Politecnico di Milano
Facoltà di Ingegneria dell'Informazione

Data Mining and Text Mining
Tecniche di Apprendimento Automatico

Prof. Pier Luca Lanzi & Ing. Daniele Loiacono
July 14th 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

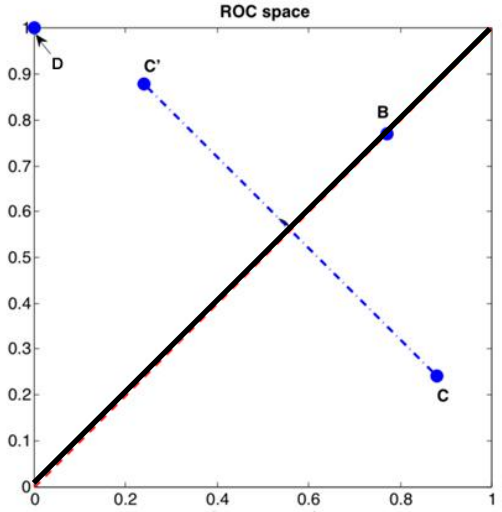
Students who completed the term project don't have to answer to problem 7.

Problem 1. NASA wants to be able to discriminate between Martians (M) and Humans (H) based on the following characteristics: Green$\in${N,Y}, Legs$\in${2,3}, Height$\in${S,T}, Smelly$\in${N,Y}.
Our available training data is as follows:

Species	Green	Legs	Height	Smelly
M	N	3	S	Y
M	Y	2	T	N
M	Y	3	T	N
M	N	2	S	Y
M	Y	3	T	N
H	N	2	T	Y
H	N	2	S	N
H	N	2	T	N
H	Y	2	S	N
H	N	2	T	Y

Apply the learn-one-rule algorithm to the class N.

Problem 2. Consider the example of ROC curve reported below.



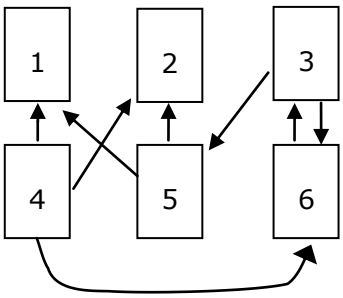
1. What's reported in the x and y axes?
2. What's the meaning of the solid line?
3. What does B represent?
4. What do C and C' represent?
5. What does D represent?

Problem 3. Explain model based clustering. Is k-means an instance of model-based clustering? (briefly motivate the answer).

Problem 4. Shortly describe one of the algorithms for mining decision rules considered during the course. Is there any relation between some of the algorithms for decision rule mining, decition trees, and clustering?

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 which is the best classification model.

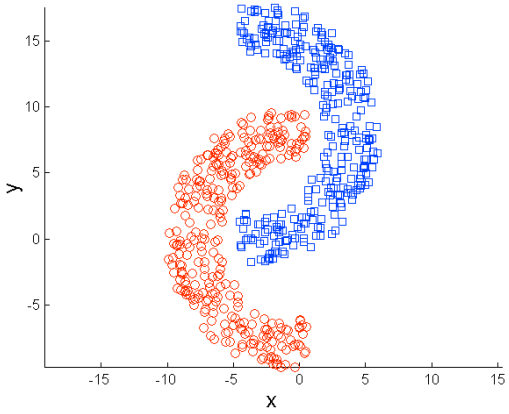
Problem 6. Consider a base set with the following six pages:



Where an arrow an edge between page A and B means that page A has a link to page B.

1. Compute the authority weight and the hub weight for each page in the base set after two iterations of the Hyperlink-Induced Topic Search. Assume that all the weights are initially set to 1 before the first iteration.
2. On the basis of the computed weights, which are the best two authoritative pages and the best two hub pages among the six pages in the base set.

Problem 7. A company presents you with the following data identified by two attributes x and y:



They wish to apply clustering to find two clusters that are clearly visible. They wish to use k-means or EM.

1. In your opinion would k-means or EM be effective? If yes, tell why, if no explain why.
2. Give an example of the result you would expect for k-means and for EM
3. Assuming that k-means or EM work, the company asks you for an alternative algorithm, which one would you suggest and why?