STAT 4360 (Introduction to Statistical Learning, Fall 2021)

Mini Project 3

Instructions:

- Due date: March 23rd, 2023.
- Total points = 30 + (bonus points 10)
- Submit a typed report.
- It is OK to discuss the project with other students in the class, but each student must write their own code and answers. If your submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: R or Python code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.
- Section 1 of the report must be limited to six pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the diabetes dataset from Mini Project 2. We will take `Outcome` as the response, the other variables as predictors, and all the data as training data. We would like to understand how the predictors are related with the response.

   (a) Perform an exploratory analysis of data.

   (b) Build a "reasonably good" logistic regression model for these data. There is no need to explore interactions. Carefully justify all the choices you make in building the model.

   (c) Write the final model in equation form. Provide a summary of estimates of the regression coefficients, the standard errors of the estimates, and 95% confidence intervals of the coefficients. Interpret the estimated coefficients of at least two predictors. Provide training error rate for the model.

2. Consider the diabetes dataset from #1. Use all predictors for all the models considered for this problem.

   (a) Fit a logistic regression model using all predictors in the data. Provide its error rate, sensitivity, and specificity based on training data.

   (b) Write your own code to estimate the test error rate of the model in (a) using LOOCV.

   (c) Verify your results in (b) using a package. You can use `cv.glm` in R, or you may use the `caret` package (`https://topepo.github.io/caret/`) for doing so as it is not restricted to the GLMs. For Python, take a look at the `sklearn.model_selection.LeaveOneOut`. Make sure the two results match.

(d) For the logistic regression model you proposed in #1, estimate the test error rate using LOOCV.

(e) Repeat (d) using LDA from Mini Project #2.

(f) Repeat (d) using QDA from Mini Project #2.

(g) Fit a KNN with K chosen optimally using the LOOCV estimate of test error rate. Repeat (d) for the optimal KNN. (You may explore `tune.knn` function for finding the optimal value of K but this is not required.)

(h) Compare the results from the various classifiers. Which classifier would you recommend? Justify your answer.

3. Consider the oxygen saturation data stored in `oxygen_saturation.txt` file. The data consist of measurements of percent saturation of hemoglobin with oxygen in 72 adults, obtained using an oxygen saturation monitor (`OSM`, method 1) and a pulse oximetry screener (`POS`, method 2). You can read about oxygen saturation on Wikipedia, `https://en.wikipedia.org/wiki/Oxygen_saturation_(medicine)`. We are primarily interested in evaluating *agreement* between the two methods for measuring oxygen saturation.

   (a) Make a scatterplot of the data and superimpose the 45^o line. Next, make a boxplot of absolute values of differences in the measurements from the two methods. Comment on the extent of agreement between the methods. Note that the methods would have *perfect agreement* if all the points in the scatterplot fell on the 45^o line, or equivalently, all the differences were zero.

   (b) Let Y_1 and Y_2 denote the population of observations of methods 1 and 2, respectively, and $D = Y_1 - Y_2$ denote their difference. Let θ be the *total deviation index* (TDI) between the two methods. For a given large probability p, it is defined as the pth quantile of $|D|$. Here we will take $p = 0.90$. Argue that smaller values for θ imply better agreement.

   (c) Provide a point estimate $\hat{\theta}$ of θ. (Tip: If the population parameter is a quantile, what should be its natural estimator?)

   (d) Write your own code to compute (nonparametric) bootstrap estimates of bias and standard error of $\hat{\theta}$, and a 95% *upper confidence bound* for θ computed using the percentile method. Interpret the results.

   (e) Repeat the computation in (d) using `boot` package in R, or check `bootstrap` function from the SciPy library (`scipy.stats.bootstrap`) in Python, and compare your results.

   (f) State your conclusion about the extent of agreement between the two methods. Would you say that the methods agree well enough to be used interchangeably in practice? Justify.

4. Bonus Question: Recall the pdf/pmf for the distributions of exponential family

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

   for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$.

   (a) Verify: $E(Y) = b'(\theta)$.

   (b) Verify: If $Y \sim \text{Poisson}(\mu)$, we have $\theta = \log(\mu), \phi = 1$, and

$$b(\theta) = \exp(\theta), a(\phi) = \phi, c(y,\phi) = -\log(y!).$$