Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico

Prof. Pier Luca Lanzi & Ing. Daniele Loiacono
September 19th 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.

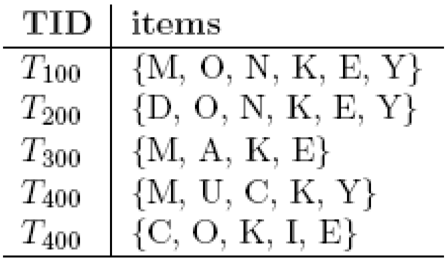
Grades

□ **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

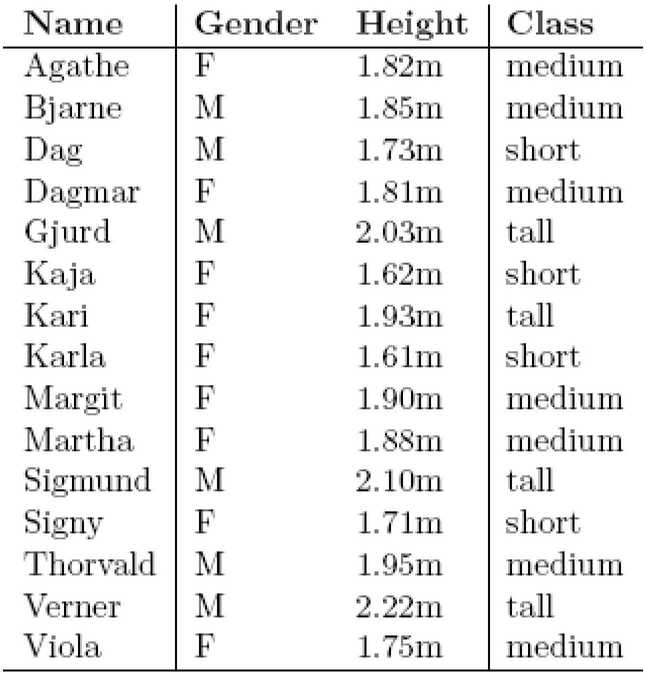
Students who completed the term project don't have to answer to problem 7.

Problem 1. Given the following dataset, let the minimum support threshold be 60% and the minimum confidence threshold be 80%. Find all frequent itemsets and list the strong association rules.



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}

Problem 2. Given the dataset below, find the decision tree that the basic top-down decision-tree induction algorithm using the information-gain measure. Do not use the Name attribute and do not perform any pruning.



Name	Gender	Height	Class
Agathe	F	1.82m	medium
Bjarne	M	1.85m	medium
Dag	M	1.73m	short
Dagmar	F	1.81m	medium
Gjurd	M	2.03m	tall
Kaja	F	1.62m	short
Kari	F	1.93m	tall
Karla	F	1.61m	short
Margit	F	1.90m	medium
Martha	F	1.88m	medium
Sigmund	M	2.10m	tall
Signy	F	1.71m	short
Thorvald	M	1.95m	medium
Verner	M	2.22m	tall
Viola	F	1.75m	medium

Problem 3. What is overfitting? The statement “Overfitting is more likely when the set of training data is small” is true or false? (Justify the answer).

Problem 4. Discuss the difference between partition-based and hierarchical clustering.

Problem 5. What is Bagging? Is there any relation between Bagging and Boostrap? If yes, which one? If no, why?

Problem 6. Suppose you have to evaluate and compare the performance of two classification algorithms. Ilustrate the main steps required to complete this task.

Problem 7. You have run the a-priori algorithm to find association rules in a grocery store transaction database. It takes an unexpectedly long time to complete. On completion, the following is one (of many) rules:

<milk, butter, cheese, bread, flour, sugar, salt, chocolate, apples>) vanilla

Based on seeing the above rule, you should be able to make a good guess as to why the algorithm took a long time. Explain why.