Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico

Prof. Pier Luca Lanzi & Ing. Daniele Loiacono
July 9th 2008

NAME

MATRICOLA

Solve the following problems and write the answer **inside** the problem box. Answers must be clearly written. Pencils are not allowed.

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.

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.** Consider the following market-basket data represented in a two-attribute table (where T# is the transaction identifier). Specify all of the association rules that can be deduced from this data with **Support>0.3** and **Confidence>0.5**. To limit your search, only consider association rules that have exactly one item on the left-hand side and one item on the right-hand side.

| T# | item |
|---|---|
| 1 | cookies |
| 1 | milk |
| 2 | beer |
| 2 | pretzels |
| 2 | cookies |
| 2 | eggs |
| 3 | beer |
| 3 | pretzels |
| 4 | beer |
| 4 | cookies |
| 4 | milk |
| 5 | beer |
| 5 | cookies |

**Problem 2.** Given below is a set of instances from a medical diagnosis domain with two attributes **Blood** pressure and **Height** and the class **Disease** that identifies whether the person suffered from a disease. Given the set of instances shown below, calculate the information gain for the attributes Blood and Height.

| Instance | Blood | Height | Disease |
|---|---|---|---|
| x1 | Normal | Normal | Yes |
| x2 | High | Tall | No |
| x3 | Normal | Small | Yes |
| x4 | Normal | Tall | No |
| x5 | High | Normal | Yes |
| x6 | Low | Tall | No |
| x7 | Low | Normal | No |
| x8 | High | Small | No |
| x9 | High | Small | No |
| x10 | Low | Small | Yes |

**Problem 3.** Briefly illustrate how FP-growth works.

**Problem 4.** Shortly explain what is density-based clustering and what are the advantages and or disadvantages with respect to k-means.

**Problem 5.** In the context of Text Mining, briefly illustrate the vector space model.

**Problem 6.** Briefly illustrate what is a social network and describe at least three link mining tasks that are involved in the mining of social networks.

**Problem 7.** You are hired by a company to implement a whole KDD process for their product management cycle. The company is looking for interesting knowledge from their database collecting customers data. More precisely, the company wants to have a model that will predict whether the customers spending level (which can be high, medium or low) based on the other customer information they have. The company tells you that they have a lot of domain knowledge that you should exploit during the process. In particular, they have information about what variables (attributes) are important and they also have information about relations between variables (for instance, they know that the customer age and the customer car model influence the customer spending level). Based on these requirements, how would you organize your solution?