STAT 6340 (Statistical and Machine Learning, Fall 2024)

Mini Project 5

Instructions:

- Due date: Nov 13, 2024 (Use Submission Link on eLearning to upload by 11:59 pm)
- Total points = 40
- Submit a typed report.
- It is OK to consult with others regarding this project or use any publicly available resources, but you must write code and answers on your own. Violation of this policy will be considered academic dishonesty, and you will referred to appropriate university authorities. If you don't do the work by yourself, how will you learn?
- 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: 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 software outputs should be provided in this section that are referred to in the report.
- You can code in `R` or `Python`.
- You are encouraged to use `R Markdown` or `Jupyter Notebook` to write your report.
- Bonus (4 points): Use *both* `R` and `Python` to code, compare the results, and provide an explanation if the results are not identical. This work should be described on a separate page titled Bonus Work with two sections — one to describe the comparison and the other for the code. The first section should be limited to one page.

---

1. (10 points) First, read carefully about *latent semantic analysis*—an application of PCA—from a PDF of the same name posted on eLearning. This is an excerpt from the book *Advanced Data Analysis from an Elementary Point of View* by C. R. Shalizi. The excerpt presents an analysis of a dataset consisting of 102 news stories from the *New York Times*. We will divide the dataset into a training and test dataset consisting of 80 and 22 news stories, respectively. These are available as `nyt.training.csv` and `nyt.test.csv`.

   (a) (2 points) Perform the analysis described in the excerpt but using only the training data. Your results won't match exactly because you are working with a subset of the data. However, your conclusions must be reported along the same lines and must include an analog of Figure 15.6 for the training data. Additionally, comment on whether the total number of PCs is consistent with what you expect.

   (b) (2 points) What does your analysis in (a) say about using only the first two PCs of a story to predict whether the story is about art or music?

(c) (2 points) Use the training data to fit a logistic regression model that predicts the class of a story using its scores on the first two PCs from (a) as predictors. Report estimate of the training error rate and superimpose the decision boundary in the figure from (a). Comment on the result.

(d) (2 points) Compute scores on the first two PCs of the stories in the test data and use them to get their class predicted by the model in (c). Compute test error rate and class-specific error rates. Comment on the results. [*Note*: You may use `predict` function to compute the PC scores for test data. If you compute them directly, be sure to take into account of the centering done in (a).]

(e) (2 points) The above analysis used only two PCs but we can work in general with M PCs. Right? How would you choose M? How does this approach compare with PCR?

2. (10 points) Consider the `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.

   First, we consider an unsupervised problem with data only on the *predictor variables*. The goal is to perform a principal components analysis (PCA) of the data.

   (a) (2 points) Do you think standardizing the variables before performing the analysis would be a good idea?

   (b) (4 points) 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) (4 points) Focus on the first two PCs obtained in (b). Compute correlations of the standardized quantitative variables with the two components. Display the results using an appropriate graph. Also, display the scores on the two components and the loadings on them using a biplot. Interpret the results.

3. (10 points) Consider again an unsupervised problem with the same `Hitters` data as in #2 but with the goal of clustering the players.

   (a) (1 point) Do you think standardizing the variables before clustering would be a good idea?

   (b) (1 point) Would you use metric-based or correlation-based distance to cluster the players?

   (c) (3 points) 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) (3 points) 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) (2 points) Compare conclusions from the two clustering algorithms, including an examination of the players in the two clusters. Which algorithm would you recommend? Explain.

4. (10 points) Consider now a supervised problem with the `Hitters` data from #2 — a regression model with log(`Salary`) as response (due to skewness in `Salary`) and the remaining 19 variables as predictors. 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) (2 points) Fit a linear regression model. Estimate the test MSE of the model.

(b) (2 points) Fit a PCR model with M chosen optimally via LOOCV. Estimate the test MSE of the model.

(c) (2 points) Fit a PLS model with M chosen optimally via LOOCV. Estimate the test MSE of the model.

(d) (2 points) Fit a ridge regression with penalty parameter chosen optimally via LOOCV. Estimate the test MSE of the model.

(e) (2 points) Compare the four models. Which model(s) would you recommend? Justify.