PAC 1

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. En posteriors apartats, quan es parli de "small.csv" es sobreentendrà que treballareu amb el resultat del tractament de dades que faceu aquí.

L'arxiu small.csv no conté valors absents i, per tant, no s'han de tractar. Els atributs PRICE, MAINT, DOORS i PERSONS són numèrics amb valors molt dispersos i distints entre ells. Per tal de normalitzar les dades numèriques aplicarem ranging, tot i que també es podria aplicar estandardització.

L'atribut SAFETY és ordinal amb possibles valors low, med i high. Per a representar-lo adequadament, emprarem codificació additiva de la següent forma: low=00,med=10,high=11

La darrera columna és la classe i no s'ha de tractar. Ara bé, per a facilitar la notació, substituirem unacc per 0 i acc per 1.

A continuació es mostren les dades d'entrada tractades segons s'ha indicat.

PRICE	MAINT	DOORS	PERSONS	SAFETY1	SAFETY0	CLASS
1,00	1,00	0,00	0,00	0	0	0
0,67	0,00	0,33	1,00	1	1	1
0,00	1,00	1,00	0,67	1	1	1
0,33	0,33	0,00	0,67	1	0	1
1,00	1,00	1,00	0,00	1	1	0
0,67	0,33	0,67	0,00	1	1	0
0,67	1,00	0,33	0,67	1	1	0
0,33	0,67	0,00	0,00	1	0	0
0,00	0,00	0,33	1,00	1	0	1
1,00	0,33	0,00	0,67	1	1	1

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?

Aplicam k-means nítid amb dues categories i distància euclidiana a totes les columnes de les dades tractades a l'apartat anterior excepte la última columna, que és la classe. Seleccionam els dos primers exemples com a centroides inicials, tot i que algun altre criteri amb dos centroides de classes distintes també seria vàlid.

Calculem les distàncies euclidianes entre els centroides i tots els exemples, assignant cada exemple al centroide més proper. Així, obtenim que els exemples 1 i 8 s'assignen al primer centroide i els exemples 2,3,4,5,6,7,9 i 10 al segon.

Tornam a calcular els centroides fent la mitjana dels atributs. Els nous centroides ara són:

0.67 0.83 0.00 0.00 0.50 0.00

0.54 0.50 0.46 0.58 1.00 0.75

Ja que els centroides han canviat, continuam iterant. De nou, calculem les distàncies euclidianes entre tots els exemples i aquests nous centroides i assignem cada exemple al centroide més proper. L'assignació no canvia respecte de la iteració anterior. Això significa que, si calculéssim els centroides tampoc hi hauria canvis. Per tant, l'algorisme ha convergit. Les categories finals són:

Categoria 1 = [1,8]

Categoria 2 = [2,3,4,5,6,7,9,10]

Si suposem que classe 0 és la categoria 1 i la classe 1 és la categoria 2 aleshores s'ha encertat en els exemples 1,2,3,4,8,9 i 10. És a dir, s'ha encertat en 7 exemples de 10. Per tant, la precisió és de 0.7 o del 70%.

És fàcil veure que si es fes l'assignació contrària (classe 0 a categoria 2 i classe 1 a categoria 1) tindríem una precisió del 30%. Ja que

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

l'assignació classe-categoria és arbitrària, considerarem que la precisió és del 70%.

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 les següents variàncies acumulades:

0.37, 0.68, 0.85, 0.93, 0.99, 1.00

Per tant, per a preservar el 95% de la variància necessitem cinc components. Sota aquesta condició, el resultat seria:

-3.34 -1.81 -0.05 0.84 0.49
1.51 0.10 1.41 0.43 0.46
0.75 1.76 -1.70 0.89 0.00
0.89 -1.56 -0.27 -0.26 -0.33
-1.40 2.21 -0.08 -0.49 0.19
-0.23 1.10 0.28 -1.06 0.75
-0.15 0.86 0.20 0.68 -1.02
-0.43 -1.19 -0.94 -1.19 -0.66
2.21 -1.41 -0.71 0.17 0.60
0.20 -0.05 1.85 -0.00 -0.48

Si apliquem k-means nítid amb les mateixes condicions que s'ha fet abans però emprant ara la sortida del PCA, tenim que els centroides finals són:

-3.34 -1.81 -0.05 0.84 0.49
0.37 0.20 0.01 -0.09 -0.05

Les categories resultants són:

Categoria 1 = [1]
Categoria 2 = [2,3,4,5,6,7,8,9,10]

En aquest cas la s'encerten 6 de 10 exemples i, per tant, la precisió és de 0.6 o del 60%.

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 calcularem la matriu de distàncies. Després podríem convertir-la en una matriu de semblances o treballar directament amb les distàncies com farem nosaltres. Aquesta matriu, que es mostra a continuació, s'obtindrà calculant la distància euclidiana entre cada parella d'exemples.

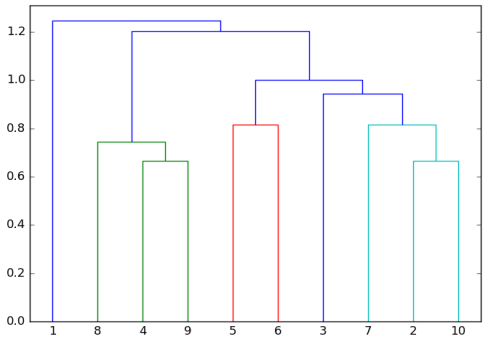
	1	2	3	4	5	6	7	8	9	10
1	0.00	2.05	2.11	1.53	1.73	1.73	1.63	1.25	2.03	1.70
2	2.05	0.00	1.41	1.20	1.60	1.11	1.05	1.63	1.20	0.67
3	2.11	1.41	0.00	1.60	1.20	1.20	0.94	1.63	1.60	1.56
4	1.53	1.20	1.60	0.00	1.83	1.41	1.29	0.75	0.67	1.20
5	1.73	1.60	1.20	1.83	0.00	0.82	1.00	1.60	2.11	1.37
6	1.73	1.11	1.20	1.41	0.82	0.00	1.00	1.29	1.63	1.00
7	1.63	1.05	0.94	1.29	1.00	1.00	0.00	1.33	1.60	0.82
8	1.25	1.63	1.63	0.75	1.60	1.29	1.33	0.00	1.29	1.41
9	2.03	1.20	1.60	0.67	2.11	1.63	1.60	1.29	0.00	1.53
10	1.70	0.67	1.56	1.20	1.37	1.00	0.82	1.41	1.53	0.00

Com es pot veure, la matriu de distàncies és simètrica (ja que la distància de A a B és la mateixa que de B a A) i té zeros a la diagonal (ja que la distància d'un exemple a ell mateix és zero).

Comencem amb el lligam simple. Es pot veure que la distància mínima és de 0.67 i apareix entre 4 i 9 i també entre 2 i 10. Seleccionem per exemple 4 i 9. Ara s'haurien d'eliminar les files de 4 i de 9 i afegir una nova fila que representés el grup 4,9. La distància d'aquest nou grup a un exemple x seria la mínima entre 4,x i 9,x. Posteriorment tornaríem a

cercar el mínim i es repetiria el procés fins haver format tots els grups.

El dendrograma resultant és el següent:



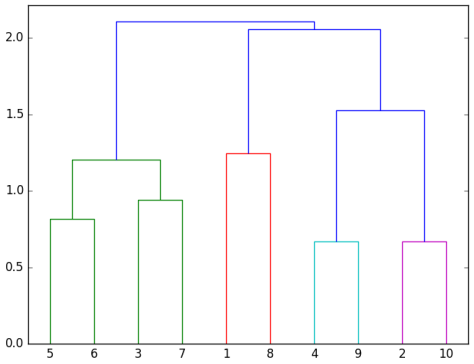
Si considerem dues categories, ens queda:

Categoria 1 = [1]

Categoria 2 = [2,3,4,5,6,7,8,9,10]

Això suposa una precisió de 0.6 o del 60%

Pel que fa al lligam complet, l'algorisme és el mateix però ara, per tal de calcular les distàncies de l'agregació no seleccionem la distància mínima sinó la màxima. En aquest cas el dendrograma resultant és:



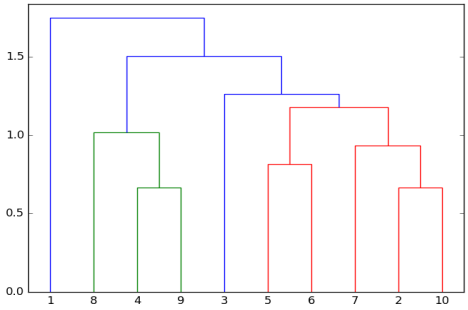
Si considerem dues categories, tendríem:

Categoria 1 = [3,5,6,7]

Categoria 2 = [1,2,4,8,9,10]

Això representa una precisió de 0.7 o del 70%

Respecte a la mitja com a criteri de lligam, el que es faria és considerar que la distància de cada nou grup a la resta d'exemples o grups és la mitja aritmètica de les distàncies de cada element del grup a l'exemple corresponent. Segons aquest criteri, el dendrograma resultant seria:



En aquest cas, si considerem dues categories tenim:

Categoria 1 = [1]

Categoria 2 = [2,3,4,5,6,7,8,9,10]

Això suposa, com en el primer cas, una precisió de 0.6 o del 60%.

Exercici 3

En aquest exercici treballareu amb l'arxiu de dades gran (car.csv). El primer que heu de fer és aplicar a aquestes dades el mateix tractament previ que heu aplicat al primer apartat de l'Exercici 1. Després haureu d'emprar una biblioteca de Python per tal de categoritzar les dades de car.csv tractades tal i com s'ha indicat.

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 esmentades, el mètode k-means (anomenat KMeans) amb dos criteris distints per a triar els centroides inicials i un altre mètode a escollir, distint a KMeans. Per a cada categorització heu de mostrar, com a mínim, precisió (accuracy), nombre d'exemples categoritzats erròniament i matriu de comfusió.

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 fet una sèrie de proves amb distints criteris per a seleccionar els centroides inicials. En primer lloc, hem emprat els dos primers exemples com a centroides:

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

En segon lloc, hem emprat la tècnica k-means++ per a la selecció de centroides inicials:

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

També hem provat una selecció aleatòria de centroides inicials:

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

Pel que fa al mètode distint a KMeans hem seleccionat SpectralClustering:

```
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 de les diverses execucions realitzades en cada cas. Les matrius de confusió es mostren en format [[TN,FP],[FN,TP]].

	Precisió	Exemples erronis	Matriu confusió
K-means	63.31%	634	[[576,634], [0,518]]
K-means++	71.3%	496	[[933,277], [219 299]]
K-means-random	71.3%	496	[[933,277],[219 299]]
SpectralClustering	71.3%	496	[[933,277],[219 299]]

Cal destacar que els millors casos trobats per K-means++, random i Spectral Clustering coincideixen en la matriu de confusió (i per tant en precisió i nombre d'exemples erronis). Tot i que el component aleatori d'aquests mètodes podria ser el causant d'aquests resultats idèntics, cal dir que és significatiu que amb les execucions fetes el millor cas detectat coincideixi. És molt probable que aquests resultats es deguin a que no és possible trobar una millor classificació amb els atributs disponibles.

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?