1 Introducció

En aquesta pràctica treballarem l'anàlisi d'agrupaments. Aplicarem alguns algoritmes d'aquest àmbit per analitzar dades multivariants. Tal com s'ha fet a les pràctiques anteriors, cada grup d'estudiants ha de realitzar el guió de la pràctica a continuació, realitzant els càlculs i gràfics necessaris en l'entorn R. Els resultats obtinguts i les respostes a les preguntes s'han de recollir en un document amb la vostre solució, fet per exemple en R markdown, o bé en Microsoft Word, o latex. Procureu posar els vostres noms i cognoms i número de grup a l'inici del document. De matrius grans nomes cal incloure les primeres 5 files a l'informe. Empleneu la mateixa numeració dels ítems de l'enunciat al document amb la vostre solució. Lliureu la solució del qüestionari **en format PDF**, pujant-la a l'entorn Atenea (atenea.upc.edu), a l'apartat de la tasca corresponent, abans de la data final de la tasca.

2 Qüestionari

En aquesta pràctica treballem amb dades de llavors de blat. S'han pres diferents mesures dels llavors mitjançant radiografies dels llavors. Les variables quantitatives recollides en l'estudi són:

- àrea (A)
- perímetre (P)
- compactesa ($C = 4\pi A/P^2$)
- longitud del gra
- amplada del gra
- coeficient d'asimetría
- longitud de la ranura del gra
- varietat

Els llavors pertanyen a tres varietats de blat diferent, anomenades "Kama", "Rosa" i "Canadian", i n'hi han 70 de cada varietat. Les dades s'han extret d'un arxiu de dades per machine learning (archive.ics.uci.edu). Feu els càlculs i els gràfics que es demanen a continuació.

1. Carregueu el fitxer `seedsdataset.dat` a l'entorn R. Etiqueteu les columnes de la matriu amb els noms de les variables segons l'ordre de la llista.
2. Separeu la variable grup (la varietat) de la resta guardant-la per separat, creant una matriu que conté únicament les variables quantitatives.
3. (1p) Feu boxplots de cadascuna de les variables per detectar possibles anomalies. Existeixen anomalies extremes?
4. (1p) Utilitzeu `pairs` per fer les diagrames bivariants i valorar si n'han anomalies bivariants.

5. (1p) Calculeu la matriu de distàncies euclidianes entre llavors, després d'estandarditzar les dades. Doneu les primers 5 files i columnes de la matriu.

6. (1p) Feu un agrupament jeràrquic amb el mètode del veí més proper utilitzant la funció `hclust`. Genereu el dendrograma amb `plot`. Resulten detectables les varietats en el dendrograma?

7. (1p) Repetiu l'agrupament jeràrquic amb el mètode del veí més lluny (farthest neighbour), generant un segon dendrograma.

8. (2p) A quina distància (aproximadament) s'hauria de tallar el dendrograma per crear tres grups? Realitzeu aquest tall amb `cutree`. Feu la taula creuada de les categories que acabeu de crear amb la variable varietat. L'anàlisi ha identificat bé les varietats? Quin tan percent dels llavors s'identifiquen correctament, suposant que els clústers detectats han de coincidir amb les varietats?

9. (1p) Proveu de la mateixa manera el "average linkage", i utilitzeu en dendrograma per fer dos grups. Feu la taula creuada de les categories que acabeu de crear amb la variable varietat. Que observeu?

10. (2p) Feu un agrupament jeràrquic amb el criteri de Ward, fent de nou el dendrograma. Talleu per fer tres grups. Feu la taula creuada de les categories que acabeu de crear amb la variable varietat. Quin percentatge dels llavors es classifiquen correctament?

11. (1p) Feu un agrupament no-jeràrquic sinó divisiu, utilitzant la funció `kmeans`. Per fer els resultats reproduibles, fixem la llavor de l'algorisme de generació de numeros aleatoris amb `set.seed(123)` abans de cridar `kmeans`. Fem servir `kmeans` fent tres grups. Feu de nou la taula creuada de la classificació. Funciona kmeans millor que el criteri de Ward per aquestes dades?

12. (2p) Installem el paquet `clusterSim` amb `install.packages`. D'aquest paquet utilitzarem la funció `index.G1` que calcula el pseudo F-statistic de Calinski-Harabasz. Fem un bucle per $k = 2$ fins a 6 clusters. A cada iteració del bucle cridem `kmeans` per fer un clustering amb k clusters, i amb els grups obtinguts calculem el pseudo F-statistic. Grafiqueu l'estadístic F en funció del número de clusters. Segons aquest criteri, quin és el número de clústers mes adequat per utilitzar amb k-means?

13. (2p) Feu agrupament basat en models, instal.lant el paquet `mclust`. Feu l'agrupament assumint una mixtura de tres components. Grafiqueu els resultats amb un plot del model obtingut, amb opció `what="classification"`. Feu de nou la taula creuada de la classificació. Quin tan per cent de les observacions són mal classificades?

14. (1p) Feu una llista de les llavors mal classificades pels tres mètodes simultàniament: jeràrquic amb criteri de Ward, kmeans i l'agrupament basat en models. Quantes llavors són? Doneu els numeros de les files corresponents de la matriu de dades.

15. (2p) Feu un anàlisi de components principals de les variables quantitatives estandaritzades. Feu el biplot de la matriu de dades utilitzant un color diferent per a cada varietat de llavor. Feu servir un símbol diferent per les llavors que els tres mètodes no classificaven bé. Que observeu?

16. (2p) Creeu que és possible fer un agrupament que produeix grups que coincideixen millor amb les tres varietats de blat? Comenteu i/o exploreu possibles vies d'anàlisi o indiqueu les limitacions.