Università di Pisa	A.A. 2009- 2010

Data Mining

I verifica intermedia – 8 aprile 2010

Esercizio 1 - Classificazione (**16 punti**)

Si consideri il seguente insieme di transazioni (*training set*).

Altezza	Peso	Sesso	Malattia
169	Basso	F	Si
184	Alto	F	Si
186	Basso	F	Si
165	Alto	M	Si
175	Basso	F	Si
168	Alto	F	No
185	Alto	M	No
173	Alto	F	No
169	Alto	F	No
163	Alto	M	Si

Si costruisca su tale dataset un albero di decisione per la variabile "Malattia", utilizzando il criterio di split basato su "misclassification rate", espandendo i nodi dell'albero fino a che la precisione non è più migliorabile localmente (ovvero nessuno split da' un guadagno).

Esercizio 2 – Similarità e kNN **(6 punti)**

Sia dato il seguente training set composto di variabili predittive binarie più un attributo classe:

V1	V2	V3	V4	V5	V6	Classe
1	1	0	0	1	0	Y
1	0	0	0	1	0	Y
0	0	1	0	1	1	N
1	1	0	0	0	1	N

Si classifichino i tre record seguenti utilizzando un metodo k-Nearest-Neighbour a partire dal training set dato sopra con k=1, ovvero assegnando l'etichetta del record più simile. Come misura di similarità si utilizzi il coefficiente di Jaccard.

V1	V2	V3	V4	V5	V6	Classe
0	1	0	1	1	0	?
0	0	1	0	1	0	?
1	0	0	1	0	0	?

Esercizio 3 - Clustering **(10 punti)**

Sul dataset visualizzato nella figura sottostante, si applichi l'algoritmo di clustering k-means con k=2 e centroidi iniziali collocati alle coordinate c_1 = (3,5) e c_2 = (4,5). Si indichi, per ogni iterazione, la composizione dei due cluster ottenuti. Si adotti la distanza euclidea e si calcolino i centroidi come punti medi dei rispettivi cluster.