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
17 Settembre 2007

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. Consider the following dataset

outlook	temperature	humidity	windy?	play?
rainy	cool	normal	Y	*no*
rainy	cool	normal	N	*yes*
rainy	mild	high	Y	*no*
rainy	mild	high	N	*yes*
rainy	mild	normal	N	*yes*
overcast	cool	normal	Y	*yes*
overcast	cool	high	Y	*no*
overcast	mild	high	Y	*yes*
overcast	hot	high	N	*yes*
overcast	hot	normal	N	*yes*
sunny	cool	normal	N	*yes*
sunny	mild	high	N	*no*
sunny	mild	normal	Y	*yes*
sunny	hot	high	Y	*no*
sunny	hot	high	N	*no*

Calculate P(play?=yes, | X = <sunny, hot, high,N>), P(play?=yes, | X = <sunny, hot, high, ?>).
How would the naive Bayes classifier classify the data instance X = <sunny, hot, high,N>?

Problem 2. Given below is a set of instances from a plant diagnosis domain with two attributes plant and type and the class "Defect" which identifies whether the product was defective within the first month. Given the set of instances shown below, calculate the information gain for the attributes Plant and Type.

Instance	Plant	Type	Defect
X1	USA	A	Yes
X2	CAN	B	No
X3	SIN	A	No
X4	SIN	B	No
X5	USA	A	Yes
X6	USA	B	No
X7	SIN	A	No
X8	CAN	B	No
X9	USA	B	No
X10	SIN	A	No

Problem 3. Shortly explain how the typical algorithm for building decision trees works.

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

Problem 5. Define MDL and the MDL principle. Shortly explain the relation between the MDL principle and the Occam's razor.

Problem 6. Name three of the data preprocessing steps that have been discussed during the course and describe them shortly.

Problem 7. Is the decision tree induction algorithm seen during the course (ID3) guaranteed to find an optimal tree (that is, a tree that best classifies the training tuples over all possible trees)? Why or why not?

3

4 Relativamente all'apprendimento supervisionato, illustrare uno degli algoritmi visto a lezione per la derivazione delle regole di decisione. Discutere inoltre se esiste qualche relazione fra gli algoritmi visti a lezione per la generazione delle regole di decisione e il clustering o gli alberi di decisione.

5. Consider the problem of evaluating models obtained by means of a supervised classification algorithm. What are the four main issues that raise?

6.

Febbraio

4 Relativamente al pruning di alberi di decisione, descrivere a parole (senza formule) come funziona il pruning con il metodo di subtree replacement.

5 MLDM Definire cos'è l'MDL e quello che a lezione è stato chiamato "MDL principle". Spiegare brevemente la relazione fra l' "MDL principle" e l'assioma del rasoio di Occam.

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

No. This is a greedy algorithm, so there is no guarantee that it will find an optimal solution. Second, the algorithm explores a subset of the possible tree-space since each node decision is based on a single attribute. A tree with compound decisions might work better.

$$P(sunny|yes) \cdot P(hot|yes) \cdot P(high|yes) \cdot P(N|yes) \cdot P(yes) = \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \cdot \frac{9}{15} = \frac{8}{1215}$$

$$P(sunny|no) \cdot P(hot|no) \cdot P(high|no) \cdot P(N|no) \cdot P(no) = \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{5}{6} \cdot \frac{2}{6} \cdot \frac{6}{15} = \frac{1}{54}$$

$$\frac{8}{1215} = 0.00658 < \frac{1}{54} = 0.0185$$

'No' wins. It happens to agree with the matching training data row.