Applied Data Mining: Final Exam

Due on 12/11/2017, 11:59pm (ET)

Instructor: Hasan Kurban

Student Name

December 7, 2017

Directions

This final exam is due Monday Dec 11, 2017 11:59p.m (ET). **OBSERVE THE TIME**. Absolutely no final exam will be accepted after that time. All the work must be your own. I am providing the LaTeX of this document too. You are not allowed to post questions related to the final exam on Canvas (Piazza/Discussion). If you think that any of the questions is ambiguous, answer it as you understand and explicitly explain your approach.

Problem 1 (10 pt.)

Let $X = \{x_1, x_2, \ldots, x_n\}$ and d be a distance metric over X. Let $\mathcal{X}$ be a partition of k blocks over X, and $\mathcal{Y}$ be a partition of $k+1$ blocks over X. The intrablock sum of distances are:

$$d_x = \sum_{b \in \mathcal{X}} \sum_{i,j \in b, i \neq j} d(i,j) \tag{1}$$

$$d_y = \sum_{b \in \mathcal{Y}} \sum_{i,j \in b, i \neq j} d(i,j) \tag{2}$$

Prove that $d_x \geq d_y$, for all $k = 1, 2, \ldots, n-1$.

Problem 2 (5 pt.)

Choose the best answer. A classification tree generally has

(a) high variance.

(b) low variance.

(c) average variance.

Problem 3 (15 pt.)

Suppose this is the given training set with features, A,B,C,D and label L:

A	B	C	L
1	2	'm'	1
2	3	'm'	1
1	2	'p'	0
3	1	'p'	1
0	0	'a'	0
4	1	'm'	1
1	1	'm'	0

(a) The entropy of the Label is:

i. minimal

ii. maximal

iii. neither maximal nor minimal

Problem 3 (15 pt.) continued on next page...

(b) Using features A,B and treating them as dimension in 2D Euclidean space, the data is

  i. linearly separable

  ii. not linearly separable

(c) Give a *reasonable* separating line for the data.

(d) Give a decision tree for the data (method is up to you).

Problem 4 (3 pt.)

Suppose you've built a classifier and have predictions $\hat{\text{L}}$ for a label L (TID is the tuple ID): What is the error rate?

TID	$\hat{\text{L}}$	L
1	1	0
2	1	1
3	0	0
4	1	1
5	0	1

Problem 5 (12 pt.)

Fill-in the confusion matrix values v_1, v_2, v_3, v_4 using the data above:

$n = 5$	$\hat{\text{L}} = 0$	$\hat{\text{L}} = 1$	
$\text{L} = 0$	v_1	v_2	$v_1 + v_2$
$\text{L} = 1$	v_3	v_4	$v_3 + v_4$
	$v_1 + v_3$	$v_2 + v_4$	

(a) Give the Accuracy

(b) Misclassification Rate

(c) True Positive Rate

(d) Specificity

Problem 6 (14 pt.)

(a) (True or False) The most important stage in the process of data mining is the problem statement.

(b) (True or False) A histogram is kind of partition.

(c) (True or False) A histogram is a kind of probability distribution function.

(d) (True or False) Outliers are always noise objects.

Problem 6 (14 pt.) continued on next page. . .

(e) (True or False) Noise objects can be outliers.

(f) Define data mining.

(g) What does over-fitting mean?

(h) What is the main difference between supervised and unsupervised learning?

Problem 7 (6 pt.)

Consider the following results from a five-fold cross validation

Fold	Error%
1	19.25
2	19.76
3	18.99
4	19.37
5	14.45

(a) Find the average error $\hat{E}$.

(b) (True or False) $\hat{E}$ is a good indicator of the true error E. Explain why/why not?

Problem 8 (5 pt.)

Fill-in the table's cell with Y (yes), N (no), or U (unknown)

Method	Parametric
Linear regression	
knn	
k-means	
decision tree	

Problem 9 (10 pt.)

In this question, you are asked to use the data set below and K-nearest neighbors to predict (X_1, X_2, X_3) = (0,0,0). Note that X_1, X_2, X_3 are the predictors and Y is the response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	0	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

(a) Calculate the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is the prediction for $K = 1$?

(c) What is the prediction for $K = 3$?

Problem 10 (20 pt.)

Load the Carseats data as follows and answer the questions below and provide the R code for each question.

```
> library(ISLR)
> attach(Carseats)
> View(Carseats)
> dim(Carseats)
[1] 400  11
```

Sales variable (1st variable in the data) is the response and the other variables are predictors.

(a) Create a training data set containing a random sample of 200 data points and a test set containing the remaining observations.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test error rate do you obtain (MSE)?

(c) Train random forests over the training set (mtry = 5, ntree = 500). What test error rate do you obtain (MSE)? Use the importance() function to determine which variables are most important (Three most important variables).