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
29 Giugno 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. Consider the following training set in the 2-dimensional Euclidean space. X and Y are the attributes, "Class" is the target class. (Suggestion: plot the data).

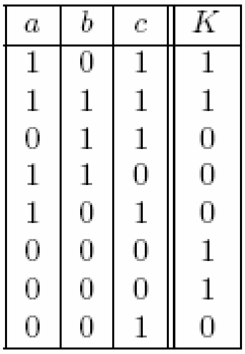
X	Y	Class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

(a) What is the class predicted by the 3-nearest-neighbor classifier for the example (1,1)?

(b) What is the class predicted by the 5-nearest-neighbor classifier for the example (1,1)?

(c) What is the class predicted by the 7-nearest-neighbor classifier for the example (1,1)?

Problem 2. Suppose you are given the following set of data with three Boolean input variables a; b; and c, and a single Boolean output variable K. Assume we are using a naive Bayes classifier to predict the value of K from the values of the other variables.



a	b	c	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

According to the naive Bayes classifier, what is the probability of class K=1 when a=1, b=1, and c=0?

According to the naive Bayes classifier, what is the probability of class K=1 when a=1, b=1, and c is unknown?

Problem 3. What are association rules? What is the goal of association rule mining?

Problem 4. What is a decision tree? Is it **true or false** that decision tree mining can be applied to any type of data? If true, how? If false, why?

Question 5. Explain the difference between apriori and fp-growth.

Question 6. What is Bagging? Is there any relation between Bagging and Boostrap? If yes, which one? If no, why?

Question 7. Your company has around 2000000 customers. In the company database, each customer is described by 400 attributes. You need a model of the high spending customers and therefore you ask to “WeMine!”, a company specialized in data mining.

People at “WeMine!” propose three solutions: one based on decision trees, one based on k-nearest-neighbor, and one based on Naive Bayes classifiers.

Knowing that the model will be deployed on a pocket pc with very limited CPU/memory resources (let’s say around 16Mb of memory for data storage), which of the three options you would choose? Why?