Machine Learning & Data Mining

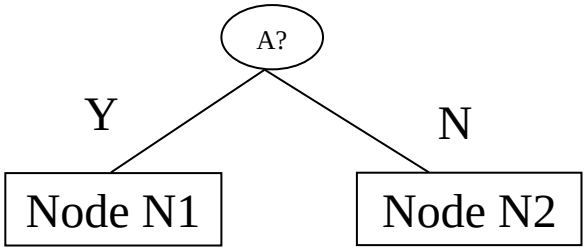
Prova scritta 1 luglio 2015

1) Concept learning

Dare le definizioni di General boundary e Specific bounday nella rappresentazione basata su Version space

2) Alberi di decisione

2.a) Descrivere tre modi differenti per gestire missing values presenti negli esempi di training durante la costruzione di un albero di decisione
2.b) Facendo riferimento all'ipotesi h rappresentata da un albero di decisione, dare la definizione di overfitting per h
2.c) Si consideri:



dove il parent node A ha esempi così distribuiti: C1 = 7 e C2 = 4
e i nodi figli hanno la seguente distribuzione: per N1 C1 = 4 e C2 = 0;per N2 C1 = 5 e C2 = 2
Si ha riduzione di impurità nello split su A usando il *misclassification error*? Motivare adeguatamente la risposta.

3) Valutazione di algoritmi

Descrivere il metodo di cross-validazione denominato *leave-one-out*. Quali sono i vantaggi e svantaggi?

4) Reti Neurali

Descrivere come avviene l'apprendimento nel *unthresholded perceptron* e indicare quali sono le sue proprietà di convergenza.

5) Reti Bayesiane

Descrivere il metodo di apprendimento *Gradient Ascent* per Bayes nets

6) Clustering

6.a) Descrivere i vantaggi e gli svantaggi di trattare un problema di clustering come un problema di ottimizzazione.

6.b) Dato il seguente dataset

Punto	P1	P2	P3	P4	P5	P6	P7
x	4	9	1	14	8	9	10
y	5	2	3	10	8	4	1

Calcolare la matrice delle distanze utilizzando la distanza Euclidea; calcolare il single link e group average hierarchical clustering mostrando i vari passi di scelta ed i rispettivi dendrogrammi.

7) Ensemble Methods

Descrivere vantaggi e svantaggi dell'algoritmo stacking rispetto al bagging

8) Regressione Lineare

Dato il seguente dataset:

x	0	1	2	4	5	6	8	10	11	14	15
y	4	0	8	11	10	15	16	23	26	31	30

Si ipotizza una relazione tra le grandezze del tipo:

$$y=ax+b$$

Si richiede di stimare i coefficienti della retta mediante il metodo dei minimi quadrati e di valutare il grado di adattamento del modello ai dati.

9) SVM

Dati i seguenti 12 elementi del nostro dataset (in R^2), calcolare l'equazione della retta che rappresenta la SVM lineare che separa le due classi massimizzando il margine; calcolare il margine e rappresentare sul piano R^2 i punti del dataset e la retta. Indicare quali sono i support vector.

x	1	1	2	3	4	8	8.5	10	11	14	15	17
y	9	14	3	14	18	13	24	18	32	27	25	28
Class	A	A	B	A	A	B	A	B	A	B	B	B