STAT 435
SPRING QUARTER 2018

Homework # 3
Due Friday, April 20, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

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, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

1. A random variable X has an Exponential(λ) distribution if its probability density function is of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases},$$

where $\lambda > 0$ is a parameter. Furthermore, the mean of an Exponential(λ) random variable is $1/\lambda$.

Now, consider a classification problem with $K = 2$ classes and a single feature $X \in \mathbb{R}$. If an observation is in class 1 (i.e. $Y = 1$) then $X \sim$ Exponential(λ_1). And if an observation is in class 2 (i.e. $Y = 2$) then $X \sim$ Exponential(λ_2). Let π_1 denote the probability that an observation is in class 1, and let $\pi_2 = 1 - \pi_1$.

(a) Derive an expression for $\Pr(Y = 1 \mid X = x)$. Your answer should be in terms of x, λ_1, λ_2, π_1, π_2.

(b) Write a simple expression for the Bayes classifier decision boundary, i.e., an expression for the set of x such that $\Pr(Y = 1 \mid X = x) = \Pr(Y = 2 \mid X = x)$.

(c) **For part (c) only**, suppose $\lambda_1 = 2$, $\lambda_2 = 7$, $\pi_1 = 0.5$. Make a plot of feature space. Clearly label:

i. the region of feature space corresponding to the Bayes classifier decision boundary,

ii. the region of feature space for which the Bayes classifier will assign an observation to class 1,

iii. the region of feature space for which the Bayes classifier will assign an observation to class 2.

(d) Now suppose that we observe n independent training observations,

$$(x_1, y_1), \dots, (x_n, y_n).$$

Provide simple estimators for λ_1, λ_2, π_1, π_2, in terms of the training observations.

(e) Given a test observation $X = x_0$, provide an estimate of

$$P(Y = 1 \mid X = x_0).$$

Your answer should be written *only* in terms of the n training observations $(x_1, y_1), \dots, (x_n, y_n)$, and the test observation x_0, and *not* in terms of any unknown parameters.

2. We collect some data for students in a statistics class, with predictors $X_1 =$ number of lectures attended, $X_2 =$ average number of hours studied per week, and response $Y =$ receive an A. We fit a logistic regression model, and get coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.

(a) Write out an expression for the probability that a student gets an A, as a function of the number of lectures she attended, and the average number of hours she studied per week. Your answer should be written in terms of X_1, X_2, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.

(b) Write out an expression for the minimum number of hours a student should study per week in order to have at least an 80% chance of getting an A. Your answer should be written in terms of X_1, X_2, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.

(c) Based on a student's value of X_1 and X_2, her predicted probability of getting an A in this course is 60%. If she increases her studying by one hour per week, then what will be her predicted probability of getting an A in this course?

3. When the number of features p is large, there tends to be a deterioration in the performance of K-nearest neighbors (KNN) and other approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality. We will now investigate this curse.

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

(c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

(d) Using your answers to parts (a)-(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1$, 2, and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$*, a hypercube is simply a line segment, when* $p = 2$ *it is a square.*

4. Pick a data set of your choice. It can be chosen from the `ISLR` package (but not one of the data sets explored in the Chapter 4 lab, please!), or it can be another data set that you choose. Choose a binary qualitative variable in your data set to be the response, Y. (By *binary qualitative variable*, I mean a qualitative variable with $K = 2$ classes.) If your data set doesn't have any binary qualitative variables, then you can create one (e.g. by dichotomizing a continuous variable: create a new variable that equals 1 or 0 depending on whether the continuous variable takes on values above or below its median). I suggest selecting a data set with $n \gg p$.

(a) Describe the data. What are the values of n and p? What are you trying to predict, i.e. what is the meaning of Y? What is the meaning of the features?

(b) Split the data into a training set and a test set. Perform LDA on the training set in order to predict Y using the features. What is the training error of the model obtained? what is the test error?

(c) Perform QDA on the training set in order to predict Y using the features. What is the training error of the model obtained? what is the test error?

(d) Perform logistic regression on the training set in order to predict Y using the features. What is the training error of the model obtained? what is the test error?

(e) Perform KNN on the training set in order to predict Y using the features. What is the training error of the model obtained? what is the test error?

(f) Comment on your results.