DATA 558
SPRING QUARTER 2022

Homework # 1
Due Via Online Submission to Canvas: Wed, April 13 at 5pm

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

1. Suppose that you are interested in performing regression on a particular dataset, in order to answer a particular scientific question. You need to decide whether to take a parametric or a non-parametric approach.

   (a) In general, what are the pros and cons of taking a parametric versus a non-parametric approach?

   (b) What properties of the data or scientific question would lead you to take a parametric approach?

   (c) What properties of the data or scientific question would lead you to take a non-parametric approach?

   Explain your answers.

2. In each setting, would you generally expect a flexible or an inflexible statistical machine learning method to perform better? Justify your answer.

   (a) Sample size n is very small, and number of predictors p is very large.

   (b) Sample size n is very large, and number of predictors p is very small.

   (c) Relationship between predictors and response is highly non-linear.

   (d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

3. For each scenario, determine whether it is a regression or a classification problem, determine whether the goal is inference or prediction, and state the values of n (sample size) and p (number of predictors).

(a) I want to predict each student's final exam score based on his or her homework scores. There are 50 students enrolled in the course, and each student has completed 8 homeworks.

(b) I want to understand the factors that contribute to whether or not a student passes this course. The factors that I consider are (i) whether or not the student has previous programming experience; (ii) whether or not the student has previously studied linear algebra; (iii) whether or not the student has taken a previous stats/probability course; (iv) whether or not the student attends office hours; (v) the student's overall GPA; (vi) the student's year (e.g. freshman, sophomore, junior, senior, or grad student). I have data for all 50 students enrolled in the course.

4. This problem has to do with the bias-variance trade-off and related ideas, in the context of regression. For (a) and (b), it's okay to submit hand-sketched plots: you are not supposed to actually compute the quantities referred to below on data; instead, this is a thought exercise.

   (a) Make a plot, like the one we saw in class, with "flexibility" on the x-axis. Sketch the following curves: squared bias, variance, irreducible error, reducible error, expected prediction error. Be sure to label each curve. Indicate which level of flexibility is "best".

   (b) Make a plot with "flexibility" on the x-axis. Sketch curves corresponding to the training error and the test error. Be sure to label each curve. Indicate which level of flexibility is "best".

   (c) Describe an $\hat{f}$ that has extremely low bias, and extremely high variance. Explain your answer.

   (d) Describe an $\hat{f}$ that has extremely high bias, and zero variance. Explain your answer.

5. We now consider a classification problem. Suppose we have 2 classes (labels), 25 observations per class, and $p = 2$ features. We will call one class the "red" class and the other class the "blue" class. The observations in the red class are drawn i.i.d. from a $N_p(\mu_r, I)$ distribution, and the observations in the blue class are drawn i.i.d. from a $N_p(\mu_b, I)$ distribution, where $\mu_r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is the mean in the red class, and where $\mu_b = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}$ is the mean in the blue class.

   (a) Generate a training set, consisting of 25 observations from the red class and 25 observations from the blue class. (You will want to use the R function `rnorm`.) Plot the training set. Make sure that the axes are properly labeled, and that the observations are colored according to their class label.

   (b) Now generate a test set consisting of 25 observations from the red class and 25 observations from the blue class. On a single plot, display both the training and test set, using one symbol to indicate training observations (e.g. circles) and another symbol to indicate the test observations (e.g. squares). Make sure that the axes are properly labeled, that the symbols for training and test observations are explained in a legend, and that the observations are colored according to their class label.

(c) Using the `knn` function in the library `class`, fit a k-nearest neighbors model on the training set, for a range of values of k from 1 to 20. Make a plot that displays the value of $1/k$ on the x-axis, and classification error (both training error and test error) on the y-axis. Make sure all axes and curves are properly labeled. Explain your results.

(d) For the value of k that resulted in the smallest test error in part (c) above, make a plot displaying the test observations as well as their true and predicted class labels. Make sure that all axes and points are clearly labeled.

(e) Recall that the Bayes classifier assigns an observation to the red class if $Pr(Y = red|X = x) > 0.5$, and to the blue class otherwise. The *Bayes error rate* is the error rate associated with the Bayes classifier. What is the value of the Bayes error rate in this problem? Explain your answer.

6. We will once again perform k-nearest-neighbors in a setting with $p = 2$ features. But this time, we'll generate the data differently: let $X_1 \sim \text{Unif}[0, 1]$ and $X_2 \sim \text{Unif}[0, 1]$, i.e. the observations for each feature are i.i.d. from a uniform distribution. An observation belongs to class "red" if $(X_1-0.5)^2+(X_2-0.5)^2 > 0.15$ and $X_1 > 0.5$; to class "green" if $(X_1 - 0.5)^2 + (X_2 - 0.5)^2 > 0.15$ and $X_1 \leq 0.5$; and to class "blue" otherwise.

(a) Generate a training set of $n = 200$ observations. (You will want to use the R function `runif`.) Plot the training set. Make sure that the axes are properly labeled, and that the observations are colored according to their class label.

(b) Now generate a test set consisting of another 200 observations. On a single plot, display both the training and test set, using one symbol to indicate training observations (e.g. circles) and another symbol to indicate the test observations (e.g. squares). Make sure that the axes are properly labeled, that the symbols for training and test observations are explained in a legend, and that the observations are colored according to their class label.

(c) Using the `knn` function in the library `class`, fit a k-nearest neighbors model on the training set, for a range of values of k from 1 to 50. Make a plot that displays the value of $1/k$ on the x-axis, and classification error (both training error and test error) on the y-axis. Make sure all axes and curves are properly labeled. Explain your results.

(d) For the value of k that resulted in the smallest test error in part (c) above, make a plot displaying the test observations as well as their true and predicted class labels. Make sure that all axes and points are clearly labeled.

(e) In this example, what is the Bayes error rate? Justify your answer, and explain how it relates to your findings in (c) and (d).

7. This exercise involves the `Boston` housing data set, which is part of the `ISLR2` library.

   (a) How many rows are in this data set? How many columns? What do the rows and columns represent?

   (b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

   (c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

   (d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

   (e) How many of the suburbs in this data set bound the Charles river?

   (f) What are the mean and standard deviation of the pupil-teacher ratio among the towns in this data set?

   (g) Which suburb of Boston has highest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

   (h) In this data set, how many of the suburbs average more than six rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.