STATS 415 Homework 5

Due by 11:59pm on Thursday, Oct 15, 2020

1. Suppose we fit logistic regression to predict the probability a Stats 415 student gets an A in the class, from two variables. The variable are average hours of study per week (X_1) and GPA in other statistics courses taken (X_2). The model estimates $\beta_0 = -4, \beta_1 = 0.05, \beta_2 = 1$. (Note: these are made up numbers! Do not try to predict your grades with them. 10 points per question.)

   (a) Predict the probability of getting an A for a student who studies 5 hours a week and has a GPA of 3.5 in other statistics courses.

   (b) What are the odds that this student will get an A?

   (c) How many hours a week does this student need to study for the model to predict a 50% chance of getting an A?

2. This question continues Q2 from Homework 4. Use the same data, the same split into training and test, and the same four variables you chose to use as predictors. (10 points per question.)

   (a) Perform logistic regression on the training data in order to predict `mpg01` using the four quantitative variables you chose in Homework 4. Comment on the significance of predictors.

   (b) Report the training and the test errors for logistic regression. Make a plot similar to the plots in HW4, showing true and predicted class labels from logistic regression plotted against the same two variables you used before.

   (c) Using your fitted model, estimate the probability of a car having mpg above 25 if its four predictors you used are all at the median values for the training dataset.

   (d) Perform KNN classification on the training data. Make plots of the training classification error and the test classification error as a function of the number of neighbors K (or $1/K$; if you use $1/K$, make sure the x-axis is on the log scale). Which K gives the best performance on the training data? On the test data?

(e) Report the training and the test errors for KNN with your choice of K. Make a plot similar to the plots in HW4, showing true and predicted class labels from KNN plotted against the same two variables you used before.

(f) Describe how KNN can be applied to estimate the probability in (c). What are the potential factors that can jeopardize the estimation?

(g) Compare and contrast the performance of LDA, QDA (take from HW 4), logistic regression, and KNN on this dataset. What do your results suggest about the distribution of the data? About the nature of the boundary between classes?

Please limit your answer to Q2 to 8 pages, organized into a coherent typed data analysis report. Answers to Q1 may be either typed or handwritten. Please clearly write your name, your UMID, and your GSI/lab number on the homework.