Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico

Prof. Pier Luca Lanzi & Ing. Daniele Loiacono
January 25th 2010

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. Consider the following distance matrix.

$$\begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 5 & 7 & 0 & & \\ 7 & 9 & 4 & 0 & \\ 8 & 5 & 3 & 1 & 0 \end{pmatrix}$$

Build the dendrogram derived by applying hierarchical clustering with complete linkage.

Problem 2. Briefly illustrate K-Means and discuss its limitations. Then, consider the following examples represented by two real-valued attributes x & y.

ID	X	Y
1	0,5	1,0
2	1,5	3,0
3	2,0	5,0
4	3,0	6,0
5	3,0	1,0
6	1,0	4,0

Show the clusters obtained by applying k-means for three iterations on the data. Also show the centroids for each cluster. For this purpose hypothesize that the number of cluster is 2 and that the initial centroids are examples with ID 1 & 3).

Problem 3. Briefly discuss the similarities and the differences between sequential discovery methods and methods that derive rules from trees.

Problem 4. Briefly describe possible metrics to be used for measuring accuracy of supervised methods.

Problem 5. You are given a paper that contains 15 times the word "set", 8 times the word "computing" and 6 times the word "mining". The paper is part of a collection of 39 papers. In this collection, the word "set" appears in 20 papers, the word "computing" appears in 10 papers, while the word "mining" appears in 5 papers. Compute the Inverse Document Frequency (IDF) for the terms "set", "computing" and "mining" and say, according to IDF, which is the most important keyword among them.

Problem 6. Briefly explain what is the VC-dimension (is the Vapnik-Chervonenkis dimension) of a classifier and why it is important to compare different classification approaches.

Problem 7. Two consultants (A & B) want to sell you their software tools which, according to the two consultants, can support the typical knowledge discovery process from large databases. The two consultants state that:

Consultant A	Consultant B
My product takes as input a table representing the data and can produce either a decision tree, or a set of decision rules, or a naive Bayes classifier, or a set of clusters.	My product takes as input a table representing the data and can apply data selection and cleaning procedure. It can also compute either a decision tree, or a set of decision rules, or a naive Bayes classifier, or a set of clusters. The results can be later validated using several methods.
The system is specialized on machine learning methods.	The system is specialized on supervised learning methods.
Crossvalidation can be applied to all the methods.	Crossvalidation can be applied to all the supervised learning methods.

According to what you know from the course, which one of the two consultants is more convincing? Why? (briefly motivate your decision).