

Machine Learning & Data Mining

Prova intermedia 10 maggio 2021

1) Concept learning

1.a) Dare le definizioni (formali) di General boundary G e Specific boundary S del Version space.

1.b) Quali sono le condizioni necessarie affinché l'algoritmo Candidate-Elimination possa convergere ad un'ipotesi che descrive correttamente il target concept?

2) Alberi di decisione

2.a) L'uso di Info Gain nella costruzione di un albero di decisione tende a far preferire split che generano un alto numero di nodi successori (figli). Perché questo può essere un problema e come risolverlo?

2.b) Nella costruzione di un albero di decisione, quali sono le condizioni per considerare un nodo come nodo-foglia (senza figli)?

3) Valutazione di algoritmi

3.1) A cosa servono gli intervalli di confidenza per $\text{Errors}_s(h)$?

3.2) Quali sono le curve ROC ottimale e ROC peggiore? Cosa rappresentano i singoli punti sulla curva ROC?

3.3) Qual è il vantaggio di usare la metrica F2 score rispetto alla metrica F1 score?

4) Reti Neurali

4.1) Descrivere l'algoritmo di discesa del gradiente incrementale (stocastico)

4.2) Quali sono le condizioni di terminazione dell'algoritmo Back Propagation

5) Learning Bayesiano

5.1) Si consideri una nuova istanza x con valori binari di classificazione "yes" e "no", uno spazio di 3 ipotesi e un training set D tali che:

$$P(h_1|D) = 0.1, P(h_2|D) = 0.4, P(h_3|D) = 0.5$$

$$P(\text{"no"}|h_1) = 0, P(\text{"yes"}|h_1) = 1$$

$$P(\text{"no"}|h_2) = 0, P(\text{"yes"}|h_2) = 1$$

$$P(\text{"no"}|h_3) = 1, P(\text{"yes"}|h_3) = 0$$

Qual'è il valore v^* calcolato dal classificatore Bayesiano ottimo per x ?

5.2) Specificare la formula che definisce il valore di classificazione v^* del classificatore Bayesiano ottimo.