DATA SCIENCE METHODS LAB 3: CROSS-VALIDATION AND BOOTSTRAPPING

Exercise 1

In this exercise, we will work with a dataset about home loan eligibility. A housing finance company provides home loans for the houses which were present across all urban, semi-urban and rural areas for their customers. The company validates the eligibility of loan after customer applies for the loan. However, it consumes lot of time for the manual validation of eligibility process. Our aim is to create a predictive model that will give us possible outcome of a loan application and fasten the procedure for the finance company. Since acceptance is a binary decision, we will use logistic regression in our predictions. After running the logit model, we will calculate how many misclassifications we have. Then we will use cross validation to calculate test error rate. In the last part of the exercise we will apply bootstrapping to calculate standard errors of coefficients. We have an analytical formula for standard errors and glm() function gives them automatically. We will compare standard errors from theory with the standard errors from bootstrapping.

1. Load the dataset "homeloan.csv" and summarize it. Eliminate rows with missing values. Convert type of categorical variables to factor.
   (Hint: use **factor()** command)

2. Estimate a logit regression using LoanAmount, Self_Employed, Education, Married, Gender, ApplicantIncome, Credit_History and Property_Area as independent variable and Loan_Status as dependent variable. Summarize estimation results.
   (Hint: use **glm(, family=binomial)** as command)

3. Based on the results of this estimation make predictions using 0.5 as threshold.
   (Hint: use **predict.glm(, type= "response")** to obtain predicted probability of acceptance.)

4. Write a for loop to implement LOOCV using following steps:

   - Create and empty array of size number of observations by 1.
   - Say for loop has an index of i, estimate model without i^{th} observation.
   - Predict outcome of i^{th} observation using estimation results.
   - Store predicted outcome in the array you created.
   - Calculate error rate with these new predictions.

5. Compare error rates in Q3 and Q4.

6. Write a for loop to implement bootsrapping using following steps:

- Create an empty array of size 1000 and number of regressors +1 to store estimation results.
- For each iteration in the for loop, draw a random sample from your data with replacement at the same size with your dataset.
- Run a logit regression on this new dataset and store its coefficients in the empty array you created.
- Find standard deviation of coefficients stored in the array.

7. Compare the standard errors from theory with the standard errors calculated by bootstrapping.