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

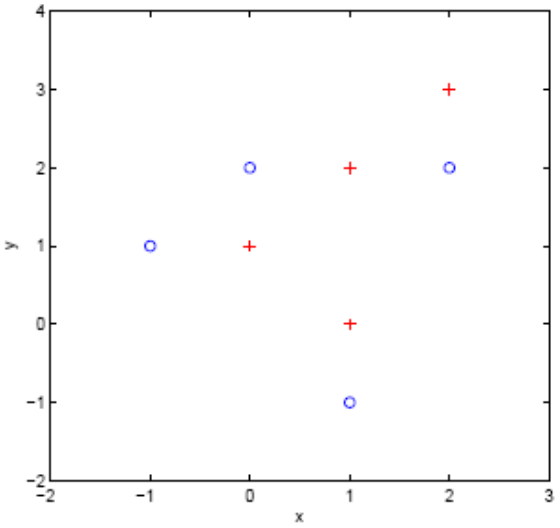
☐ **Data Mining and Text Mining**
Problems 1, 2, 5, 6, and 7

☐ **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. Consider the following training set in the 2-dimensional Euclidean space:

x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$



- What is the prediction of the 3-nearest-neighbor classifier at the point (1,1)?

- What is the prediction of the 5-nearest-neighbor classifier at the point (1,1)?

- What is the prediction of the 7-nearest-neighbor classi¯er at the point (1,1)?

Problem 2. A database has four transactions. Let min_sup = 60% and min_conf = 80%.

TID	Date	Items_bought
T100	10/15/05	{K, A, D, B}
T200	10/15/05	{D, A, C, E, B}
T300	10/19/05	{C, A, B, E}
T400	10/22/05	{B, A, D}

Find all frequent items using Apriori.

Problem 3. Briefly describe Apriori.

Problem 4. Define what is subtree replacement, why it is used, and how it works.

Problem 5. Define what is subtree replacement, why it is used, and how it works.

Problem 6. What are the main challenges and the major steps involved in microarray data analysis?

Problem 7. Answer the following short questions.

- Give one advantage of hierarchical clustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering.

- True or false (and why): Given m data points, the training error converges to the true error as $m \rightarrow \infty$.

- True or false (and why): A classifier trained on less training data is less likely to overfit.

- True or false (and why): In n-fold cross-validation each data point belongs to exactly one test fold, so the test folds are independent. Are the error estimates of the separate folds also independent? So, given that the data in test folds i and j are independent, are e_i and e_j, the error estimates on test folds i and j, also independent?