



PAC 1

Presentació

Primera activitat d'avaluació continuada del curs. En aquesta PAC es practicaran els algorismes bàsics de categorització.

Objectius

L'objectiu d'aquesta prova d'avaluació és categoritzar les dades dels arxius adjunts relacionats amb problemes de fertilitat. Volem agrupar els pacients segons si tenen l'esperma normal o alterat.

L'arxiu de dades fertility.csv té un format tipus taula, on cada fila correspon a un exemple. L'última columna és la classe i la resta corresponen als atributs de l'exemple. L'arxiu adjunt attributes.txt conté la descripció d'aquests atributs.

Aquests arxius pertanyen al problema "Fertility" del repositori d'aprenentatge de l'UCI:

<http://archive.ics.uci.edu/ml/>

Solució de la PAC

Exercici 1

L'objectiu d'aquest exercici és mirar si es pot categoritzar l'arxiu petit de dades (small.csv). En concret, se us demana:

1. Efectueu, si és necessari, el tractament previ de les dades. Justifiqueu totes les decisions que prengueu.

Ens proporcionen un arxiu amb dades relacionat amb clients d'un banc.

1	1	0	0.8	0	0.88	N
0	1	0	0.8	1	0.31	O
1	1	0	1.0	-1	0.38	N
1	0	0	0.8	-1	0.50	O



0	1	0	1.0	-1	0.38	N
0	1	0	0.8	-1	0.25	O

No hi ha valors absents, no els hem de tractar. Tots els atributs són numèrics i els seus valors estan entre -1 i 1, no hem de normalitzar.

2. Utilitzeu el k-means (nítid) per categoritzar les dades del arxiu esmentat en dues categories, ignorant les columnes no pertinents. Quin és el nivell de precisió¹ del resultat?

Apliquem el k-means nítid amb la distància euclídea i dues categories a totes les columnes excepte l'última (classe).

Hem seleccionat els dos primers exemples com a centroides inicials. Calculem les distàncies euclídees entre els centroides i tots els exemples. Els exemples pertanyen a la categoria del centroide més proper. Tornem a calcular els centroides fent la mitjana dels valors dels atributs (atribut a atribut) de tots els exemples de cada categoria. Repetim aquest procés fins que no es produeixin canvis en les categories. Obtenim els centroides finals:

0.5	0.75	0	0.9	-1.0	0.3775
0.5	1.00	0	0.8	0.5	0.5950

I com a categories finals:

Cat.1: {3, 4, 5, 6}

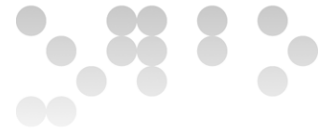
Cat.2: {1, 2}

Si assignem la classe "N" a la categoria 1 i la classe "O" a la categoria 2, obtenim una $precisió = \frac{1+2}{6} = 50\%$.

El k-means depèn molt dels centroides inicials escollits. Si agafem com a centroides inicial els dos últims exemples obtenim els centroides:

0.5	0.75	0	0.9	-1.0	0.3775
0.5	1.00	0	0.8	0.5	0.5950

¹ Per calcular la precisió heu de comparar la categoria resultant per cada exemple amb la seva classe (última columna de l'arxiu).



I les categories:

Cat.1: {1, 2, 3, 4}

Cat.2: {5, 6}

I, per tant, la mateixa precisió que abans.

3. Apliqueu l'algorisme del PCA per reduir la dimensionalitat del conjunt anterior conservant el 95% de la variància. Utilitzeu el k-means (nítid) sobre el conjunt reduït de la mateixa forma que en l'apartat anterior. S'obtenen resultats comparables?

Per realitzar aquest exercici, hem utilitzat el PCA des de l'R. A continuació teniu les ordres concretes:

```
dt <- read.table("small.csv", head=F, sep=",")
pca.dt <- princomp(dt[,1:6])
vars.dt <- pca.dt$sdev ^ 2
v <- vars.dt / sum(vars.dt)
v
v[1]+v[2]
v[1]+v[2]+v[3]
```

Amb la penúltima línia ens dona un 88% i amb l'última un 98%. Per tant, hem de menester 3 components. A partir d'aquí fer el que segueix per generar l'arxiu amb aquestes components:

```
dt2 <- cbind(pca.dt$scores[,1:3], as.character(dt$V7))
write.table(dt2, "small.pca.csv", row.names=FALSE,
  col.names=FALSE, sep=",", quote=FALSE)
```

El resultat és:

```
0.3862758  0.63754049 -0.35687523  N
1.5792840 -0.07223096  0.20743016  O
-0.5717919  0.18730193 -0.37010943  N
-0.7651724  0.58004798  0.52854625  O
-0.3185035 -0.65316982 -0.03284547  N
-0.3100920 -0.67948961  0.02385372  O
```

Apliquem el k-means comen l'exercici anterior. Obtenim els centroides:

```
-0.4913900 -0.1413274  0.03736127
0.9827799  0.2826548 -0.07472254
```

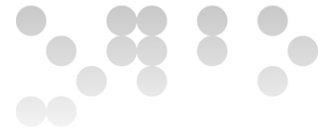
I les categories:

Cat.1: {3, 4, 5, 6}



Cat.2: {1, 2}

Si assignem la classe “N” a la categoria 1 i la classe “O” a la categoria 2, obtenim una $precisió = \frac{1+2}{6} = 50\%$.



Exercici 2

L'objectiu d'aquest exercici és categoritzar les dades amb el mètode aglomeratiu. En concret, se us demana la construcció de tres dendrogrames per al conjunt de dades de l'exercici anterior (small.csv) utilitzant la distància euclídea i els mètodes del lligam simple, el lligam complet i la mitja com a criteris d'aglomeració. Doneu les precisions de la mateixa forma que en l'exercici anterior.

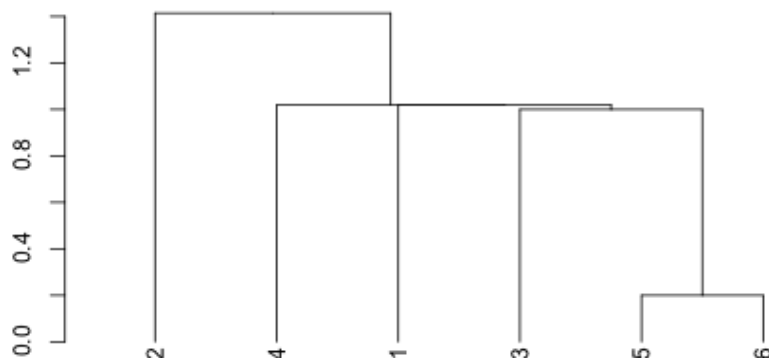
El primer que hem de fer és calcular la matriu de distàncies. Podeu convertir la matriu en una de semblances o treballar directament amb distàncies. Aquestes distàncies les obtindrem de cada parell d'objectes aplicant la distància euclídea.

	1	2	3	4	5
2	1.414214				
3	1.019804	2.244994			
4	1.414214	2.449490	1.019804		
5	1.428286	2.009975	1.000000	1.428286	
6	1.414214	2.000000	1.019804	1.414214	0.200000

A partir d'aquest moment hem de separar en tres processos en funció del mètode d'agregació. Per a l'enllaç simple obtindrem les agrupacions següents:

$\{5, 6\}$
 $\{3, 5, 6\}$
 $\{1, 3, 5, 6\}$
 $\{1, 3, 4, 5, 6\}$
 $\{1, 2, 3, 4, 5, 6\}$

i el dendrograma que segueix:





i per tant, considerant dues categories per l'ordre d'agrupació queden els grups:

Cat.1: {2}

Cat.2: {1, 3, 4, 5, 6}

Que correspon a una precisió de:

$$precisió = \frac{1+3}{6} = 67\%$$

Per l'enllaç complet obtenim les agrupacions:

{5, 6}

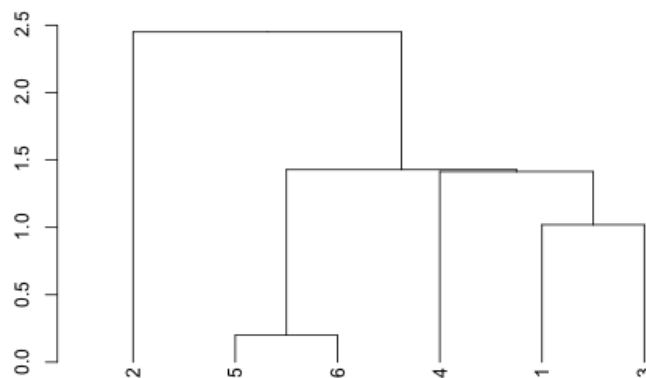
{1, 3}

{1, 3, 4}

{1, 3, 4, 5, 6}

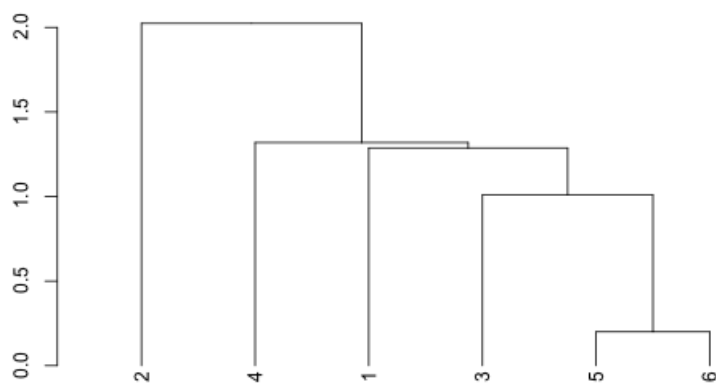
{1, 2, 3, 4, 5, 6}

I el dendrograma que segueix:



i, per tant, les mateixes categories i precisió.

Per l'enllaç mitja obtenim les mateixes agrupacions que per al primer cas i, per tant, les mateixes categories i precisió; i el dendrograma que segueix:





Exercici 3

L'objectiu d'aquest exercici és utilitzar una eina per a categoritzar l'arxiu adjunt fertility.csv en dues categories. Aquesta eina s'anomena Weka i la teniu a la seva plana Web:

<http://www.cs.waikato.ac.nz/ml/weka>

En concret, se us demana l'aplicació del mètode del k-means (el trobareu com a SimpleKMeans) i un altre mètode a escollir sobre les dades de l'arxiu esmentat. Heu de mostrar els resultats comparant-los amb l'última columna. Apliqueu el PCA per reduir la dimensionalitat conservant el 95% de la variància i torneu a aplicar els mateixos algorismes que al conjunt original.

Hem realitzat una sèrie de proves per als algorismes SimpleKMeans, Dendrograma amb enllaç simple, FarthestFirst i el EM. No hem aplicat cap tractament als atributs. A continuació tenim taules que resumeixen els resultats obtinguts i les matrius de confusió:

Algorisme	Precisió	Erronis	Iteracions
2-means	51%	49	3
Dendrograma	84%	16	N/A
FarthestFirst	64%	36	N/A
EM	51%	49	N/A

Algorisme	Matriu
2-means	44 44 5 7
Dendrograma enllaç simple	84 4 12 0
FarthestFirst	58 30 6 6
EM	44 44 7 5

I aplicant el PCA (ens queden 6 atributs):

Algorisme	Precisió	Erronis	Iteracions
2-means	67%	33	8
Dendrograma	79%	21	N/A
FarthestFirst	66%	34	N/A
EM	79%	21	N/A



Algorisme	Matriu
2-means	28 60 7 5
Dendrograma enllaç simple	11 77 2 10
FarthestFirst	60 28 6 6
EM	15 73 6 6



Exercici 4

Realitzeu una valoració global comparant els mètodes i els diferents exercicis i redacteu unes conclusions globals sobre l'aplicació dels mètodes a aquests conjunts de dades. Els criteris de correcció de la PAC invaliden una A si tots els processos no estan ben justificats i comentats.

En aquest apartat s'espera que extrèieu conclusions generals sobre l'exercici. Aquestes conclusions dependran dels resultats obtinguts. A mode d'exemple enumerarem algunes de les qüestions sobre les que podeu argumentar:

- La possibilitat de que sigui pràctic l'aplicació d'algun dels mètodes sobre el problema donat. En cas que no, que faltaria afegir.
- Comparativa dels diferents mètodes emprats. Enumeració dels avantatges i inconvenients en funció de: aplicació del PCA, precisió, eficiència, categories, models...
- La representació del problema. Com es comporten els atributs? És una bona representació? Com afecta el preprocés de les dades al funcionament dels algorismes.
- Avantatges dels models que generen els diferents mètodes. Comparativa dels models generats durant tot l'exercici.
- En general, intent de justificació i/o explicació dels resultats que es van obtenint: fixant-se no només en la precisió.
- Com es comporten els algorismes en funció del nombre d'exemples d'entrenament que es disposen?
- Quin cost computacional té cadascun dels mètodes?