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 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

Using Naïve Bayes, determine the class of the following tuples,

1. <Green=?,Legs=3,Height=S,Smelly=Y>

2. <Green=Maybe,Legs=2,Height=T,Smell=N>

3. <Green=Y,Legs=2,Height=S,Smelly=N>

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 explain how the typical algorithm for building decision trees works. Compare Information Gain with Gini Index.

Problem 4. Shortly describe pruning using subtree replacement.

Problem 5. Consider the following sequence of items:

ABBAACBDCDAABCADCBDC

Apply the "Lossy Counting" algorithm to find the items with at least a support of 30% (a 20% error is considered acceptable). Which are the frequent items according to the Lossy Counting algorithm? Report the final estimates of the items' frequency computed to motivate your answer.

Problem 6. Briefly describe the most typical link mining tasks that can be identified in the analysis of the social networks.

Problem 7. Your company is trying to solve the following problem. You have some data for which you have no additional information apart from the data themselves. You want to extract as much information as possible.

You interview three data mining consultants for an opening in your firm. The three consultants propose the following approaches.

Consultant A: First we apply a decision tree. Then we prune it as much as possible so that we can identify large portions of the problem space. Then, we apply hierarchical clustering on the subspaces identified by the decision tree to find good description of the subproblems.

Consultant B: I disagree with A. We should first apply clustering and then apply decision tree using the results of the clustering process. In this way we can extract a description of the clusters.

Consultant C: They are both wrong. First apply a hierarchical clustering so that you can find some structure in the data. Then, on each cluster we just found we apply k-mean so that we can actually have a compact description of the clusters.

Which solution is the best? Why the other ones are worse?