Exam 1

Name: _____ ID: _____

Instructions: In order to receive partial/full credit in any problem a complete answer must be provided including any procedures needed, using only the methods covered in the lectures. Guesses receive no credit.

Problem 1

Explain whether each scenario is a classification or regression problem, and provide the number of observations, n and the number of variables, p.

(a) [10 points] A data consulting firm collect a set of data on the top 100 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

(b) [10 points] We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 25 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and five other variables.

Problem 2

The vocabulary “richness" of a text can be quantitatively described by counting the words used once, the words used twice, and so forth. Based on these counts, a linguist proposed the following distances between chapters of the Old Testament book Lamentations.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ .76 & 0 & & & \\ 2.97 & .80 & 0 & & \\ 4.88 & 4.17 & .21 & 0 & \\ 3.86 & 1.92 & 1.51 & .51 & 0 \end{pmatrix} \end{array}$$

(a) [15 points] On the basis of this dissimilarity matrix, cluster the chapters of Lamentations using **complete linkage** and draw the dendrogram that results from hierarchically clustering these five observations. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) [5 points] Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?

Problem 3

(a) [10 points] Suppose we want to compute 10-Fold Cross-Validation error on 200 training examples. We need to compute error N_1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N_2, and test the model on the data of size N_3. What are the appropriate numbers for N_1, N_2, and N_3?

(b) [5 points] A classifier that attains 100% accuracy on the training set and 80% accuracy on test set is better than a classifier that attains 80% accuracy on the training set and 85% accuracy on test set. True or False? Why?

(c) [10 points] Suppose we want to predict the average credit card debt for a number of individuals using the predictor gender (i.e. males and females). Since the qualitative predictor has only two levels, create a dummy variable that takes on two possible numerical values and use this variable as a predictor to write out the regression equation.

Problem 4

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

where $\lambda \geq 0$ is the tuning parameter. Use the above equation to answer the following questions:

(a) [5 points] What happens when $\lambda = 0$.

(b) [5 points] What happens when $\lambda \to \infty$.

(c) [5 points] How do we select a good value for λ?

Problem 5

A credit card company developed a classifier for predicting whether or not an individual will default on the basis of credit card balance and student status. We evaluate the classifier on 10,000 test observations. Here is the confusion matrix:

		True default status	
		No	Yes
Predicted default status	No	9, 644	252
	Yes	23	81

(a) [5 points] Compute the accuracy of the classifier.

(b) [5 points] Compute the precision of the classifier.

(c) [5 points] Compute the recall of the classifier.

(d) [5 points] Compute the F_1 score.