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 els guanys d'una persona a partir de la informació del cens. Volem agrupar persones segons si superen o no els 50k dòlars anuals.

Els arxius de dades tenen un format tipus taula, on cada fila correspon a un exemple. L'última columna representa la classe i la resta corresponen als atributs de l'exemple. L'arxiu adjunt "txt" conté la descripció dels atributs.

Aquests arxius pertanyen al problema "Census Income Data Set" 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 guany anuals de persones a partir de dades del cens. No hi ha valors absents, no els hem de tractar.

Els atributs són molt diferents. Tenim un atribut numèric (*age*), un cardinal (*education*) i tres nominals o binaris (*marital-status*, *occupation* i *native-country*). El numèric el tractarem aplicant *ranging* ((valor – mínim) / (màxim – mínim)). Al cardinal li assignarem 0, 0.5 o 1 en funció del seu valor i als nominals 0 o 1. El resultat global serà:

```
0.17 0.5 1 0 0 0.71 0 pobre
0.48 1 1 0 0.71 0 1 pobre
1.0 0 1 0 0.71 0 1 ric
0.28 1 0 0 0 0.71 1 ric
0.66 0.5 1 0 0.71 0 1 ric
```

```
0.0 0.5 0 0.71 0 0 1 pobre
```

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 euclidea i dues categories a totes les columnes excepte l'última (classe).

Hem seleccionat els dos primers exemples com a centroides inicials. Calculem les distancies euclidees 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 0.50 1.00 0.00 0.00 0.71 0.00
0.48 0.60 0.60 0.14 0.42 0.14 1.00
```

I com a categories finals:

Cat.1: {1}

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

Si assignem la classe "pobre" a la categoria 1 i la classe "ric" a la categoria 2, obtenim una $precisió = \frac{1+3}{6} = 67\%$.

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

```
0.58 0.50 1.00 0.00 0.53 0.18 0.75
0.14 0.75 0.00 0.35 0.00 0.35 1.00
```

I les categories:

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

Cat.2: {4, 6}

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

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:

```
0.50 0.28 0.14 0.07 0.00 0.00
```

I acumulant:

```
0.50 0.79 0.93 1.00 1.00 1.00 1.00 1.00
```

El resultat és:

```
0.26 1.06 0.18 -0.08
-0.41 -0.13 -0.46 -0.34
-0.88 -0.14 0.32 0.32
0.84 -0.10 -0.43 0.35
-0.60 -0.13 -0.07 -0.05
0.78 -0.57 0.46 -0.21
```

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

```
0.26 1.06 0.18 -0.08
-0.05 -0.21 -0.04 0.02
```

I les categories:

Cat.1: {1}

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

Si assignem la classe "pobre" a la categoria 1 i la classe "ric" a la categoria 2, obtenim una $precisió = \frac{1+3}{6} = 67\%$.

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 euclidea 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 euclidea.

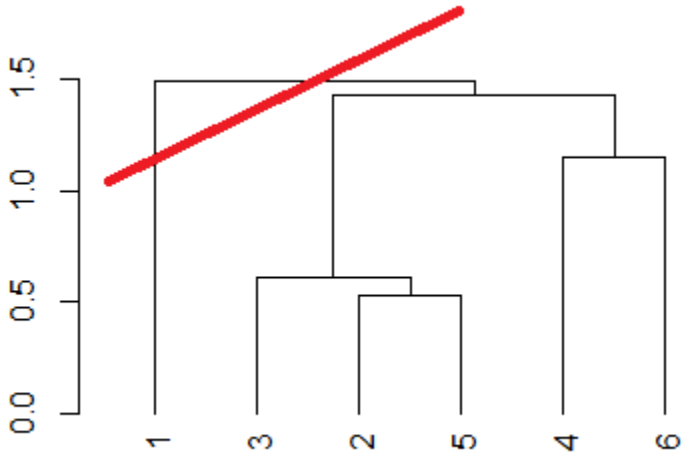
```
          1         2         3         4         5
2 1.5317682
3 1.7131547 1.1258502
4 1.5035630 1.4292677 1.8773321
5 1.4943413 0.5288918 0.6073764 1.5472157
```

```
6 1.7406110 1.5757715 1.8027756 1.1515641 1.5586054
```

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:

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

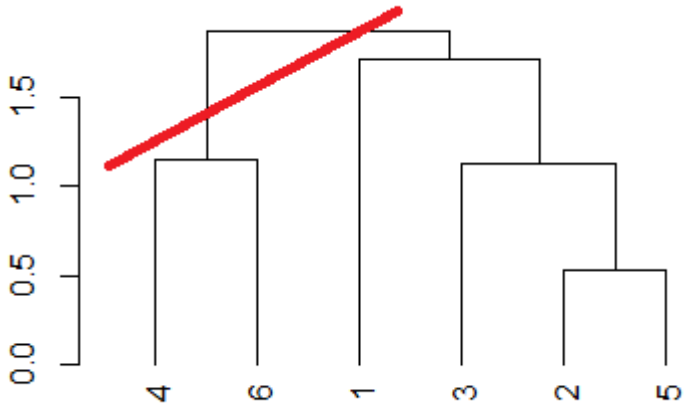
Que correspon a una precisió de:

$$precisió = \frac{1+3}{6} = 67\%$$

Per l'enllaç complert obtenim les agrupacions:

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

i el dendrograma que segueix:



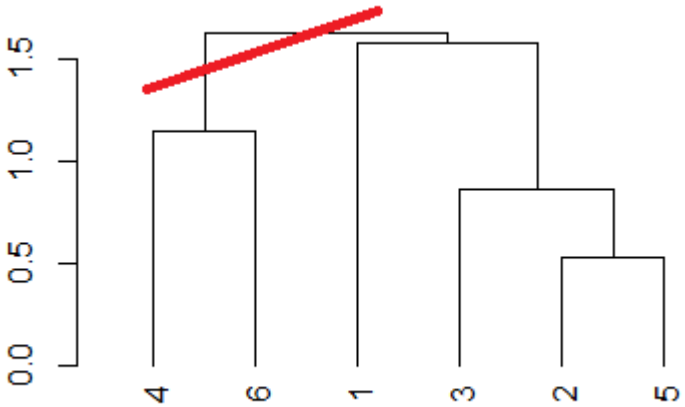
i per tant, considerant dues categories per l'ordre d'agrupació queden els grups:

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

Que correspon a una precisió de:

$$precisió = \frac{2+1}{6} = 50\%$$

Per l'enllaç mitja obtenim les mateixes agrupacions que el cas anterior i el dendrograma que segueix:



i per tant, les mateixes categories i precisió.

Exercici 3

L'objectiu d'aquest exercici és utilitzar una eina per a categoritzar l'arxiu adjunt "adult.data.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.

Hem realitzat una sèrie de proves per als algorismes SimpleKMeans, FartherFirst i el EM. Hem aplicat els filtres ReplaceMissingValues a tots els atributs i el Normalize als numèrics i el NominalToBinary als nominals.

A continuació tenim taules que resumeixen els resultats obtinguts i les matrius de confusió:

Algorisme	Precisió	Erronis	Iteracions
2-means	61%	9392	8
FarthestFirst	63%	8699	N/A
EM	62%	9255	N/A

Algorisme	Matriu
2-means	16484 8236 1156 6685
FarthestFirst	23601 1119 7580 261
EM	16663 8057 1198 6643

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?