Corso di Laurea Specialistica in Ingegneria Informatica
Data Mining e Scoperta di Conoscenza
Esame del 10 gennaio 2006

Si consideri il seguente dataset

	x	*y*	*U*
1	0	1	-1
2	1	4	-1
3	10	0	1
4	0	6	-1
5	0	2	-1
6	3	10	1
7	6	6	1
8	10	10	1
9	1	5	-1
10	8	9	1

Esercizio 1 (4 punti).
Utilizzando il lagrangiano descritto dal vettore

$$[0\ 0\ 0.023802\ 0\ 0\ 0.074711\ 0\ 0\ 0.098512\ 0]^T$$

(a) Si idenfichino graficamente il decision boundary e i suoi margini;
(b) Si descriva analiticamente l'equazione corrispondente.

Esercizio 2 (2 punti).
Nell'ipotesi in cui si voglia costruire un albero di decisione C4.5, si determini la radice dell'albero e lo split corrispondente.

Esercizio 3 (2 punti).
Si applichi l'algoritmo DBScan (utilizzando MinPts=3 e fissando opportunamente il valore di ε) all'intero dataset, utilizzando la distanza del coseno. Cosa cambia se invece si utilizza la distanza euclidea?

Esercizio 4 (4 punti).
Assumendo *x* continua e *y* e *U* discrete,
(a) discretizzare il dataset utilizzando l'algoritmo ChiMerge con α=95% (χ^2=3.8414).
(b) Calcolare le regole associative multidimensionali utilizzando supporto 30% e confidenza 100%.

SOLUZIONE

Esercizio 1
Il lagrangiano primale del problema è dato da

$$L_p(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - 1 \right]$$

Dove $\mathbf{w}$ e b caratterizzano l'iperpiano di separazione, e $\boldsymbol{\alpha}$ rappresenta il lagrangiano. Le condizioni di ottimalità sono date dai valori della funzione che soddisfano:

$$\frac{\partial L_p}{\partial w_j} = 0, j = 1...m$$
$$\frac{\partial L_p}{\partial b} = 0$$
$$\frac{\partial L_p}{\partial \alpha_i} \le 0, i = 1...N$$
$$y_i \ge 0, i = 1...N$$
$$\alpha_i \left[y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - 1 \right] = 0, i = 1...N$$

Semplificando, le condizioni possono essere riscritte in

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$
$$\sum_{i=1}^{N} \alpha_i y_i = 0$$
$$y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) \le 1, i = 1...N$$
$$y_i \ge 0, i = 1...N$$
$$\alpha_i \left[y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - 1 \right] = 0, i = 1...N$$

L'ultima condizione specifica che, ove α_i non sia uguale a 0, allora deve valere la condizione $y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) = 1$. Nel nostro caso, $\boldsymbol{\alpha}$ è dato dal vettore

$$[0\ 0\ 0.023802\ 0\ 0\ 0.074711\ 0\ 0\ 0.098512\ 0]^T$$

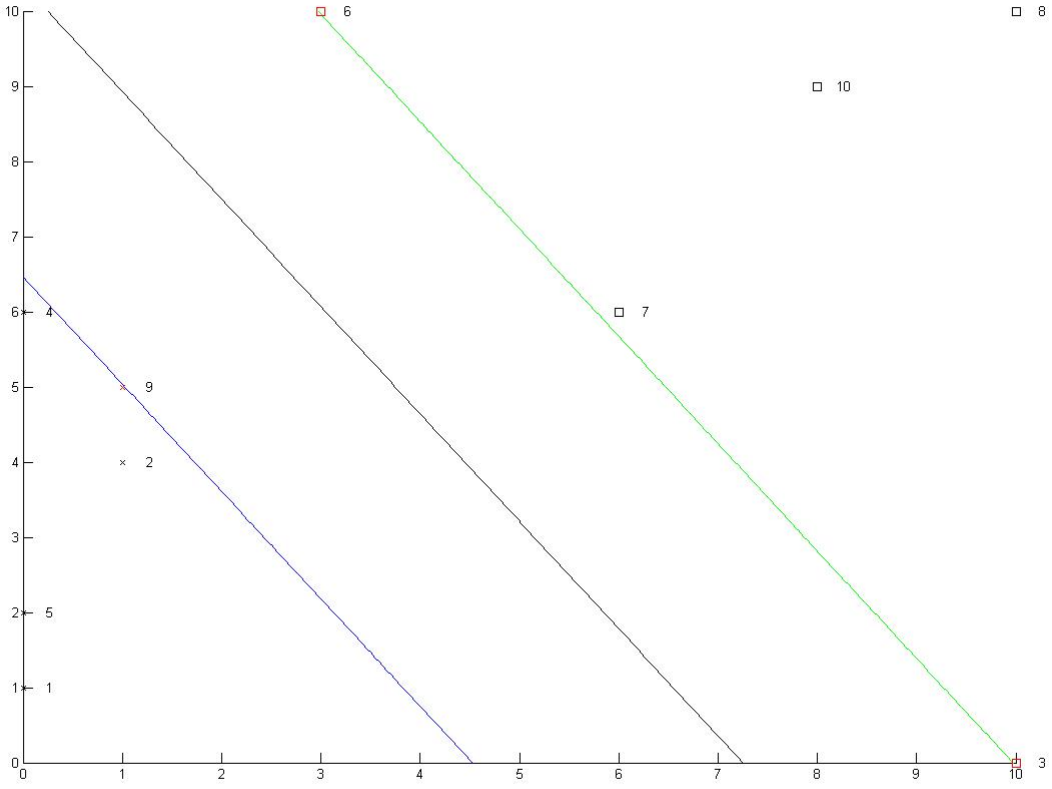
che caratterizza le tuple $\mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_9$ come vettori di supporto. Le equazioni corrispondenti quindi sono date da

$$\mathbf{w} = \alpha_3 y_3 \mathbf{x}_3 + \alpha_6 y_6 \mathbf{x}_6 + \alpha_9 y_9 \mathbf{x}_9$$
$$y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) = 1, i = 3,6,9$$

Analiticamente, i coefficienti del decision boundary sono

$$w_1 = 0.023802 \times 10 + 0.074711 \times 3 - 0.098512 \times 1 = 0.363641$$
$$w_2 = 0.023802 \times 0 + 0.074711 \times 10 - 0.098512 \times 5 = 0.25455$$
$$\left.\begin{array}{l} 0.363641 \times 10 + 0.25455 \times 0 + b = 1 \\ 0.363641 \times 3 + 0.25455 \times 10 + b = 1 \\ 0.363641 \times 1 + 0.25455 \times 5 + b = -1 \end{array}\right\} b = -2.6364$$

Graficamente, il decision boundary e i suoi margini sono raffigurati come segue:



Esercizio 2
È facile verificare che l'attributo che dà il maggior guadagno informativo è l'attributo *x*, tagliato al punto 1.

Esercizio 3
Dal grafico si vede che i vettori $\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5$ hanno distanza 0, che i gruppi di vettori $\mathbf{x}_2, \mathbf{x}_6, \mathbf{x}_9$ e $\mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_{10}$ hanno una similarità molto alta, mentre $\mathbf{x}_3$ è un outlier. Ponendo ε al valore massimo di distanza tra i gruppi, DBScan otterrebbe 3 clusters. Si noti che il primo e il secondo cluster sono molto vicini, per cui un settaggio leggermente più ampio di ε fonderebbe i due cluster.

Esercizio 4
Dalla tabella si desumono i seguenti valori riassuntivi:

x	**+**	**-**
0	0	3
1	0	2
3	1	0
6	1	0
8	1	0
10	2	0

È facile vedere che l'algoritmo ChiMerge effettuerebbe i seguenti passi:

x	**+**	**-**
0	0	3
1	0	2
3	1	0
6-8	2	0
10	2	0

x	**+**	**-**
0	0	3
1	0	2
3	1	0
6-10	4	0

x	**+**	**-**
0-1	0	5
3	1	0
6-10	4	0

x	**+**	**-**
0-1	0	5
3-10	5	0

Il dataset discretizzato così risulterà

	x	***y***	***U***
1	[0,1]	1	-1
2	[0,1]	4	-1
3	[3,10]	0	1
4	[0,1]	6	-1
5	[0,1]	2	-1
6	[3,10]	10	1
7	[3,10]	6	1
8	[3,10]	10	1
9	[0,1]	5	-1
10	[3,10]	9	1

Da tale dataset, otteniamo i seguenti itemset frequenti:

- Itemsets di lunghezza 1: x=[0,1] (supp. 5), x=[3,10] (supp. 5), U=1 (supp. 5) U=-1 (supp. 5);
- Itemsets di lunghezza 2: x=[0,1], U=1 (supp. 5) e x=[3,10], U=-1 (supp. 5).

Da tali itemsets otteniamo le regole

x=[0,1] $\rightarrow$ U=1
U=1 $\rightarrow$ x=[0,1]
x=[3,10] $\rightarrow$ U=-1
U=-1 $\rightarrow$ x=[3,10]