STA 545 Statistical Data Mining I, Fall 2020

Homework 2, due: Wednesday 9/16/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (40 points) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

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

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbors.

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

   (b) What is our prediction with $k = 1$?

   (c) What is our prediction with $k = 3$?

   (d) Suppose we only use features X_2 and X_3. Please draw the decision boundary of the k-nearest neighbor classifier with $k = 1$ in a figure.

   (e) If the Bayes decision boundary in this problem is highly nonlinear, would we expect the best value for k to be large or small? Why?

2. (60 points) Data for this question come from the handwritten ZIP codes on envelopes from U.S. postal mail. Each image is a segment from a five digit ZIP code, isolating a single digit. The images are 16×16 eight-bit grayscale maps, with each pixel ranging in intensity

from 0 to 255. The images have been normalized to have approximately the same size and orientation. The task is to predict, from the 16×16 matrix of pixel intensities, the identity of each image (0, 1, . . . , 9) quickly and accurately. The zipcode data are available from the book website www-stat.stanford.edu/ElemStatLearn. Please consider only the 2's and 3's in the data.

(a) Fit a linear regression model where we code $Y = 1$ if the label of the image is 2, and $Y = -1$ if the label of the image is 3. Show both the training misclassification error and test misclassification error for this binary classification problem.

(b) Consider the k-nearest neighbor classifiers with $k = 1$, 3, 5, 7 and 15. Show both the training error and test error for each choice.