Sample Midterm Exam

Section A: Multiple-Choice questions

1. Suppose we would like to convert a nominal attribute X with 4 values to a data table with only binary variables. How many new attributes are needed?
   A. 1
   B. 2
   C. 4
   D. 8
   E. 16

2. It was shown that the Naive Bayesian method
   A. can be much more accurate than the optimal Bayesian method
   B. is always worse off than the optimal Bayesian method
   C. can be almost optimal only when attributes are independent
   D. can be almost optimal when some attributes are dependent
   E. None of the above

3. In a medical application domain, suppose we build a classifier for patient screening (True means patient has cancer). Suppose that the confusion matrix is from testing the classifier on some test data.

		Predicted	
		True	False
Actual	True	*TP*	*FN*
	False	*FP*	*TN*

   Which of the following situations would you like your classifier to have?
   A. $FP >> FN$
   B. $FN >> FP$
   C. $FN = FP \times TP$
   D. $TN >> FP$
   E. $FN \times TP >> FP \times TN$
   F. All of the above

4. Consider discretizing a continuous attribute whose values are listed below:
   3, 4, 5, 10, 21, 32, 43, 44, 46, 52, 59, 67
   Using equal-width partitioning and four bins, how many values are there in the first bin (the bin with small values)?
   A. 1
   B. 2
   C. 3
   D. 4
   E. 5
   F. 6

5. Which of the following statements about Naive Bayes is incorrect?
   A. Attributes are equally important.
   B. Attributes are statistically dependent of one another given the class value.
   C. Attributes are statistically independent of one another given the class value.
   D. Attributes can be nominal or numeric
   E. All of the above

6. What are the axes of an ROC curve?
   A. Vertical axis: % of true negatives; Horizontal axis: % of false negatives
   B. Vertical axis: % of true positives; Horizontal axis: % of false positives
   C. Vertical axis: % of false negatives; Horizontal axis: % of false positives
   D. Vertical axis: % of false positives; Horizontal axis: % of true negatives
   E. None of the above

Section B: Long questions

7. Consider the following dataset of a credit card promotion database. The credit card company has authorized a new life insurance promotion similar to the existing one. We are interested in building a classification data mining model for deciding whether to send the customer promotional material.

Customer ID	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	Life Insurance Promotion
1	Y	N	N	M	N
2	Y	Y	Y	F	Y
3	N	N	N	M	N
4	Y	Y	Y	M	Y
5	Y	N	N	F	Y
6	N	N	N	F	N
7	Y	Y	Y	M	Y
8	N	N	N	M	N
9	Y	Y	Y	M	N
10	N	Y	N	F	Y

(a) Build a 3-level decision tree (the root node counts as level-one) using gain ratio. Show your calculations of entropy and information gain.

(b) Build a Naive Bayes classifier for this dataset, by filling in the following with counts and probabilities.

		Life Insurance Promotion (*LIP*)	
		Y	N
Magazine Promotion (*MP*)	Y		
	N		
***P*(*MP* = ? | *LIP* = ?)**	Y		
	N		

		Life Insurance Promotion (*LIP*)	
		Y	N
Watch Promotion (*WP*)	Y		
	N		
***P*(*WP* = ? | *LIP* = ?)**	Y		
	N		

		Life Insurance Promotion (*LIP*)	
		Y	N
Credit Card Insurance (*CCI*)	Y		
	N		
***P*(*CCI* = ? | *LIP* = ?)**	Y		
	N		

		Life Insurance Promotion (*LIP*)	
		Y	N
Sex (*S*)	M		
	F		
***P*(*S* = ? | *LIP* = ?)**	M		
	F		

	Life Insurance Promotion (*LIP*)	
	Y	N
Counts		
***P*(*LIP* = ?)**		

(c) Let $X = (MP = \text{Y}, WP = \text{Y}, CCI = \text{N}, S = \text{F})$. Calculate the conditional probabilities $P(LIP = \text{Y} \mid X)$ and $P(LIP = \text{N} \mid X)$ in terms of $P(X)$.
Hint: Use Bayes theorem

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

and assume class conditional independence, i.e.,

$P(X \mid H) = P(MP = \text{Y} \mid H) \times P(WP = \text{Y} \mid H) \times P(CCI = \text{N} \mid H) \times P(S = \text{F} \mid H)$

where H is $LIP = \text{Y}$ or $LIP = \text{N}$.

Additional space for part (b) answer

Additional space for part (b) answer

(d) Use the Naive Bayes classifier obtained in part (a) and results in part (b) to determine the value of Life Insurance Promotion for the following instance:

Magazine Promotion = Y
Watch Promotion = Y
Credit Card Insurance = N
Sex = F
Life Insurance Promotion = ?

Explain your answers clearly.

~ End ~