Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico

Prof. Pier Luca Lanzi & Ing. Daniele Loiacono
June 22nd 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

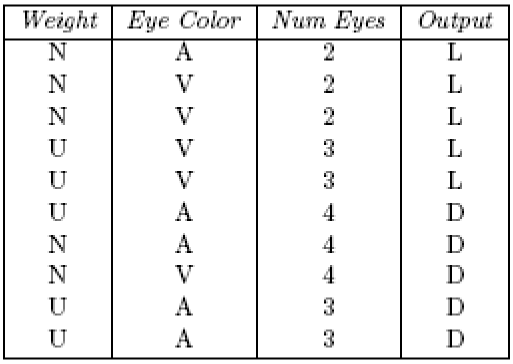
Students who completed the term project don't have to answer to problem 7.

Problem 1. Given the following dataset, apply the first step of the FP-growth algorithm by building the FP-tree (in case of items with the same frequency, sort the items the lexicographical order).

TID	items
T_{100}	{M, O, N, K, E, Y}
T_{200}	{D, O, N, K, E, Y}
T_{300}	{M, A, K, E}
T_{400}	{M, U, C, K, Y}
T_{400}	{C, O, K, I, E}

Then, shortly explain what the next step would be.

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. Illustrate how the typical knowledge discovery process is structured.

Problem 4. Discuss the differences between Naïve Bayes Classifiers and Bayesian Belief Networks. Can you provide an example when the two approaches are actually equivalent? If the answer is yes, illustrate the example. If the answer is no, explain why.

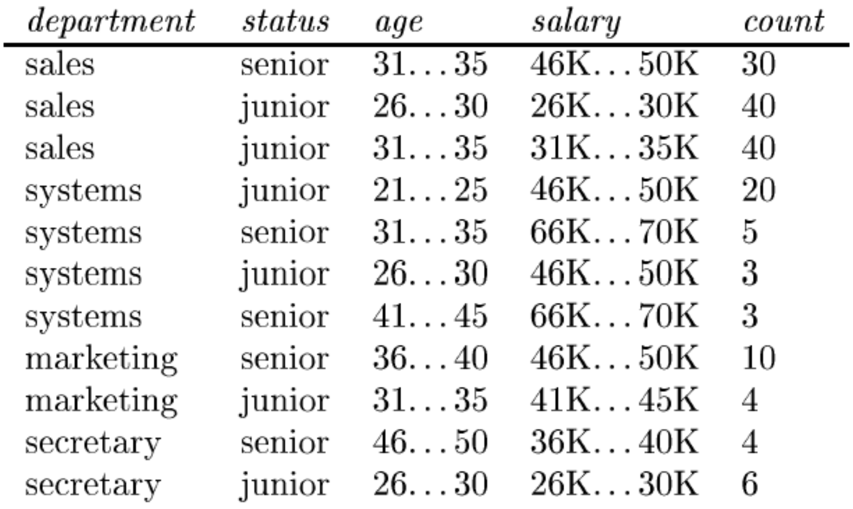
Problem 5. Give one example of an eager learning algorithm and briefly explain how it is or is not incremental.

Problem 6. You are given a classification problem in which there are four possible labels (**right**, **left**, **forward**, **backward**). You are also given six SVM models which have been trained to solve the following binary classification problems:

- one SVM has been trained to discriminate between **right** and **left**
- one SVM has been trained to discriminate between **right** and **forward**
- one SVM has been trained to discriminate between **right** and **backward**
- one SVM has been trained to discriminate between **left** and **forward**
- one SVM has been trained to discriminate between **left** and **backward**
- one SVM has been trained to discriminate between **forward** and **backward**

How would you use these six models to classify an instance **x** as **right**, **left**, **forward**, **backward** (note that, the models discriminate between two classes only, but **x** should be classified using four labels)?

Problem 7. A company has a database in which many instances are duplicated. Therefore, to save space in the company database, it decides to represent a set of instances with the same attribute values by adding an extra attribute called "count". The new attribute represents the number of instances which have the attribute values. For instance, in the following table,



department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

considering the first row, there are 30 instances with a value "sales" for attribute department, a value of "status" equal to senior, a value of "age" of "31…35", and a value of "salary" equal to "46K…50K".

How would you modify the basic decision tree algorithm and the information gain criterion to take into account the new attribute count? (for instance, explain any modification in the algorithm, in the attribute selection, etc.)