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 fallida de projectes empresarials. Volem predir la fallida en funció de les seves propietats.

L'arxiu de dades Qualitative_Bankruptcy.data.txt té un format tipus taula, on cada fila correspon a un exemple. L'última columna és la classe i la resta corresponen als atributs de l'exemple. L'arxiu adjunt Qualitative_Bankruptcy.info.txt conté la descripció d'aquests atributs.

Aquests arxius pertanyen al problema "Qualitative Bankruptcy" 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.txt. 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.txt.

No hi ha valors absents, no els hem de tractar. Tots els atributs són ordinals, podem substituir el valors N, A i P per 0, 0.5 i 1 respectivament. D'aquesta forma, tots el valors queden normalitzats.

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

```
          [,1]      [,2]
[1,] 1.581139 1.4142136
[2,] 1.224745 1.4142136
[3,] 1.118034 1.1180340
[4,] 1.658312 0.8660254
[5,] 1.936492 0.5000000
[6,] 1.581139 1.2247449
```

Per a 1NN, assignarem les classes 'NB' al primer exemple de test i 'B' 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 ('NB', 'NB', 'NB') i ('NB', 'B', 'B') que corresponen a les prediccions 'NB' i 'B' 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.txt. Un cop tingueu el model, heu de classificar els exemples de l'arxiu test.txt.

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:

```
   Ind_Risk Man_Risk Fin_Flex Credib Competit Oper_Risk
NB     0.75     0.75      0.5    0.5      0.5      0.75
NB     0.00     0.00      0.5    0.5      0.5      0.00
B      0.50     0.00      0.0    0.0      0.0      0.25
B      0.50     0.50      0.0    0.0      0.0      1.00
```

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,] 1.299038 1.198958
[2,] 1.224745 1.414214
[3,] 1.600781 1.030776
```

```
[4,] 1.936492 0.500000
```

Que corresponen a les prediccions 'NB' i 'B' respectivament. Això equival a un 100% de precisió.

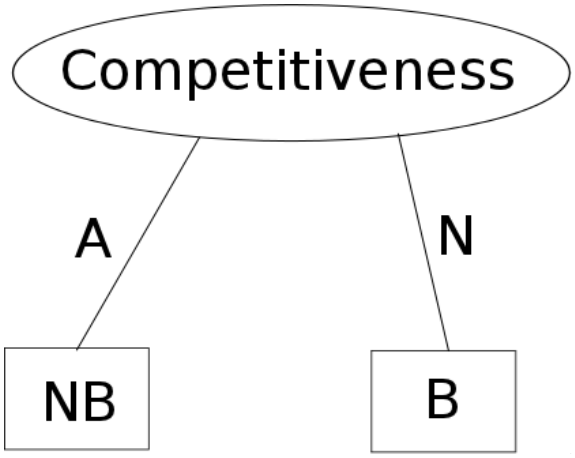
c) Construïu un arbre de decisió a partir de l'arxiu train.txt i classifiqueu amb ell els exemples de l'arxiu test.txt.

Per a construir arbres de decisió utilitzarem l'arxiu amb les dades originals, utilitzant els atributs nominals, ja que només tenint tres possibles valors no aporta gaire el ús dels punts de tall. Així, obtenim les bondats:

```
Industrial_Risk         83% = (3+1+1)/6
Management_Risk         67% = (1+2+1)/6
Financial_Flexibility 100% = (3+3)/6
Credibility            100% = (3+3)/6
Competitiveness        100% = (3+3)/6
Operating_Risk          50% = (1+1+1)/6
```

El millor que obtenim són els atributs: Financial_Flexibility, Credibility i Competitiveness, que aconsegueix un 100% de bondat. Podem escollir qualsevol dels tres. Per sota de l'atribut assignem NB a A i B a N.

Tots tres atributs particionen correctament el conjunt d'exemples. Així, hem enllestit la construcció de l'arbre, que quedaria:



Les prediccions són respectivament 'NB' i 'B'. 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.

Tenint en compte el tractament comentat en el primer apartat, podem tractar els atributs com a numèrics i, per tant, té sentit aplicar el PCA. Per realitzar aquest exercici, hem utilitzat el PCA des de l'R[1].

Amb dues components obtenim un 93,7% i amb tres un 99,3% de la variància. Per tant, hem de menester 3 components. El resultat és:

```
          Comp.1     Comp.2       Comp.3
[1,] -1.0413820 -0.1563740  0.151140072
[2,]  0.5885120 -0.6535221 -0.158495856
[3,] -0.2264350 -0.4049480 -0.003677892
[4,]  0.3304628  0.4164625 -0.013939541
[5,] -0.2886853  0.5577006 -0.257525970
[6,]  0.6375275  0.2406809  0.282499188
```

per al conjunt d'entrenament i per al de test[2]:

---

[1] Consulteu el document/tutorial que hi ha disponible a la carpeta exemples del fòrum de l'assignatura.

```
          Comp.1     Comp.2     Comp.3
[1,]  0.03370460 -0.9634281  0.6099192
[2,] -0.09288639  0.4503646 -0.6559304
```

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

```
          [,1]      [,2]
[1,] 1.4204315 1.3853297
[2,] 0.9971533 1.3893562
[3,] 0.8695251 1.0838949
[4,] 1.5431666 0.7697572
[5,] 1.7805136 0.4567103
[6,] 1.3862483 1.2075271
```

Per a 1NN, assignarem les classes 'NB' i 'B' 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 ('NB', 'NB', 'B') i ('NB', 'B', 'B') que corresponen a les prediccions 'NB' i 'B' respectivament; que equival a una precisió del 100% ($= \frac{1+1}{2}$).

[2] Per calcular el de test, s'ha de projectar utilitzant les dades del conjunt d'entrenament. En el cas de l'R, s'utilitza la comanda *predict*.

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 Qualitative_Bankruptcy.data.txt conté 175 exemples de projectes empresarials amb dues classes. Aquest arxiu està en un dels formats d'entrada del Weka.

Se us demana l'estudi de cara a construir un classificador. 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.
- Apliqueu el PCA per reduir la dimensionalitat conservant el 95% de la variància i torneu a aplicar els mateixos algorismes que al conjunt original.

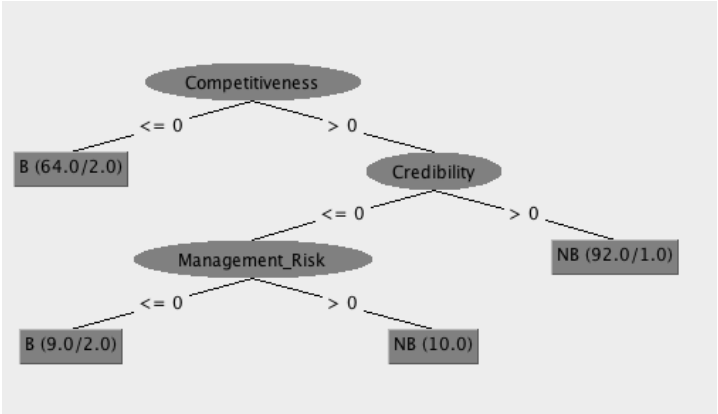
Comenteu els resultats obtinguts i justifiqueu tot el que feu.

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. Utilitzarem l'última columna com a classe.

La taula següent mostra els resultats aplicant als atributs els comentats a l'exercici anterior. 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)	0,02	0,01	0	0	0,03	0,09	1,5
Oks	171	165	165	171	167	169	167
Precision	97,7%	94,3%	94,3%	97,7%	95,4%	96,6%	95,4%
F-mesure NB	0,981	0,954	0,952	0,981	0,962	0,972	0,962
F-mesure B	0,971	0,925	0,930	0,971	0,943	0,957	0,943
Kappa stat.	0,9519	0,8792	0,8815	0,9522	0,9048	0,9282	0,9048

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	105 0 | a = NB 4 66 | b = B
Decision Stump	103 2 | a = NB 8 62 | b = B
1NN	99 6 | a = NB 4 66 | b = B
3NN	104 1 | a = NB 3 67 | b = B
J48	101 4 | a = NB 4 66 | b = B
AdaBoost.M1	103 2 | a = NB 4 66 | b = B
Perceptron	101 4 | a = NB 4 66 | b = B

A continuació tenim les dades de les Support Vector Machines:

	SMO			
	lineal	quadratic	rbf ($\gamma = 0.01$)	rbf ($\gamma = 1$)
Temps (s)	0,08	0,03	0,14	0,06
Oks	171	168	170	168
Precision	97,7%	96,0%	97,1%	96,0%
F-mesure NB	0,981	0,967	0,977	0,967
F-mesure B	0,971	0,950	0,964	0,950
Kappa stat.	0,9519	0,9165	0,94	0,9165

SMO	lineal	105 0 | a = NB 4 66 | b = B
	quadratic	102 3 | a = NB 4 66 | b = B
	rbf ($\gamma = 0.01$)	104 1 | a = NB 4 66 | b = B
	rbf ($\gamma = 1$)	102 3 | a = NB 4 66 | b = B

En aplicar el PCA ens quedem amb 6 components per obtenir el 95% de variància. No aplicarem el PCA.

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:

- 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

Doneu una definició intuïtiva i curta de les màquines de vectores de suport estructurades ("Structured Support Vector Machines") i descriviu un exemple de problema al que aplicar-les.

Les màquines de vectors de suport estructurades és una generalització de les màquines de vectors de suport en que les classes de sortida són una estructura.

Un exemple d'aplicació seria l'anàlisi sintàctica en el processament del llenguatge natural. Aquest algorisme podria tenir com a sortida l'arbre corresponent a l'anàlisi sintàctica d'una frase en català.