PAC 2

Presentació

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

Objectius

L'objectiu d'aquesta prova d'avaluació és classificar les dades dels arxius adjunts relacionats amb la distribució d'aliments. Volem predir o el tipus de venda (al detall o a l'engròs) o la ciutat del client (Lisboa, Oporto o altres).

Els arxius de dades "csv" tenen un format tipus taula, on cada fila correspon a un exemple. La primera i/o segona columna són la classe 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 Custormers" del repositori d'aprenentatge de l'UCI:

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

Solució de la PAC

Exercici 1

a) Construïu els models de classificació basats en el veí més proper per valors de k u i tres a partir de l'arxiu train.csv. És a dir, heu de construir els models: 1NN i 3NN. Un cop tingueu el model, heu de classificar els exemples de l'arxiu test.csv.

Ens proporcionen un arxiu amb dades de clients d'una distribuïdora 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. Els valors dels promitjos i desviacions són respectivament:

```
4010.5 4912.8 8320.8 553.3 3834.5 1893.3
3917.5 4013.4 7681.7 365.5 4287.2 960.8
```

En estandarditzar el conjunt de test hem d'utilitzar els promitjos i desviacions del conjunt d'entrenament.

Per aplicar el kNN, calculem les distàncies euclidees de tots els exemples de test a tots els exemples de train:

```
          [,1]      [,2]
[1,] 6.000373 2.6626176
[2,] 7.710386 4.2653440
[3,] 3.824833 5.1148720
[4,] 4.849979 3.6767085
[5,] 5.625313 2.8606337
[6,] 6.686189 0.7271603
```

Per a 1NN, assignarem la classe 2 al primer exemple i la 1 al segon, ja que els més petits (marcats en vermell) són d'aquestes classes, que correspon a una precisió del 100% ($= \frac{1+1}{2}$).

Per al 3NN els vots serien (2, 2, 2) i (1, 1, 1) que corresponen a les prediccions 2 i 1 respectivament; que equival a una precisió del 100% ($= \frac{1+1}{2}$).

b) Construïu els models de classificació basats en el k–means per a un valor de k de dos a partir de l'arxiu train.csv. Un cop tingueu el model, heu de classificar els exemples de l'arxiu test.csv.

Apliquem en primer lloc el 2-means dues vegades: una per als 3 exemples de cadascuna de les classes. Agafant com a centroides inicials els dos primers exemples en cada cas ens queden els quatre centroides:

```
   Fresh      Milk         Grocery    Frozen     Detergents Delicassen
1  1.0052397 -0.004443497 -0.9106441  1.0455502 -0.8588404  0.0381641
1 -0.8649707 -1.210416999 -1.0653616 -1.3086476 -0.8927789 -1.9623285
2 -0.5160214  1.439485519  1.4500834 -0.8763835  1.8059674  0.2702712
2 -0.3147437 -0.110090763  0.7182832  0.0469654  0.4022462  0.8078646
```

Apliquem 1NN agafant com a training els quatre centroides obtinguts en el pas anterior i com a test el conjunt de test. El tractarem com en l'exercici anterior. Obtenim les distàncies següents:

```
          [,1]     [,2]
[1,] 6.434933 1.379182
[2,] 8.450218 4.702481
[3,] 4.187207 5.166582
[4,] 5.228734 3.098284
```

Que corresponen a les prediccions 2 i 1 respectivament. Això equival a un 100% de precisió.

c) Construïu un arbre de decisió a partir de l'arxiu train.csv i classifiqueu amb ell els exemples de l'arxiu test.csv.

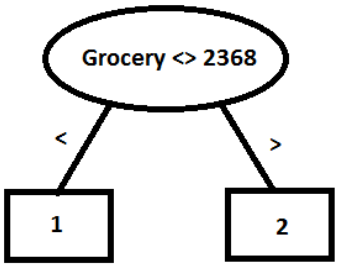
Per a construir arbres de decisió no cal normalitzar, donat que no treballarem amb distancies.

Tots els atributs són numèrics. Utilitzarem els punts de tall per tractar-los. Els punts de tall es calculen ordenant tots els valors d'un atribut i calculant la mitjana

aritmètica de cada dos valors consecutius. A continuació es mostra els valors ordenats sense repeticions, els punts de tall i la seva bondat per cadascun dels atributs:

```
Fresh
622    1173,5 67%    =(1+3)/6
1725   1857   50%    =(1+2)/6
1989   2726   67%    =(2+2)/6
3463   3646,5 50%    =(2+1)/6
3830   8132   67%    =(3+1)/6
12434
Milk
55     297,5  67%    =(1+3)/6
540    2095,5 83%    =(2+3)/6
3651   4471   67%    =(2+2)/6
5291   7270,5 50%    =(2+1)/6
9250   9970   67%    =(3+1)/6
10690
Grocery
137    210    67%    =(1+3)/6
283    1325,5 83%    =(2+3)/6
2368   7595   100%   =(3+3)/6
12822  13838,5       83%    =(3+2)/6
14855  17157,5       67%    =(3+1)/6
19460
Frozen
75     154    67%    =(1+3)/6
233    275    50%    =(1+2)/6
317    548    67%    =(2+2)/6
779    801,5  50%    =(2+1)/6
824    958    67%    =(3+1)/6
1092
Detergents_Paper
3      5      67%    =(1+3)/6
7      154,5  83%    =(2+3)/6
302    2363   100%   =(3+3)/6
4424   5559   83%    =(3+2)/6
6694   9135,5 67%    =(3+1)/6
11577
Delicassen
8      817,5  67%    =(1+3)/6
1627   1890   83%    =(2+3)/6
2153   2155   67%    =(2+2)/6
2157   2195   50%    =(2+1)/6
2233   2707,5 67%    =(3+1)/6
3182
```

El millor que obtenim són els punts de tall 2368 de l'atribut 'Grocery' i el 302 de 'Detergents_Paper', que aconsegueixen un 100% de bondat. Podem escollir qualsevol dels dos. Nosaltres escollim el Grocery. Tenim particionat correctament el conjunt d'exemples. Així, hem enllestit la construcció de l'arbre, que quedaria:



La prediccions són respectivament '2' i '1'. Per tant, obtindrem una predicció del 100%.

d) Apliqueu l'algorisme del PCA per reduir la dimensionalitat dels conjunts anteriors conservant el 95% de la variància. Utilitzeu el 1-NN i 3-NN sobre els conjunts reduïts de la mateixa forma que en el primer apartat. S'obtenen resultats comparables? Expliciteu les diferències en aplicar el PCA sobre el conjunt d'entrenament i sobre el de test.

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
```

I projectant el test:

```
-3.30 2.16  0.30 2.58
 2.00 1.48 -0.09 0.16
```

Un cop obtingudes les dades projectades, fem el mateix que al primer apartat, calcular la matriu de distàncies:

```
          [,1]      [,2]
[1,] 5.286345 2.6193420
[2,] 7.266409 4.2614830
[3,] 3.098355 5.1069024
[4,] 3.681336 3.5790035
[5,] 5.093048 2.8581007
[6,] 6.236847 0.7196443
```

Per a 1NN, assignarem les classes '2' i '1' respectivament, ja que els més petits (marcats en vermell) són d'aquestes classes, que correspon a una precisió del 100% ($=\frac{1+1}{2}$).

Per al 3NN els vots serien ('2', '2', '2') i ('1', '2', '1') que corresponen a les prediccions '2' i '1' respectivament; que equival a una precisió del 100% ($= \frac{1+1}{2}$).

Exercici 2

L'objectiu d'aquest segon exercici és la construcció d'un model amb un conjunt de dades proper al que utilitzaríem en una aplicació pràctica per a un cas real. Per realitzar aquest exercici, teniu a la vostra disposició una eina anomenada Weka a la Web; la seva direcció és:

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

L'arxiu adjunt "Wholesale customers.csv" conté 440 exemples de clients d'una distribuïdora d'aliments amb dos columnes de classe que corresponen a dos problemes. Aquest arxiu està en un dels formats d'entrada del Weka.

Se us demana l'estudi de cara a construir dos classificadors, un per a cada problema. Per a realitzar-los heu de:

- Provar els algorismes per validació creuada (cross-validation). És a dir, no utilitzar un sol arxiu de train i un de test. Aquesta opció la permet realitzar el Weka de forma automàtica.
- Provar almenys els algorismes: naïve bayes, algun tipus d'arbre de decisió (mostrant a l'informe l'arbre generat), AdaBoost, xarxes neuronals (perceptró multicapa) i algun altre algorisme.
- Provar les Support Vector Machines (SMO i/o libSVM al Weka) amb diferents tipus de kernel.

Comenteu els resultats obtinguts i justifiqueu tot el que feu.

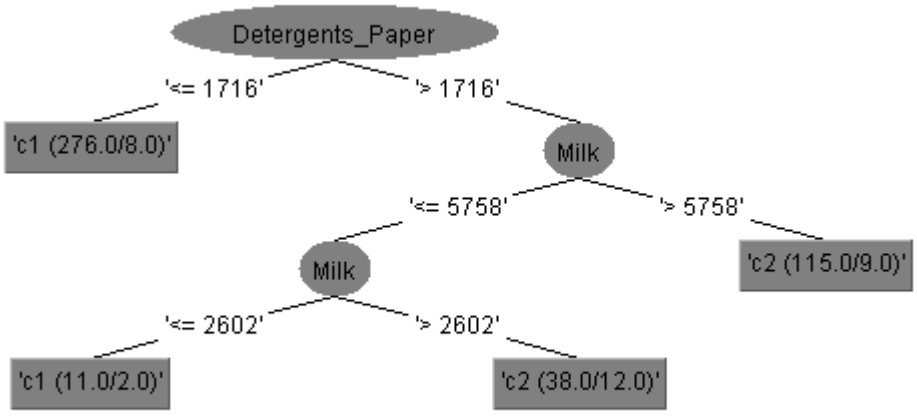
Hem escollit realitzar les proves que demana l'enunciat amb un 10-fold crossvalidation sobre els algorismes: Naïve Bayes, 1NN (IB1), 3NN (IB3), Decision Stumps, l'arbre de decisió J48, l'AdaBoost.M1 i el perceptró multicapa. No hem realitzat cap tipus de tractament previ als atributs.

La taula següent mostra els resultats per a la classe 'channel'. Concretament mostra el temps de construcció dels models, la precisió global, una mesura per cadascuna de les classes i les dades de l'estadístic Kappa (mesura la proporció de la precisió entre les diferents classes):

	NB	DecisionStump	1NN	3NN	J48	AdaBoost.M1	Perceptron
Temps (s)	0,03	0,03	0,01	0	0,16	0,08	1,46
Oks	392	402	388	397	396	405	405
Precision	89%	91%	88%	90%	90%	92%	92%
F-mesure 2	0,827	0,876	0,818	0,852	0,851	0,884	0,878

F-mesure 1	0,92	0,934	0,912	0,927	0,925	0,939	0,941
Kappa stat.	0,7477	0,8101	0,7306	0,7793	0,7762	0,8242	0,819

A continuació es mostra l'arbre que s'ha generat amb el mètode d'arbres de decisió J48:



i a continuació les matrius de confusió:

Naïve Bayes	115 27 | a = c2 21 277 | b = c1
Decision Stump	134 8 | a = c2 30 268 | b = c1
1NN	117 25 | a = c2 27 271 | b = c1
3NN	124 18 | a = c2 25 273 | b = c1
J48	126 16 | a = c2 28 270 | b = c1
AdaBoost.M1	134 8 | a = c2 27 271 | b = c1
Perceptron	126 16 | a = c2 19 279 | b = c1

A continuació tenim les dades de les Support Vector Machines:

	SMO			
	lineal	quadratic	rbf ($\gamma = 0.01$)	rbf ($\gamma = 1$)
Temps (s)	0,15	0,25	0,94	0,11
Oks	386	328	298	392
Precision	88%	75%	68%	89%
F-mesure 2	0,779	0,349	0	0,812
F-mesure 1	0,915	0,842	0,808	0,923
Kappa stat.	0,6969	0,2662	0	0,7369

SMO	lineal	95 47 | a = c2 7 291 | b = c1
	quadratic	30 112 | a = c2 0 298 | b = c1
	rbf ($\gamma = 0.01$)	0 142 | a = c2 0 298 | b = c1
	rbf ($\gamma = 1$)	104 38 | a = c2 10 288 | b = c1

Ara hem de repetir tot el procés amb la classe 'region'.

	NB	DecisionStump	1NN	3NN	J48	AdaBoost.M1	Perceptron
Temps (s)	0	0	0	0	0	0,01	1,3
Oks	209	316	236	292	316	316	314
Precision	48%	72%	54%	66%	72%	72%	71%
F-mesure 3	0,628	0,836	0,691	0,794	0,836	0,836	0,833
F-mesure 2	0,113	0	0,085	0,068	0	0	0
F-mesure 1	0,229	0	0,199	0,234	0	0	0
Kappa stat.	-0,0047	0	-0,0284	0,0694	0	0	-0,0062

A continuació es mostra l'arbre que s'ha generat amb el mètode d'arbres de decisió J48:

i a continuació les matrius de confusió:

Naïve Bayes	177 123 16 | a = c3 45 28 4 | b = c1

	26	17	4	|	c = c2
Decision Stump	316	0	0	|	a = c3
	77	0	0	|	b = c1
	47	0	0	|	c = c2
1NN	216	62	38	|	a = c3
	56	16	5	|	b = c1
	37	6	4	|	c = c2
3NN	275	32	9	|	a = c3
	61	15	1	|	b = c1
	41	4	2	|	c = c2
J48	316	0	0	|	a = c3
	77	0	0	|	b = c1
	47	0	0	|	c = c2
AdaBoost.M1	316	0	0	|	a = c3
	77	0	0	|	b = c1
	47	0	0	|	c = c2
Perceptron	314	0	2	|	a = c3
	77	0	0	|	b = c1
	47	0	0	|	c = c2

A continuació tenim les dades de les Support Vector Machines:

	SMO			
	lineal	quadratic	rbf ($\gamma = 0.01$)	rbf ($\gamma = 1$)
Temps (s)	0,07	0,21	0,11	0,27
Oks	316	316	316	316
Precision	72%	72%	72%	72%
F-mesure 3	0,836	0,836	0,836	0,836
F-mesure 2	0	0	0	0
F-mesure 1	0	0	0	0
Kappa stat.	0	0	0	0

SMO	lineal	316	0	0	| a = c3
		77	0	0	| b = c1
		47	0	0	| c = c2
	quadratic	316	0	0	| a = c3
		77	0	0	| b = c1
		47	0	0	| c = c2
	rbf ($\gamma = 0.01$)	316	0	0	| a = c3
		77	0	0	| b = c1
		47	0	0	| c = c2
	rbf ($\gamma = 1$)	316	0	0	| a = c3
		77	0	0	| b = c1
		47	0	0	| c = c2

Exercici 3

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 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: 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? Tant en el procés de training com en el de test.

En comparar els resultats, és important notar que els conjunts de dades tenen diferent número de classes.

Exercici 4

Descriviu de forma clara i curta en què es diferencien el "Multinomial Naïve Bayes" i el "Bernoulli Naïve Bayes" del "Gaussian Naïve Bayes", posant èmfasi en l'aplicació pràctica.

Les tres variants es basen en l'asumpció de distribucions estadístiques diferents. Les diferències venen donades pel tipus d'atribut que volen tractar. En el cas del "Gaussian" es vol tractar valors numèrics, en el "Multinomial" freqüències d'aparició i en el "Bernoulli" variables booleanes o valors binaris.