Data Mining, Spring 2017: Mid-term Exam

April 20, 12pm – 1:20pm, 2017

Name : _____

Student ID # : _____

Department : _____

You can write your answers in English or in Korean.

	Max	Score
Problem 1	24	
Problem 2	10	
Problem 3	20	
Problem 4	16	
Problem 5	12	
Problem 6	6	
Problem 7	12	
Total	100	

1. (24 pts) True or False (No point for unexplained answer. Explain your answer briefly.)

   (a) For ordinal variables, we cannot perform inequality test.

   (b) The test time (runtime) for K-NN algorithm does not depend on the number of training samples.

   (c) The test time (runtime) for Naïve Bayes algorithm does not depend on the number of training samples.

   (d) Logistic regression model has a linear decision boundary.

   (e) Decision tree algorithm cannot be applied to a dataset having mixed types of variables.

   (f) In linear regression, the more variables we have, the better the general performance is.

2. (10 pts) Short questions.

   (a) Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

   (b) What are the advantages and disadvantages of each cross validation method below?

   - Holdout (test-set) validation

   - LOOCV

   - k-fold CV

3. (20 pts) Consider the data set shown in the following table. (Assume $x_i \in \{0, 1, 2\}$)

	x_1	x_2	x_3	y
s_1	0	2	0	+
s_2	2	0	2	+
s_3	0	0	2	-
s_4	1	2	0	-
s_5	2	2	1	-

(a) Predict the class label for a test sample ($x_1 = 0, x_2 = 2, x_3 = 1$) using the Naive Bayes approach.

(b) Briefly explain what 'Naïve' means in Naïve Bayes algorithm. That is, what is the basic difference between the Naïve Bayes classifier and the exact Bayes classifier?

(c) Describe how a 3-NN algorithm would classify the test example $X_{test} = (0, 2, 1)$ if we use a Euclidean distance.

(d) Now suppose we use the Hamming distance (= the number of attributes on which the two samples disagree) in k-NN. Describe how a 3-NN algorithm would classify the test example $X_{test} = (0, 2, 1)$.

4. (16 pts) Suppose two screening methods for a disease, Model A and Model B, have been tried on the same group of people and we obtained the following confusion matrices. Here, positive class means the patient group.

Model A		Predicted class	
		Yes	No
Actual class	Yes	10	90
	No	0	900

Model B		Predicted class	
		Yes	No
Actual class	Yes	80	20
	No	100	800

(a) Compute sensitivity, specificity, precision, and recall for each model A and B.

(b) How many false negatives and how many false positives are there for each model A and B?

(c) Compare Model A and Model B in terms of accuracy and F-measure, respectively.

(d) Suppose the penalty for missing a true patient is high (e.g., the disease easily spreads, or the disease is fatal). Which model would you like to use, and why?

5. (12 pts) For variable selection, we perform best-subset, forward, and backward selection on a single data set. For each approach, we select $p + 1$ models, containing 0, 1, 2, . . . , p predictors. That is, in the best subset, or in forward selection, we start with the model with 0 predictor, select the best model with 1 predictor based on training RSS, then the best model with 2 predictors, etc. In case of backward selection, we start from the model with p predictors, and then select the best model with $p - 1$ predictors, and so on. Explain your answers:

   (a) Which of the three variable selection models with the same k predictors has the smallest training RSS?

   (b) True or False: The predictors in the k-variable model identified by forward selection are a subset of the predictors in the $(k + 1)$-variable model identified by forward selection.

   (c) Show at least two performance measures (other than training RSS) that can be used for model selection in linear regression.

6. (6 pts) This problem has to do with odds.

   (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

   (c) True or False: in logistic regression, the log odds is modeled as a linear function of predictors.

7. (12 pts) Suppose we have a dataset with five predictors:

$$
\begin{aligned}
X_1 &= GPA \\
X_2 &= IQ \\
X_3 &= \text{Gender (1 for Female and 0 for Male)} \\
X_4 &= \text{Interaction between GPA and IQ (that is, } X_4 = X_1 \times X_2) \\
X_5 &= \text{Interaction between GPA and Gender (that is, } X_5 = X_1 \times X_3)
\end{aligned}
$$

We want to predict the starting salary after graduation using linear regression. Suppose we get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$ from least squares fit.

(a) Which answer among the followings is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0. .

(c) True or False: Since the coefficient $\hat{\beta}_4$ for the GPA/IQ interaction term is very small, there is little evidence of an interaction effect. Justify your answer.

———— SCRATCH PAPER ————-