Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Machine Learning and Data Mining
Tecniche di Apprendimento Automatico
per Applicazioni di Data Mining
Prof. Pier Luca Lanzi
February 2nd 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

NAME

MATRICOLA

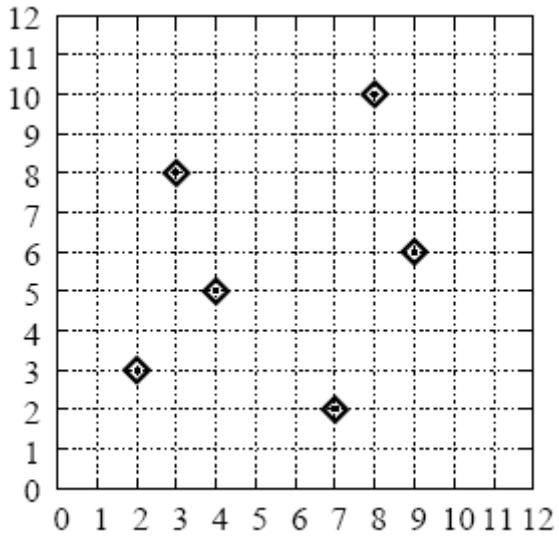
Grades

☐ **Machine Learning and Data 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. Apply K-means with K = 2 to the data above. Use Euclidean distance as your distance measure. Show the steps and the final clusters.



Problem 2. Given below is a set of instances from a medical diagnosis domain with two attributes blood pressure and height and 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. Shortly describe one of the algorithms for mining decision rules considered during the course. Is there any relation between some of the algorithm for decision rule mining, decition trees, and clustering?

Problem 4. With respect to the pruning of decision trees, shortly describe pruning using subtree replacement.

Problem 5. Suppose you need to evaluate one or more classification algorithms. What are the major decisions you have to take to be able to measure the performance of an algorithm and to decide what is the best classification model.

Problem 6. Shortly illustrate what is boosting and how a typical boosting algorithm works.

Problem 7. Assume that the largest frequent itemset is of size k.

How many passes does the apriori algorithm need in worst case? (shortly justify the answer)

A. $k - 1$
B. k
C. $k + 1$
D. k^2
E. 2^k
F. $2^k - 1$

Which of the following is true? (shortly justify the answer)

A. If $A \rightarrow B$ is an association rule, A and B are positively correlated.
B. If $A \rightarrow B$ is an association rule, A and B are at least not negatively correlated.
C. If both $A \rightarrow B$ and $B \rightarrow A$ are association rules, A and B are positively correlated.
D. If both A and B are correlated, then $A \rightarrow B$ is a strong association rule.
E. Association does not imply correlation.