Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico
per Applicazioni di Data Mining
Prof. Pier Luca Lanzi
Ing. Daniele Loiacono
23 Giugno 2008

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 added. No sheet can be removed.

NAME

MATRICOLA

Grades

☐ **Data Mining and Text 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. Suppose you have to apply Naïve Bayes to a dataset described by five nominal attributes (A,B,C,D,E) and the class attribute "CL". (1) What parameters do you need to compute to apply Naïve Bayes classification? (2) Would it be different if the five attributes (A,B,C,D,E) were real-valued? If no, why? If yes, how? (3) Draw the Bayesian Belief network that corresponds to the naïve Bayes classifier.

Problem 2. Given the following dataset, where “inflated” is the class attribute, compute the first two level of the decision tree using the information gain criterion.

Color	size	act	inflated
YELLOW	SMALL	STRETCH	T
YELLOW	SMALL	STRETCH	T
YELLOW	SMALL	DIP	T
YELLOW	SMALL	DIP	T
YELLOW	SMALL	STRETCH	T
YELLOW	SMALL	STRETCH	T
YELLOW	SMALL	DIP	T
YELLOW	SMALL	DIP	T
YELLOW	LARGE	STRETCH	F
YELLOW	LARGE	STRETCH	F
YELLOW	LARGE	DIP	F
YELLOW	LARGE	DIP	F
PURPLE	SMALL	STRETCH	F
PURPLE	SMALL	STRETCH	F
PURPLE	SMALL	DIP	F
PURPLE	SMALL	DIP	F
PURPLE	LARGE	STRETCH	F
PURPLE	LARGE	STRETCH	F
PURPLE	LARGE	DIP	F
PURPLE	LARGE	DIP	F

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

Problem 4. Briefly explain what is overfitting.

Problem 5. Briefly explain what Support Vector Machines (SVMs) are. With respect to what illustrated during the course, explain what are the advantage of SVMs in applications involving unsupevised learning.

Problema 6. What are the differences and the similarities between decision trees, bagging, boosting, and random forests?

Problema 7. A company asks you for help. They have to select the best clustering algorithm for their data, among a set of ten algorithms. They can provide you with two sets of data. One set of data consists in raw data about their customers. The other set contains data that have been labeled by a company expert who labeled customer records as “HIGH” or “LOW” based on their spending level. How would you organize the comparison? Which data would you use? And how would you use the different data?

```
Class,A,B,C,D,E,F
2,*,*,*,*,*,2
1,2,*,*,*,*,1
1,1,2,*,*,*,1
1,1,1,*,*,*,1
1,1,3,2,2,*,1
1,*,*,*,*,4,1
2,1,4,*,*,1,1
2,1,4,*,*,2,1
2,1,4,*,*,3,1
2,1,3,1,1,1,1
2,1,3,1,1,2,1
2,1,3,1,2,1,1
2,1,3,1,2,2,1
1,1,3,1,1,3,1
2,1,3,1,2,3,1
```