UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI INFORMATICA E AUTOMAZIONE

Test di Uscita - Text Mining

Data: *03/03/2008*

Claudio Biancalana
claudio.biancalana@dia.uniroma3.it
http://www.dia.uniroma3.it/~biancal/

1) Si consideri il seguente insieme di dati, dove *tipo* è da considerarsi l'attributo di classe.

Nome	**Numero Zampe**	**Vola**	**Tipo**
corvo	2	Y	uccello
tirannosauro	2	N	dinosauro
cane	4	N	mammifero
pegaso	4	Y	animale leggendario

a. Costruire un albero di classificazione basato sull'algoritmo ID3. Qual è l'errore di sostituzione che si ottiene?

b. Quale tasso di errore si ottiene col metodo del leave one out cross validation?

2) Supponiamo di eliminare l'attributo *nome* dall'insieme di dati di cui sopra, e di addestrare in questo modo un classificatore Bayesiano naive.

    a. Data una generica istanza "numero zampe = n, vola = v", qual è la probabilità che la classe predetta sia un mammifero?

    b. Poiché probabilità nulle sono spesso indesiderate, quale metodo può essere utilizzato per eliminare questo inconveniente?

3) Descrivere in maniera chiara e concisa cosa è una regola associativa (è sufficiente limitarsi al solo caso di regole associative booleane mono-dimensionali). Supporto e confidenza sono sempre buone misure oggettive per determinare quali sono le regole interessanti? (motivare la risposta).

4) Si consideri il seguente insieme di dati:

ID	X	Y
A	0	0
B	0	0.5
C	2	1
D	0.2	0.5
E	0	-0.8
F	2	2

Applicare un algoritmo di raggruppamento gerarchico agglomerativo e disegnare il dendogramma risultante. Il raggruppamento deve avvenire solo sulla base degli attributi x e y. Lo studente può scegliere a suo piacimento il tipo di misura di distanza da utilizzare.

5) Sia dato il seguente insieme di dati

Animale	Vola	Acquatico	Zampe	Mammifero
Colomba	S	N	S	N
Mosca	S	N	S	N
Delfino	N	S	N	S
Squalo	N	S	N	N
Lucertola	N	N	S	N
Cane	N	N	S	S
Pipistrello	S	N	S	S

Si supponga di addestrare con questi dati un classificatore Bayesiano naive, considerando l'attributo mammifero come attributo di classe e ignorando l'attributo animale. Quale classe verrebbe predetta in presenza della seguente istanza?

Animale	Vola	Acquatico	Zampe
Negumi	S	S	S

6) Si consideri il seguente insieme di dati:

ID	X	Y
A	0	0
B	0	0.5
C	2	1
D	0.2	0.5
E	0	-0.8
F	2	2

Applicare l'algoritmo k-means con k=2, scegliendo come centri iniziali i punti A e B. Il raggruppamento deve avvenire solo sulla base degli attributi x e y. Lo studente può scegliere a suo piacimento il tipo di distanza da utilizzare.

7) Sia dato il seguente insieme di dati, nel quale l'attributo *stipendio* è da considerare l'attributo di classe:

Nome	**Età**	**Anni servizio**	**Dipartimento**	**Stipendio**
Gianluca	30	10	Ricerca	basso
Carla	50	20	Ricerca	alto
Lucia	50	25	Vendite	alto
Michela	40	10	Vendite	alto

Dopo aver normalizzato i dati di tipo numerico, calcolare il tasso di errore utilizzando il metodo di leave one out cross validation e l'algoritmo 1-nearest neighbour

8) Sia dato il seguente insieme di dati:

ID	X	Y
A	0	0
B	0	1
C	2	1
D	2	2
E	2	3

Applicare l'algoritmo k-mediods con k=2, considerando i punti C ed E come medioidi iniziali.

9) Descrivere gli algoritmi di bagging e boosting usati per la combinazione di classificatori.

10) Dati i seguenti documenti:

D1 = 3T1 + 5T2 + 3T3
D2 = 1T1 + 7T2 + 3T3

Rappresentati in uno spazio vettoriale per IR basato su vocabolario ridotto a 3 termini, e data la query

Q = 1T1 + 1T2 + 0T3

Quale tra D1 e D2 è più simile a Q, secondo la misura di somiglianza del coseno?
Giustificare la risposta attraverso la descrizione dei calcoli effettuati.

11) Avendo questi tre documenti:

Doc1.
Questo documento tratta della segmentazione temporale. In questo documento viene inoltre trattato lo spazio delle versioni.

Doc2.
La rappresentazione vettoriale della conoscenza presume una segmentazione spaziale

Doc3.
Lo spazio vettoriale ha delle conseguenze sullo spazio delle versioni?

Rappresentare attraverso l'indicatore TFxIDF lo spazio vettoriale.

12) A partire dalla descrizione degli algoritmi di classificazione Rocchio e k-NN, descriverne le differenze. (Aiutarsi con i grafici sul piano cartesiano).

13) Descrivere in linguaggio naturale, la metodologia dell'indicizzazione semantica latente (LSI).