STA4241 Homework 3, Fall 2021

Please turn in your own work, though you may discuss these problems with your classmates, professor, and TA. The assignment is due on Wednesday, October 13th at 11:59pm.

(1) Read in the Crabs data set off of the course website to answer the following questions. The variable y is the outcome of interest, and the remaining variables are to be used as predictors.

(i) What types of variables do you have as predictors? How do you intend to include them in any statistical model that you might run? Does this choice depend on which statistical approach you are using?

(ii) Hold out a portion of the data set to be the testing data set, and use the rest as training data. Use support vector machines with a radial kernel and assess the sensitivity of results to the choice of the γ tuning parameter. Does the model begin to overfit to the training data?

(iii) Perform the same exercise as in (ii), but use a polynomial kernel and vary the degree of the polynomial.

(iv) Let's return to the KNN algorithm for classification. For this problem, do not use the hold out data from the previous two parts, i.e. the training data is your entire data set. I want you to run 10-fold cross validation on the KNN algorithm to find the optimal value of K in this data set. You must code this by hand. Do not use any pre-existing cross validation codes that are built into R to find the optimal value of K. For this question, please paste your R code for running the 10-fold cross validation along with your findings.

(v) In this question, we will compare the performance of all of the classification algorithms that we have used so far. This includes logistic regression, LDA, QDA, KNN, SVMs with a polynomial kernel, and SVMs with a radial kernel. For any method that has tuning parameters, select them using cross-validation. Note for this question, you may use built in cross-validation functions (such as tune() for SVMs) for all algorithms. I want you to randomly select 25 subjects in the data to be the validation data and the rest to be your training data. Apply each algorithm on the training data and use them to classify the testing data outcomes. Keep track of the classification error rate for each algorithm. Do this 100 times, where each time you randomly generate 25 observations to be the testing data.

(a) Which algorithm has the best performance, on average across the 100 testing data sets?

(b) Make a boxplot that shows the distribution of error rates for each estimator across the 100 data sets. This plot should look similar to the plot on slide 53 of lecture 5. Comment on your findings.

(2) Suppose that we observe data X_i for $i = 1, \ldots, 100$ and that these variables are uniformly distributed on the interval $[0, 10]$. Now suppose that I am interested in performing inference on the 0.8 quantile. That means that I would like to estimate, and provide a confidence interval, for the value $q_{0.8}$ that satisfies $P(X_i < q_{0.8}) = 0.8$.

(i) In this case we know the distribution of X_i. What is the value of $q_{0.8}$?

(ii) Suppose we don't know the distribution of X_i. What is a good estimator of $q_{0.8}$?

(iii) Simulate one data set as above. Construct an estimate and confidence interval for $q_{0.8}$ using the bootstrap.

(iv) Now run a simulation study to assess the performance of your bootstrapped confidence intervals by finding the 95% interval coverage provided by your intervals. I would like you to assess the performance of both the percentile method for constructing confidence intervals, as well as the method that finds the standard error of the estimator and then uses the standard confidence interval formula that adds or subtracts 1.96 times the standard error. In what percentage of your simulations do the bootstrap intervals cover the true parameter?

(v) Run the same simulation as in (iv) but now perform inference on $q_{0.99}$. Comment on any differences you find, and explain why you think there are differences.

(3) Let q_α be the α quantile of the bootstrap replicates $\widehat{\theta}^{(b)}$. Prove that the confidence interval from the bootstrap given by $(2\widehat{\theta} - q_{1-\alpha/2}, 2\widehat{\theta} - q_{\alpha/2})$ is approximately valid in the sense that

$$P(2\widehat{\theta} - q_{1-\alpha/2} < \theta < 2\widehat{\theta} - q_{\alpha/2}) \approx 1 - \alpha$$

An important hint is that the distribution of $\theta - \widehat{\theta}$ is approximated by $\widehat{\theta} - \widehat{\theta}^{(b)}$