STAT 4360 (Introduction to Statistical Learning, Fall 2021)

Mini Project 2

Instructions:

- Due date: Feb 16, 2023.
- Total points = 20 + 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/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 seven pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the wine data `wine.txt`. The data come from a study of Pinot Noir wine quality. The dataset contains 38 observations and 7 variables: `Quality`, `Clarity`, `Aroma`, `Body`, `Flavor`, `Oakiness`, and `Region`. We will take `Quality` as the response variable and the remaining variables (which represent features of the wine) as predictors. Be sure to treat `Region` as a qualitative predictor. The goal is to develop a model that relates the quality of Pinot Noir with its features. The model can potentially be used to predict the quality of the wine.

   (a) Perform an exploratory analysis of data. Comment on findings that interest you.

   (b) Do part (a) of Exercise 15 in Chapter 3 for these data.

   (c) Do part (b) of Exercise 15 in Chapter 3 for these data.

   (d) Based on your observation in (b) and (c), build a "reasonably good" multiple regression model for these data. Be sure to explore interactions of `Region` with other predictors. Carefully justify all the choices you make in building the model and verify the model assumptions.

   (e) Write the final model in equation form, being careful to handle the qualitative predictors and interactions (if any) properly.

   (f) Use the final model to predict the `Quality` of a wine from `Region` 1 with other predictors set equal to their sample means. Also provide a 95% prediction interval for the response and a 95% confidence interval for the mean response. Interpret the results.

2. Consider the diabetes dataset `diabetes.csv`. These data are from `https://www.kaggle.com/johndasilva/diabetes?select=diabetes.csv`. You can read more about the data, including a description of the variables, on this website. We will take `Outcome` as the response, the other variables as predictors, and all the data as training data.

(a) Perform an exploratory analysis of the data. Comment on findings that interest you.

(b) Perform an LDA of the data. Compute the confusion matrix, sensitivity, specificity, and overall misclassification rate based on 0.5 cutoff for the posterior probability. Plot the ROC curve. What do you observe?

(c) Repeat (a) using QDA.

(d) Compare the results from (a) and (b). Which classifier would you recommend? For the recommended classifier what posterior probability cutoff would you suggest? Justify your answer.

3. **Bonus problem (10 points)**

This problem gives you an idea about what will happen when the the number of parameters is a lot larger than the number of sample size. In particular, here we have p parameters and only 1 observation. In this case, the typical MLE estimator will fail because it gives a larger MSE. Recall the bias and variance trade-off pheonomenon. Here MLE is an unbiased estimator. The famous J-S estimator is a shrinkage estimator in this case, which gives smaller MSE compared to typical MLE.

In this question, we will compare the performance of James-Stein (JS) estimator and MLE. Suppose $Y \in \mathbb{R}^p \sim N(\mu, \sigma^2 I_p)$, where I_p is a p-dimensional identity matrix and $N(\mu, \sigma^2 I_p)$ indicates the multivariate normal distribution with mean vector μ and variance matrix $\sigma^2 I_p$. Suppose we have one observation $Y_1 \sim N(\mu, \sigma^2 I_p)$, then the James-Stein estimator formula

$$\hat{\mu}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|Y_1\|_2^2}\right) Y_1,$$

where $\|\cdot\|_2$ indicates the ℓ_2 norm of a vector. The MLE will just be the observation itself $\hat{\mu}_{MLE} = Y_1$.

(a) Take $p = 10$, $\sigma = 1$. $\mu = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]^T$. Generate one observation $Y_i \sim N(\mu, \sigma^2 I_p)$, compute the corresponding JS estimator $\hat{\mu}_{JS,i}$ and MLE $\hat{\mu}_{MLE,i}$. Keep doing this process for $N = 1000$ times and get $\hat{\mu}_{JS,i}$, $\hat{\mu}_{MLE,i}$, $i = 1, \ldots, N$. Compute the empirical bias and risk of these two estimators.

$$\hat{Bias}(\hat{\mu}) = \left\| \left(\frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_i \right) - \mu \right\|_2$$

$$\hat{Risk}(\hat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mu}_i - \mu\|_2^2 .$$

Comment on what you observe.

(b) Repeat (a) by replacing $\mu = a * [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]^T$ where $a = 1, \ldots, 10$. Plot the Risk of two estimators versus the value of a. What do you observe?

(c) Repeat (a) by replacing $\sigma = 0.1, 0.5, 2, 5, 10$. Plot the Risk of two estimators versus the value of σ. what do you observe?