CC by 4.0

Harjoitukset 1

Harjoituksissa käytetään aineistoja, jotka ovat vapaasti saatavilla verkossa “UCI Machine Learning Repository” tietokannasta. Aineistojen yhteydessä on .names tiedosto, josta löytyy tarkempi kuvaus kyseisestä aineistosta.

Ennen klusterointia kannattaa selvittää esimerkiksi:
• Onko aineistoissa puuttuvia arvoja ja miten puuttuvat arvot on koodattu?
• Millaisia muuttujat ovat? Onko aineistossa luokkamuuttujaa (tätä muuttujaa ei oteta huomioon klusterointia tehdessä, mutta sitä voidaan hyödyntää klusteroinnin arvioinnissa)?

1. Tutkitaan R:n cluster.datasets paketin birth.death.rates.1966 aineistoa.
a) Asenna R:n cluster.datasets paketti.
b) Ota käyttöön cluster.datasets paketti.
c) Lataa birth.death.rates.1966 aineisto käyttäen data() funktiota. Mitä aineiston ensimmäinen sarake sisältää? Entä toinen ja kolmas sarake?
d) Piirrä pisteparvi aineistosta käyttäen plot() funktiota. Näetkö pisteparvesta selkeitä klustereita? Lisää text() funktiota käyttäen pisteparveen kutakin pistettä vastaava maan nimi.
e) Suorita K-means klusterointi aineistolle käyttäen kmeans() funktiota. Klusteroi aineisto kahteen klusteriin. Käytä muiden parametrien kohdalla funktion antamia oletusarvoja.
f) Havainnollista klusterointitulosta käyttäen plot() funktiota. Aseta plot() funktion col parametrille arvoksi K-means-algoritmin lopputuloksena saatu klusterileima (muuttujan_nimi$cluster) aineiston kullekin havainnolle.

2. Tutkitaan UCI-tietokannan “Wine” aineistoa.
a) Hae UCI-tietokannasta Wine aineisto ja talleta se tietokoneellesi .data formaatissa.
b) Lataa Wine aineisto R/RStudioon read.table() funktion avulla.
c) Valitse vähintään kaksi erilaista kahden piirteen osajoukkoa aineistosta. Tarkastele valitsemiasi piirrejoukkoja graafisesti. Onko piirrejoukoissa näkyvissä selkeitä klustereita?
d) Suorita K-means klusterointi valituille piirrejoukoille. Voit valita vapaasti klusterien määrän.
e) Vertaa havainnoille lopputuloksena saatuja klusterileimoja havaintojen luokkaleimoihin käyttäen table() funktiota. Tulkitse saamasi taulukot.
f) Toista vaiheet d) ja e) käyttäen eri parametriarvoja.

3. Tutkitaan edelleen “Wine” aineistoa.
a) Valitse piirrejoukoksi aineiston kaikki muuttujat paitsi ensimmäisessä sarakkeessa oleva luokkamuuttuja.
b) Suorita esiprosessointia muuttujille (esim. scale() funktio R:ssä).
c) Suorita K-means klusterointi esiprosessoidulle ja esiprosessoimattomalle piirrejoukolle. Voit valita parametriarvot haluamallasi tavalla.
d) Muodosta sekaannustaulukko table() funktiolla. Saatko eroaesiprosessoidun ja esiprosessoimattoman datan sekaannustaulukkojen välille?

CC by 4.0

4. Jatketaan “Wine” aineiston tutkimista.
a) Muodosta kuvaaja, missä x-akselin arvo vastaa klusterien lukumäärää ja y-akseli kuvaa tot.withinss arvoa (saadaan K-means-algoritmin lopputuloksena). Muodosta siis pistejoukko, missä kunkin pisteen ensimmäinen komponentti vastaa klusterien lukumäärää ja jälkimmäinen komponentti on klusterien lukumäärää vastaava tot.withinss arvo. Voit valita vapaasti tutkittavat muuttujat (piirteet) sekä tutkittavat klusterien lukumäärät.
b) Määritä kuvaajasta taitekohta eli “polvi”. Taitekohdasta löytyy aineistolle sopiva klusterien lukumäärä.
c) Testaa kmeans_plot() funktiota, joka löytyy paketista Kmisc. Tulkitse kuvaaja.

5. Tutkitaan UCI:n “Pima Indians Diabetes” aineistoa.
a) Selvitä, miten puuttuvat arvot on esitetty ja missä muuttujissa puuttuvia arvoja on.
b) Käsittele puuttuvat arvot valitsemallasi tavalla.
c) Sovella K-means-algoritmia aineistoon käyttäen eri parametriarvoja. Analysoi tulokset.
6. Sovella K-means-algoritmia UCI:n “Letter Recognition” aineistoon ja analysoi tulokset. Voit valita piirrejoukon/piirrejoukot vapaasti.

7. Tutustu K-means++ -algoritmiin (funktio kmeanspp() LICORS paketissa) ja sovella sitä valitsemiisi aineistoihin.

8. Tutustu K-medoids-algoritmiin (Partitioning around medoids, pam() funktio cluster paketissa) ja sovella sitä valitsemiisi aineistoihin.