



PAC 1

Presentació

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

Objectius

L'objectiu d'aquesta prova d'avaluació és categoritzar les dades dels arxius adjunts relacionats amb problemes de distribució d'aliments. Volem agrupar els clients segons si el tipus (venta al detall o a l'engròs) i/o regió de procedència (Lisboa, Oporto o altres). Aquest conjunt de dades representa dos problemes en un.

Els arxius de dades "csv" tenen un format tipus taula, on cada fila correspon a un exemple. La primera i/o segona columna representen les classes de cadascun dels problemes esmentats en el paràgraf anterior i la resta corresponen als atributs de l'exemple. L'arxiu adjunt "txt" conté la descripció d'aquests atributs.

Aquests arxius pertanyen al problema "Wholesale Customers" 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'una empresa de distribució d'aliments. No hi ha valors absents, no els hem de tractar.

Els atributs són numèrics i els seus valors són molt dispersos. Decidim aplicar estandardització per tal de normalitzar. A continuació es mostren, respectivament, els promitjos i les desviacions estàndard:

4010.5, 4912.83, 8320.83, 553.33, 3834.5, 1893.33



3917.47, 4013.36, 7681.74, 365.52, 4287.18, 960.76

i el resultat global del procés:

-0.14	1.08	-0.77	0.62	-0.82	-0.28
-0.86	-1.21	-1.07	-1.31	-0.89	-1.96
-0.52	1.44	1.45	-0.88	1.81	0.27
-0.05	0.09	0.85	-0.65	0.67	1.34
-0.58	-0.31	0.59	0.74	0.14	0.27
2.15	-1.09	-1.05	1.47	-0.89	0.35

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 precisió del resultat?

Apliquem el k-means nítid amb la distància euclídea i dues categories a totes les columnes excepte la primera (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:

0.17	0.24	0.21	0.26	0.18	0.39
-0.86	-1.21	-1.07	-1.31	-0.89	-1.96

I com a categories finals:

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

Cat.2: {2}

Si assignem la classe "1" a la categoria 2 i la classe "2" a la categoria 1,

obtenim una $precisió = \frac{1+3}{6} = 66,6\%$.

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

-0.43	0.22	0.21	-0.29	0.18	-0.07
2.15	-1.09	-1.05	1.47	-0.89	0.35

I les categories:

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

Cat.2: {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?

Apliquem el PCA sobre el conjunt de l'arxiu. Obtenim els percentatges de variàncies:

0.52 0.32 0.096 0.044 0.023 0.000

I acumulant:

0.52 0.84 0.93 0.98 1 1

El resultat és:

0.61 0.02 1.57 0.03
1.43 -2.73 -0.45 0.13
-2.84 -0.09 0.16 0.50
-1.40 0.62 -0.78 0.04
-0.29 0.32 -0.10 -1.08
2.49 1.85 -0.39 0.38

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

-0.29 0.55 0.09 -0.03
1.43 -2.73 -0.45 0.13

I les categories:

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

Cat.2: {2}

Si assignem la classe "1" a la categoria 2 i la classe "2" a la categoria 1,

obtenim una $precisió = \frac{1+3}{6} = 66,6\%$.

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.



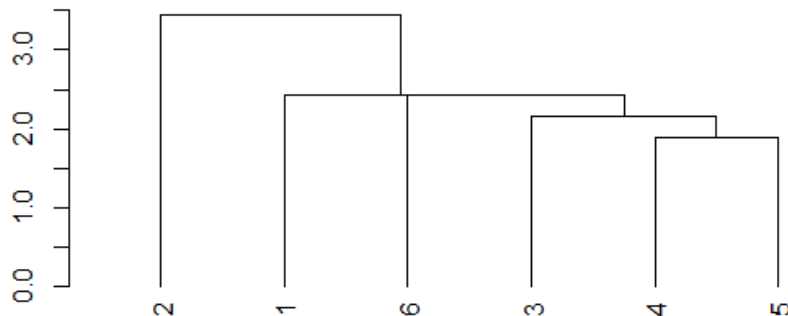
El primer que hem de fer és calcular la matriu de distàncies. Podeu convertir la matriu en una de semblances o treballar directament amb distàncies. Aquestes distàncies les obtindrem de cada parell d'objectes aplicant la distància euclídea.

	1	2	3	4	5
2	3.447821				
3	3.940444	5.177261			
4	3.324996	4.489787	2.160154		
5	2.458088	3.846205	3.036605	1.891969	
6	2.433036	3.622116	5.142614	3.729196	2.423339

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:

{4, 5}
 {3, 4, 5}
 {3, 4, 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:

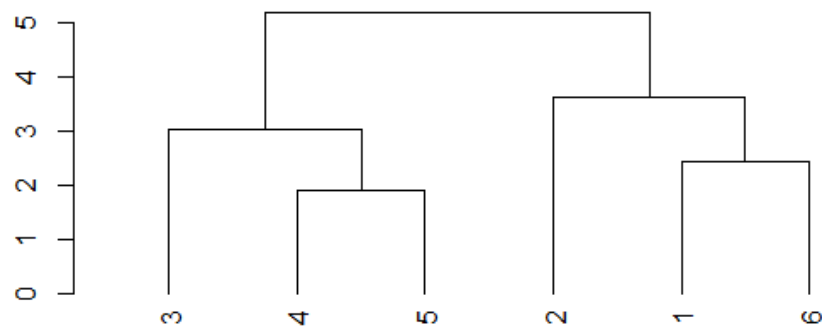
$$precisió = \frac{1+3}{6} = 67\%$$



Per l'enllaç complet obtenim les agrupacions:

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

Cat.2: {3, 4, 5}

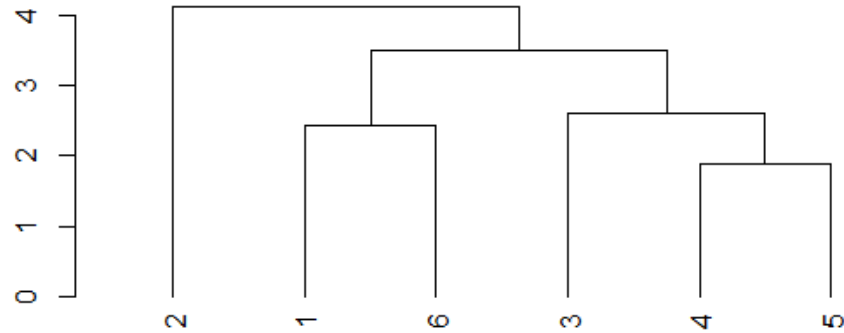
Que correspon a una precisió de:

$$precisió = \frac{3+3}{6} = 100\%$$

Per l'enllaç mitja obtenim les agrupacions:

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

$$precisió = \frac{1+3}{6} = 67\%$$



Exercici 3

L'objectiu d'aquest exercici és utilitzar una eina per a categoritzar l'arxiu adjunt "Wholesale customers.csv" en dues o tres categories, segons la primera i/o segona columna (una vegada per problema). 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.

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 per al problema representat en la primera columna ("channel"):

Algorisme	Precisió	Erronis	Iteracions
2-means	77,5%	99	14
Dendrograma	67,5%	143	N/A
FarthestFirst	68,9%	137	N/A
EM	82,5%	49	N/A

Algorisme	Matriu
2-means	97 45 296 2
Dendrograma enllaç simple	142 0 297 1
FarthestFirst	137 5 298 0
EM	102 40 37 261



I a continuació els de la segona columna ("region"):

Algorisme	Precisió	Erronis	Iteracions
3-means	58,6%	182	14
Dendrograma	71,4%	126	N/A
FarthestFirst	70,9%	128	N/A
EM	42,7%	252	N/A

Algorisme	Matriu
3-means	240 47 29 60 10 7 35 4 8
Dendrograma enllaç simple	314 1 1 77 0 0 47 0 0
FarthestFirst	311 4 1 77 0 0 45 1 1
EM	157 131 28 42 29 6 26 19 2



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:

- Les diferències entre els dos problemes, tenint en compte les dues primeres columnes.
- 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?