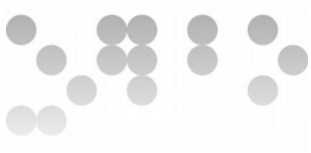


PAC 1

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 l'origen de distints vins a partir de la seva composició i color. Volem agrupar els vins en funció del seu origen.

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 "Wine Data Set" del repositori d'aprenentatge de l'UCI:

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

Solució de la PAC

Exercici 1

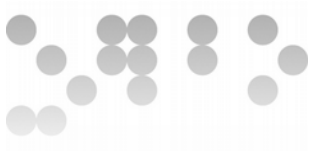
En aquest exercici haureu de mirar si es possible categoritzar l'arxiu petit de dades (small.csv). La darrera columna en aquest arxiu denota la classe. En particular, se us demana:

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

Se'ns proporciona un arxiu amb dades relacionades amb la composició de distints vins. No hi ha valors absents i, per tant, no s'han de tractar. Tots els atributs són numèrics amb valors molt dispersos, a excepció de la classe (que no s'ha de tractar). Per tal de normalitzar les dades decidim aplicar estandardització. A continuació es mostren, respectivament, els promitjos i les desviacions estàndard:

```
13.02 2.23 105.25 2.61 3.84
0.83 0.40 13.64 0.45 1.51
```

Les dades, ja estandarditzades, són les següents:



```
1.47 0.50 1.60 0.43 1.19
-0.72 1.25 -0.24 -0.92 -1.16
-0.96 -0.71 -0.97 -0.56 -0.75
1.26 0.95 1.16 -0.02 0.80
-1.28 0.18 -0.31 1.73 -0.42
0.34 -0.22 -0.83 -0.47 0.07
0.66 0.21 0.86 1.21 1.51
-0.78 -2.16 -1.27 -1.41 -1.25
```

2.- Utilitzeu el k-means nítid per a categoritzar les dades de l'esmentat arxiu en dues categories, ignorant les columnes no pertinents. Quin és el nivell de precisió[1] del resultat?

Apliquem el k-means nítid amb la distància euclidia i dues categories a totes les columnes excepte l'última, que denota la classe. Hem seleccionat els dos primers exemples com a centroides inicials. Calculem les distàncies euclidies entre els centroides i tots els exemples. Cada exemple pertany a la categoria del centroide més proper. Tornam a calcular els centroides fent la mitjana dels atributs, atribut a atribut de tots els exemples de cada categoria. Repetim aquest procés fins que no es produeixin canvis en les categories. Els centroides finals són:

```
1.13 0.55 1.20 0.54 1.17
-0.68 -0.33 -0.72 -0.32 -0.70
```

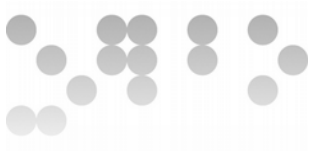
Les categories finals són:

```
Categoria 1=[1, 4, 7]
Categoria 2=[2, 3, 5, 6, 8]
```

Si assignem la classe 0 a la categoria 1 i la classe 1 a la categoria 2 tenim que:

```
precisió=(3+4)/8=87.5%
```

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



Com ja sabem, el resultat de k-means depèn molt dels centroides inicials. Si, per exemple, escollim els dos darrers exemples com a centroides inicials, obtenim els següents centroides finals:

```
0.53 0.46 0.83 0.84 0.77
-0.53 -0.46 -0.83 -0.84 -0.77
```

Les categories resultants en aquest cas són:

```
Categoria 1=[1, 4, 5, 7]
Categoria 2=[2, 3, 6, 8]
```

Si assignem la classe 0 a la categoria 1 i la classe 1 a la categoria 2 tenim que:

```
precisió=(3+3)/8=75%
```

3.- Apliqueu l'algorisme PCA per a 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. Comparau els resultats.

Si apliquem PCA sobre el conjunt de dades, obtenim els següents percentatges de variàncies:

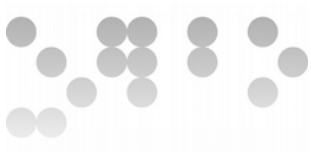
```
0.67 0.17 0.13 0.019 0.00
```

L'acumulat és:

```
0.67 0.84 0.97 0.99 1
```

Es pot veure que per tal de conservar el 95% de la variància és necessari emprar tres components. Sota aquesta condició el resultat seria:

```
-2.44981014  0.64202207  0.15467962
 0.88954245 -0.21020685 -1.86543676
 1.7905489  -0.09321994  0.10981319
-1.94186237  0.66023539 -0.51351541
```



```
 0.30405939 -2.16560609  0.32364174
 0.4776658   0.54991618  0.12487448
-2.01421005 -0.36071883  0.85073568
 2.94406601  0.97757809  0.81520744
```

Si apliquem k-means amb el conjunt reduït de dades emprant els dos primers exemples com a centroides inicials obtenim els següents centroides finals:

```
-2.14 0.31 0.16
1.28 -0.19 -0.10
```

Les categories són:

```
Categoria 1=[1, 4, 7]
Categoria 2=[2, 3, 5, 6, 8]
```

La precisió és:

```
precisió=(3+4)/8=87.5%
```

Per tant, hem aconseguit la mateixa precisió que abans.

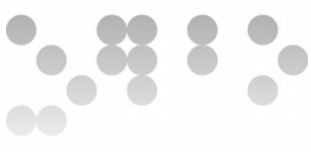
Si empram els dos darrers exemples com a centroides inicials, obtenim aquests centroides finals:

```
-1.53 -0.31 0.20
1.53 0.31 -0.20
```

Les categories són:

```
Categoria 1=[1, 4, 5, 7]
Categoria 2=[2, 3, 6, 8]
```

I la precisió:



precisió=(3+3)/8=75%

Com en el cas anterior, hem aconseguit la mateixa precisió que amb les dades originals.

Exercici 2

En aquest exercici haureu de categoritzar les dades amb el mètode aglomeratiu. En particular, heu de construir tres dendrogrames per al mateix conjunt de dades que a l'exercici anterior (small.csv) utilitzant la distància euclidia. Com a criteris d'aglomeració emprau el lligam simple per al primer dendrograma, el lligam complet per al segon i la mitja per al tercer. Calculeu les precisions de la mateixa forma que a l'exercici anterior.

En primer lloc s'ha de calcular la matriu de distàncies. Es pot convertir a una de semblances o treballar directament amb distàncies, com farem nosaltres. Aquesta matriu s'obtindrà calculant la distància euclídea entre cada parell d'exemples.

	1	2	3	4	5	6	7
2	4.01						
3	4.33	2.18					
4	0.89	3.26	3.87				
5	3.94	3.00	2.58	3.71			
6	3.11	2.31	1.63	2.62	2.85		
7	1.42	3.99	3.89	1.74	3.02	2.83	
8	5.44	3.59	1.78	5.08	4.14	2.80	5.17

A partir d'aquest punt hem d'efectuar tres càlculs distints en funció del mètode d'aglomeració. En el cas de lligam simple obtenim les següents agrupacions:

{1,4}

{1,4,7}

{3,6}

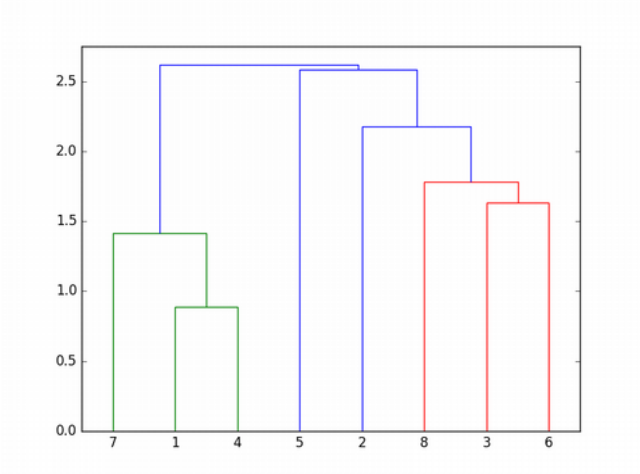
{3,6,8}

{2,3,6,8}

{2,3,5,6,8}

{1,2,3,4,5,6,7,8}

El dendrograma és el següent:



Si consideram dues categories queden els grups:

`Categoria 1=[1,4,7]`

`Categoria 2=[2,3,5,6,8]`

Això es correspon a una precisió de:

`precisió=(3+4)/8=87.5%`

Per l'enllaç complet obtenim les agrupacions:

`{1,4}`

`{3,6}`

`{1,4,7}`

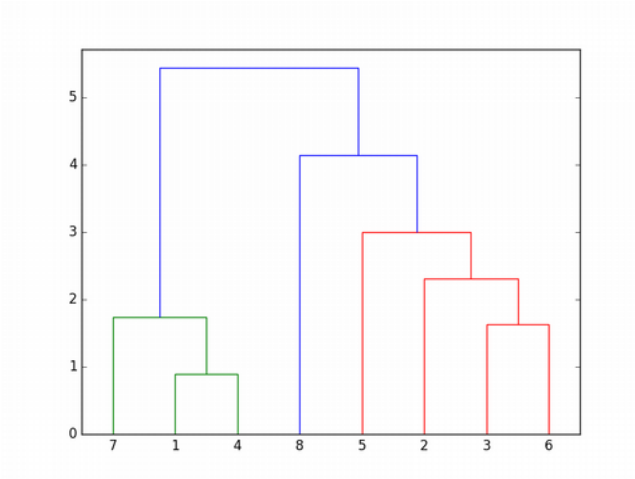
`{2,3,6}`

`{2,3,5,6}`

`{2.3.5.6.8}`

`{1,2,3,4,5,6,7,8}`

I el dendrograma:

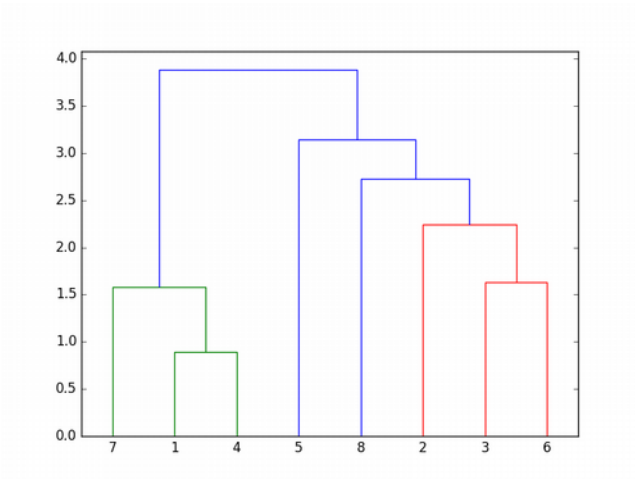


Si considerem dues categories queden els mateixos grups i, per tant, precisió que en el cas de lligam simple.

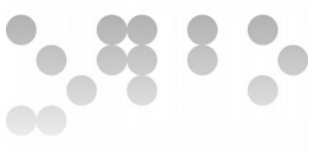
Per l'enllaç mitjà obtenim les agrupacions:

```
{1,4}
{1,4,7}
{3,6}
{2,3,6}
{2,3,6,8}
{2,3,5,6,8}
{1,2,3,4,5,6,7,8}
```

El dendrograma en aquest cas és:



Si consideram dues categories obtenim els mateixos grups (i, per tant, precisió) que en els casos anteriors.



Exercici 3

En aquest exercici haureu d'emprar unes biblioteques en Python per tal de categoritzar l'arxiu gran adjunt (wine.csv) en dues categories. La biblioteca s'anomena scikit-learn i la podeu descarregar i llegir-ne la documentació a la seva plana Web:

http://scikit-learn.org

En particular, se us demana que apliqueu, sobre les dades de l'arxiu esmentat, el mètode k-means (anomenat KMeans) amb dos criteris distints per a triar els centroides inicials i un altre mètode a escollir. Mostrau els resultats més destacables de cada categorització i calculau la precisió en cada cas. No heu d'adjuntar el codi, però sí heu de mostrar per a cada un dels tres casos (k-means amb el primer criteri, k-means amb el segon i mètode a escollir) la línia o línies que defineixen i parametritzen el categoritzador.

Hem realitzat una sèrie de proves amb KMeans amb tres inicialitzacions distintes: els dos primers exemples com a centroides inicials, el mètode k-means++ i el mètode random. També hem provat el mètode SpectralClustering.

La línia de codi corresponent a emprar els dos primers exemples com a centeroides és:

```
kMeansOut=KMeans(n_clusters=2,init=sampleData[0:2,:]).fit(sampleData)
```

La corresponent a la inicialitzcaió k-means++ és:

```
kMeansOut=KMeans(n_clusters=2,init='k-means++').fit(sampleData)
```

La corresponent a la inicialització random és:

```
kMeansOut=KMeans(n_clusters=2,init='random').fit(sampleData)
```

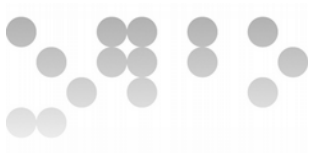
La corresponent a SpectralClustering és:

```
spClustOut=SpectralClustering(n_clusters=2).fit(sampleData)
```

Ja que, a excepció del primer cas, els mètodes retornaven resultats distints en cada execució, mostrarem els millors resultats obtinguts en cada cas:

	Precisió	Exemples erronis	Matriu confusió
K-means	93.85%	8	[[59,0],[8,63]]
K-means++	93.85%	8	[[59,0],[8,63]]
K-means-random	93.85%	8	[[59,0],[8,63]]
SpectralClustering	70.76%	38	[[28,31],[7,64]]

S'ha de dir que tant K-means++ i K-means random, produeixen resultats distints en execucions successives. En ambdós casos, tanmateix, el 93.85% de precisió és el



resultat majoritari. Pel que fa a SpectralClustering, el resultat majoritari era el 29% de precisió.

Exercici 4

Realitzau una valoració global comparant els mètodes dels exercicis anteriors i els resultats que n'heu obtingut. Redacteu unes conclusions globals. Els criteris de correcció de la PAC invaliden una A si tots els processos no estan ben justificats.

En aquest apartat s'espera que extraieu 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 pràctica 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?