STAT 6340 (Statistical and Machine Learning, Spring 2021)

Mini Project 5

Instructions:

- Due date: April 19, 2021.
- Total points = 30
- 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 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 **five** pages. Also, only those output should be provided in this section that are referred to in the report.

This project involves `Hitters` dataset from the `ISLR` package in `R`. The dataset has already been used in the course. It consists of 20 variables measured on 263 major league baseball players (after removing those with missing data). `Salary` is the response variable and the remaining 19 are predictors. Some of the predictor variables are categorical with two classes. Use a dummy representation for them.

1. Consider an unsupervised problem with data only on the predictor variables, i.e., our data table has dimension 263×19. The goal is to perform a principal components analysis (PCA) of the data.

   (a) Do you think standardizing the variables before performing the analysis would be a good idea?

   (b) Regardless of your answer in (a), standardize the variables, and perform a PCA of the data. Summarize the results using appropriate tables and graphs. How many PCs would you recommend?

   (c) Focus on the first two PCs obtained in (b). Prepare a table showing correlations of the standardized quantitative variables with the two components. Also, display the scores on the two components and the loadings on them using a biplot. Interpret the results.

2. Consider again an unsupervised problem with the same data as in #1 but with the goal of clustering the players.

   (a) Do you think standardizing the variables before clustering would be a good idea?

   (b) Would you use metric-based or correlation-based distance to cluster the players?

(c) Regardless of your answers in (a) and (b), standardize the variables and hierarchically cluster the players using complete linkage and Euclidean distance. Display the results using a dendogram. You may have to do some preprocessing of the data to make the dendogram look nice. Cut the dendogram at a height that results in two distinct clusters. Summarize the cluster-specific means of the variables. Also, summarize the mean salaries of the players in the two clusters. Interpret the clusters.

(d) Use K-means with $K = 2$ to cluster the players on the basis of standardized variables and Euclidean distance. Summarize the cluster-specific means of the variables. Also, summarize the mean salaries of the players in the two clusters. Interpret the clusters.

(e) Compare conclusions from the two clustering algorithms. Which algorithm gives more sensible results? Explain.

3. Consider now a supervised problem — a linear regression model with log(`Salary`) as response (due to skewness in `Salary`) and the remaining 19 variables as predictors. Now our data table has dimension 263×20. All data will be taken as training data. For all the models below, use leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

(a) Fit a linear regression model. Compute the test MSE of the model.

(b) Fit a PCR model with M chosen optimally via LOOCV. Compute the test MSE of the model.

(c) Fit a PLS model with M chosen optimally via LOOCV. Compute the test MSE of the model.

(d) Fit a ridge regression with penalty parameter chosen optimally via LOOCV. Compute the test MSE of the model.

(e) Compare the four models. Which model(s) would you recommend? Justify.