellä. Kannattaa laskea peruspiirteitä jossain valitussa aikaikkunassa. (Esim. 0.5–1h ikkunat?). Tavallisia piirteitä ovat (kokonais- tai osa-)kiihtyvyyden keskiarvo, keskihajonta (hyvä), "energia" eli neliösumma (määritä ensin origo! esim. ko. 1h pätkältä x:n, y:n ja z:n keskiarvot tai mediaanit – jälkimmäinen usein parempi), minimi, maksimi, xy (ym.) -korrelaatiot sekä kiihtyvyyspiikeistä lasketut heuristiset piirteet.

d-kohta

Alkutoimi: luetellaan vastauksissa esiintyvät sanat yleisyysjärjestyksessä ja käydään listaa läpi. Kaikkein yleisimpiä apusanoja (ja, tai, hän, se, jne.) ei kannata ottaa mukaan.

Helppo tapa: Määritetään listasta sopivia sanoja, joiden esiintymiä etsitään. Jos esim. eläinlajit ovat perusmuodossa, ei ole taivutusongelmaa.

Monimutkaisempi ehdotus (klusterointia): Koetetaan tunnistaa sanan vartalot edes jollain tarkkuudella. (Kokeile esim. klusterointia edit distancella tai jollain itse keksityllä etäisyysfunktiolla.) Tuloksena saatavat klusterit ovat uusia "supersanoja" (eli toivon mukaan samaa sanaa tarkoittavia, vaikka isokaan virhe ei välttämättä haittaa). Klusterien koon perusteella karsitaan kaikkein harvinaisimmat (outlierit) ym. turhina pidettyjä. Jäljelle jäävistä

klustereista tulee uudet muuttuja-arvot $c_1, \ldots, c_K$, joilla korvataan klusteriin kuuluvan sanan esiintymät.

Sen jälkeen voidaan määrittää vastaukselle piirteitä: esim. joka klusterimuuttujalle c_i, esiintyikö se vastauksessa vai ei tai montako kertaa se esiintyi (on olemassa hienostuneempiakin mittoja, jotka huomioivat sanaklusterin yleisyyden).

Nyt vastauksia voidaa klusteroida uusien piirteiden suhteen tai etsiä jotain muita hahmoja.

Vaihtoehtoisesti vastaukset voi koettaa jakaa kahteen ryhmään HITS-algoritmilla. Taas kannattaa määrittää ensin sanavartalot (sanaklusterit). (Ilman vartalon määritystäkin toimii jotenkin, mutta silloin saisi olla pitkiä dokumentteja?) HITS:issä voi toisena osapuolena olla kokonaisia vastauksia (idea: kaksi vastausta on samankaltaisia, jos ne sisältävät samoja sanoja ja kaksi sanaa samanlaisia, jos esiintyvät samanlaisissa vastauksissa). (Virkepohjaltakin voisi keksiä jotain.)

Muuta: Vastauksien pituus (merkkimäärä, sanojen määrä) tai sanojen keskimääräinen määrä virkkeessä voivat myös kertoa jotain vastaajista.

e-kohta

Standardoiduista muuttujista alussa. Kannattaa luoda "kumulatiivisia muuttujia", jotka pitävät kirjaa edellisistä tapahtumista. Esim. monastiko lehmä on edellisen 30min tai 1h aikana kirjautunut syömään? Monestiko näistä samalla automaatilla kuin nyt? Paljonko lehmä on syönyt edellisen 0.5h, 1h, 2h tms. aikana? Keskimäärin (syöntikerralla)? Tai millä keskinopeudella?

Riippuvuuksia etsiessä huomaa triviaalien riippuvuuksien ongelma (ks. alusta).

f-kohta

Liiketunnistimen havainto, olettaen ettei lemmikkejä (eikö lasketa "asukkaisiin"?). Lyhytaikaiset vedenkäytöt (poikkeaa pesukoneesta tms. mutta voi missata suihkun tai kylvyn?). Mahdollisesti sähkön kulutus (ellei ohjelmoitavia laitteita). Kun takka alkaa lämmetä (huoneenlämmöstä; voi jatkaa lämpiämistä, vaikka puita ei enää lisättäisi). Oikeassa datassa myös hiilidioksidipitoisuus – lienee hyvä (kotiviinipytty tuottanee hiilidioksidia tasaisesti koko ajan?). Koska muutokset edellisiin ajanhetkiin kiinnostavia, kannattaa luoda muutoksia kuvaavia muuttujia. Esim. Olk. $S(t)$ sähkönkulutus ajanhetkellä (aikavälillä) t, missä t:t ovat 1h aikaikkunoita. Lasketaan uudeksi muuttujaksi $d(t) = S(t) - S(t-1)S(t-1)$. Nyt ei haittaa, että talossa on

jatkuvasti päällä sähköä kuluttavia laitteita, koska luku näyttää vain muutoksen.

g-kohta

Merk. 6 eri ikäryhmää $v_1, \ldots, v_6$. Mahdollisia osavälejä kaikki ikäryhmät erikseen $v_1, \ldots, v_6$ sekä erilaiset yhdistelmät $[v_1, v_2]$, $[v_1, v_3]$, $[v_1, v_4]$, $[v_1, v_5]$, $[v_2, v_3]$ jne. Kaikkia kannnattaa tutkia, sillä emme tiedä, missä iässä geeni aktiivisin/passiivisin.

Ryhmien välisten poikkeamien tutkimiseksi kannattaa kokeilla erilaisia suhdelukuja. Esim. ryhmien $[v_1, v_3]$ ja $[v_4, v_6]$ erojen kuvaamiseksi lasketaan joka geenille g suhdeluku

$$d(g) = \frac{Avg(Expr(g|[v_1, v_3])) - Avg(Expr(g|[v_4, v_6]))}{Avg(Expr(g|[v_4, v_6]))}$$

eli erotus keskimääräisessä ekspressioarvossa ikäryhmien $[v_1, v_3]$ ja $[v_4, v_6]$ välillä jaettuna jommallakummalla (tai jollain muulla referenssiarvolla, esim. ikäryhmä nuoret aikuiset, v_4 on voitu valita referenssiksi – usein koeasettelussa on etukäteen päätetty "kontrolliryhmä", josta saadaan referenssiarvot).

Kun suhdelukuja binärisöidään (riippuvuussääntöanalyysiä varten), voidaan keskittyä vain merkitsevästi poikkeaviin arvoihin. Tehdään suhdeluvulle X standardisointi, kuten alussa on kuvattu, mutta luodaan binäärimuuttujat vain poikkeaville arvoille, kun $X > 2$ (tai $X > 3$ tms. kannattaa katsoa X:n jakaumaa) ja kun $X < -2$ (tms.). Huom! jos et tiedä sopivia raja-arvoja, voit luoda muuttujia useille vaihtoehdoille (tyyliin binäärimuuttujat arvoille $X > 2$, $X > 3$, $X > 4$) ja asettaa eheysrajoituksen, joka kieltää niitä esiintymästä samassa säännössä. Transaktiodatassa luetellaan geenin rivillä vain ne binäärimuuttujat, joiden arvo on 1 (tosi), joten mielenkiinnottomien geenien kohdalle ei tule montaa 1-arvoa. (Ne tarvitaan silti mittalukuen laskentaan.) Tuloksena saatava data on paljon kevyempää analysoida, kuin jos mukana olisi normaaliarvot eli binäärimuuttujat $-2 \leq X <\leq 2$.

Huom! Tehtävänannossa ehdotettu skaalaus samalle arvovälille ei ehkä tarpeellista. Jos sen haluaa tehdä, voi esim. määrittää uuden suhdeluvun X maksimi (yli kaikkien geenien) ja jakaa suhdeluvut sillä.