**Harjoitukset 2**

Harjoitustehtävissä käytetään seuraavia aineistoja:
• Letter Recognition dataset
• Wine dataset
• R:n “cluster.datasets” paketin “Birth and Death Rates” aineisto

1. Tutkitaan birth.death.rates.1966 aineistoa.
a) Lataa cluster.datasets paketti käyttöön käyttäen library() funktiota.
b) Lataa birth.death.rates.1966 aineisto käyttöösi käyttäen data() funktiota.
c) Muodosta aineistosta pisteparvi käyttäen plot() funktiota.
d) Muodosta etäisyysmatriisi (etäisyysmatriisi kertoo aineistossa olevien havaintojen pareittaiset etäisyydet käyttäen valittua etäisyysmittaa) käyttäen dist() funktiota. Muodosta etäisyysmatriisi käyttäen funktion oletusarvoja parametreille eli anna funktiolle syötteenä vain data.
e) Suorita hierarkkinen klusterointi käyttäen hclust() funktiota. Käytä funktion oletusarvoja eli anna funktiolle syötteenä vain viittaus muodostettuun etäisyysmatriisiin.
f) Muodosta dendrogrammi käyttäen plot() funktiota (viittaus muuttujaan, missä hierarkkisen klusteroinnin lopputulos on).
g) Lisää dendrogrammin alimmalle tasolle maiden nimet numerokoodauksen asemesta (labels parametri plot() funktiossa).
h) Tulkitse dendrogrammi.

2. Jatketaan birth.death.rates.1966 aineiston tutkimista.
(a) Tutki, miten etäisyysmitan valinta (method parametri dist() funktiossa) ja klusterien etäisyysmitan valinta (method parametri hclust() funktiossa) vaikuttavat dendrogrammin muotoon.
(b) Muodosta dendrogrammi käyttäen plot() funktiota.
(c) Sovella rect.hclust() funktiota muodostettuun dendrogrammiin. Funktio havainnollistaa muodostuneita klustereita piirtäen suorakulmiot muodostuneiden klusterien ympärille. Klusterien muodostamisen voi tehdä joko leikkaamalla dendrogrammin halutulta korkeudelta (h parametri funktiossa) tai antamalla haluttu klusterien lukumäärä (k parametri funktiossa).
(d) Tulkitse dendrogrammi.

3. Jatketaan birth.death.rates.1966 aineiston tutkimista.
a) Määritä luokkaleima aineiston havainnoille sen mukaan, mihin maanosaan kukin maa kuuluu. Luokkien koodauksen voit valita vapaasti.
b) Muodosta dendrogrammi aineistosta käyttäen dist() ja hclust() funktioita. Etäisyysmitat dist() ja hclust() funktioissa voit valita haluamallasi tavalla.
c) Sovella cutree() funktiota muodostettuun dendrogrammiin. Funktio muodostaa dendrogrammista klusterit joko leikkaamalla sen halutulta korkeudelta (h parametri funktiossa) tai käyttäjän antaman klusterien lukumäärän mukaan (k parametri funktiossa).

d) Muodosta sekaannustaulukko vertaamalla havainnoille annettuja luokkaleimoja ja havaintoja vastaavia klusterileimoja keskenään käyttäen table() funktiota. Tulkitse saamasi sekaannustaulukko.

4. Sovella hierarkkista klusterointia “Wine” aineistoon. Voit valita käytettävät muuttujat vapaasti (Huom! Dendrogrammin voit muodostaa, vaikka sinulla olisi yli kolme muuttujaa käytössä.). Voit tehdä halutessasi esiprosessointia muuttujille. Sinun ei tarvitse käyttää koko aineistoa, vaan voit ottaa halutessasi aineistosta pienemmän osajoukon tarkasteluun.

5. Sovella hierarkkista klusterointia “Letters Recognition” aineistoon. Voit valita käytettävät muuttujat vapaasti. Voit tehdä halutessasi esiprosessointia muuttujille. Sinun ei tarvitse käyttää koko aineistoa, vaan voit ottaa aineistosta pienemmän osajoukon tarkasteluun.