STAT 503 Homework 2

Due Friday, Feb 11, 2022
Submit online on Canvas by 11:59pm

Part 1.

1. This problem is about simple properties of logistic regression. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied for the final, X_2 = undergrad GPA in statistics classes, and $Y = c_1$ if the final grade is over 90% and c_2 otherwise. We fit a logistic regression and obtain estimates $\hat{\beta}_0 = -5$, $\hat{\beta}_1 = 0.1$, $\hat{\beta}_2 = 1$.

   (a) Write out the resulting equation for the probability of receiving over 90% on the final.

   (b) According to this model, how many hours would a student with an undergrad statistics classes GPA of 3.5 need to study in order to have a 50% chance of getting over 90% on the final?

2. This problem is about understanding bias and variance in a very simple setting.

   (a) Suppose you observe n i.i.d. observations $x_1, \ldots, x_n$ from a distribution with mean 0 and variance σ^2. You estimate the mean of this distribution (unknown to you) with the sample average,

   $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

   Calculate the bias, the variance, and the MSE of $\bar{x}$ as an estimator of the true mean (which is 0). What happens when $n \to \infty$?

   (b) Suppose now that the observations in your sample are all from the same distribution with mean 0 and variance σ^2, but are in fact correlated, and the correlation $\text{cor}(x_i, x_j) = \rho$ for all pairs $i \neq j$. Calculate the bias, the variance, and the MSE of $\bar{x}$ as an estimator of the true mean (which is 0). What happens when $n \to \infty$?

   (c) Suppose the n observations are in fact independent, but the distribution is contaminated: 90% of your sample came from the original distribution with mean 0 and variance σ^2, and the remaining 10% came from a distribution with mean $\mu > 0$ and the same variance σ^2. (This can happen, for example, in KNN, when the local neighborhood gets too large and includes points from another class). You can assume that $0.9n$ and $0.1n$ are integers. Calculate the bias, the variance, and the MSE of $\bar{x}$ as an estimator of the true mean (which is 0). What happens when $n \to \infty$? For what values of μ will the squared bias term be larger than the variance term?

Part 2. We will use the same dataset `College` as in Homework 1, with the same task of predicting whether the school is public or private (variable `Private`) from the other 17 variables.

1. Set aside the same 20% of the data to use as the test set as you did in HW1. Fit a KNN classifier. Plot the training and test errors as a function of the number of the nearest neighbors used, k. Comment on the results.

2. Perform LOOCV, 5-fold, and 10-fold CV and plot cross-validated error curves as a function of k. Report the best k according to each of the cross-validations and to the test data. Comment on similarities and differences.

3. Report, in a table, the training and the test errors for all CV-suggested choices of k for KNN, along with the training and test errors that you have obtained on the same problem from LDA, QDA, and logistic regression. Comment on their relative performance and on what that might suggest about the distribution of the classes and the nature of the boundary between them.