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
20 Luglio 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 data set with one real attribute X and the class attribute Y

X	Y
0	-
0	+
0	-
1	-
1	-
1	+
2	+
2	+

Using information gain, draw the first node of the decision tree for this dataset (answer must be adequately commented).

Problem 2. Consider the following dataset. Extract all the frequent itemsets with a mininum support of 3 and a minimum confidence equal to 0.8.

A	B	C	D
0	0	0	1
0	1	1	0
1	0	1	1
1	1	0	0
0	0	0	1
1	0	0	0
1	0	0	0
1	1	1	1
1	0	1	1

Problem 3. Illustrate the typical steps of a KDD process.

Problem 4. Briefly illustrate sequential covering algorithms.

Problem 5. Consider the following dataset with one real input X and one class attribute Y. Suppose we are using a 1-nearest neighbor. What is the leave one out crossvalidation error for 1-NN on this data set? (Crossvalidation error is computed as the average error of all the training set considered in the crossvalidation). The answer must be adequately commented.

X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

Question 6. What is data stream mining? What are synopses?

Question 7. Your company is trying to solve the following problem. You have some data for which you have no additional information apart from the data themselves. You want to extract as much information as possible.

You interview three data mining consultants for an opening in your firm. The three consultants propose the following approaches.

Consultant A: First we apply a decision tree. Then we prune it as much as possible so that we can identify large portions of the problem space. Then, we apply hierarchical clustering on the subspaces identified by the decision tree to find good description of the subproblems.

Consultant B: I disagree with A. We should first apply clustering and then apply decision tree using the results of the clustering process. In this way we can extract a description of the clusters.

Consultant C: They are both wrong. First apply a hierarchical clustering so that you can find some structure in the data. Then, on each cluster we just found we apply k-mean so that we can actually have a compact description of the clusters.

Which solution is the best? Why the other ones are the worst?