Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico

Prof. Pier Luca Lanzi & Ing. Daniele Loiacono
August 31st 2009

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. This is a closed-book, closed-notes exam. Only non-programmable calculators are allowed. Notes/books/mobile phones are not allowed.

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. In the context of association rule mining, consider the data set below and answer the following questions (answers must be adequately motivated):

1. What is the maximum size of frequent itemsets that can be extracted?
2. What is the maximum number of size-3 itemsets that can be derived from this data set.
3. Find an itemset (of size 2 or larger) that has the largest support.

Table 1: Market basket transactions for Question 6.

Transaction ID	Items Bought
1	{ Milk, Beer, Diapers }
2	{ Bread, Butter, Milk }
3	{ Milk, Diapers, Cookies }
4	{ Bread, Butter, Cookies }
5	{ Beer, Cookies, Diapers }
6	{ Milk, Diapers, Bread, Butter }
7	{ Bread, Butter, Diapers }
8	{ Beer, Diapers }
9	{ Milk, Diapers, Bread, Butter }
10	{ Beer, Cookies }

Problem 2. The following data set will be used to learn a decision tree for predicting whether students are lazy (L) or diligent (D) based on their weight (Normal or Underweight), their eye color (Amber or Violet) and the number of eyes they have (2 or 3 or 4).

Weight	*Eye Color*	*Num Eyes*	*Output*
N	A	2	L
N	V	2	L
N	V	2	L
U	V	3	L
U	V	3	L
U	A	4	D
N	A	4	D
N	V	4	D
U	A	3	D
U	A	3	D

Using Information Gain, what score would be assigned to each of the attributes, when evaluating which feature should be used as the root? Be sure to show your work.

Problem 3. Briefly explain how DBSCAN works.

Problem 4. In the context of Sequential Pattern Mining, briefly explain:

1. What is Sequential Pattern Mining?
2. What data are involved in the typical sequential pattern mining task?
3. Define support in this context?

Problem 5. In the context of Sequential Pattern Mining, briefly explain:

4. What is Sequential Pattern Mining?
5. What data are involved in the typical sequential pattern mining task?
6. Define support in this context?

Problem 6. To train a Support Vector Machine (SVM) it is necessary to solve an expensive optimization problem. Briefly describe how the Chunking method and the Osuna's method can be applied to speed up the training of SVMs. In what way do these methods mainly differ?

Problem 7. A company needs to select a clustering algorithm for their data. The company has two options, k-means and EM. They ask you to briefly explain the differences and the similarities of the two methods. In addition, the company specifies that their data are noisy and asks you which algorithm (in your opinion) is more robust to noise and why.