Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Machine Learning and Data Mining
Tecniche di Apprendimento Automatico
per Applicazioni di Data Mining
Prof. Pier Luca Lanzi
03 Settembre 2007

Solve the following problems and write the answer **inside** the problem box.

The final consists of 5 sheets of paper. It must be returned with all the 5 sheets. No any other sheet can be

NAME

MATRICOLA

Grades

□ **Machine Learning and Data Mining**
Problems 1, 2, 5, 6, and 7

□ **Tecniche di Apprendimento Automatico per Applicazioni di Data Mining**
Problems 1, 2, 3, 4, and 7

Students who completed the term project don't have to answer to problem 7.

Problem 1. Explain the similarities and the differences between naive Bayes classifiers and Bayesian networks.

Problem 2. NASA wants to be able to discriminate between Martians (M) and Humans (H) based on the following characteristics: Green$\in$\{N,Y\}, Legs$\in$\{2,3\}, Height$\in$\{S,T\}, Smelly$\in$\{N,Y\}.
Our available training data is as follows:

Species	Green	Legs	Height	Smelly
M	N	3	S	Y
M	Y	2	T	N
M	Y	3	T	N
M	N	2	S	Y
M	Y	3	T	N
H	N	2	T	Y
H	N	2	S	N
H	N	2	T	N
H	Y	2	S	N
H	N	2	T	Y

Draw a decision tree for the class attribute “Species” using the information gain.

Problem 3. Explain what is clustering and illustrate the k-means algorithm using pseudo-code. Finally, shortly explain how k-medoids differs from k-means.

Problem 4. Consider a naive Bayes classifier with 2 Boolean attributes X (X∈{0,1}) and Y (Y∈{0,1}) and the binary class attributes Z (Z∈{0,1}).

a) Draw the equivalent Bayesian network.

b) How many parameters must be estimated to train such a naive Bayes classifier? Which ones?

Problem 5. Explain the evaluation of classification algorithms using bootstrap. How is it different from crossvalidation?

Problem 6. What is overfitting? The statement "Overfitting is more likely when the set of training data is small" is true or false? (Justify the answer).

Problem 7. A company that sells Data Mining tools contacts you to sell a tool specialized in decision rule mining. They explain you that this tool is extremely powerful in extracting rule-based models for the clustering of nominal attributes. Considering what we discussed during the course, is this a good offer? If yes, why? If no, why?