IEOR 4574: Business Analytics for OR
Spring 2016

Professor Adam Elmachtoub
Professor Jacob Leshno

Assignment 4: Unsupervised Learning and Extracting Signals

Due date: April 7, 11:59pm

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_2.pdf` and `Obama_Barack_2.R`. Your submissions must be based on your own original work. Late submissions will not be accepted.

1. Consider the `USArrests` data, which is part of the base `R` package.

   (a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

   (b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

   (c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. (Hint: use function `scale()` to standardize the data)

   (d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?

2. In this problem, we perform K-means clustering on the data set `HW4-Q2-data.csv`. The variable `GroupLabel` indicates the true class labels, make sure you do not use it in the clustering analysis.

   (a) Perform K-means clustering of the observations with $K = 3$. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

   (b) Perform K-means clustering with $K = 2$. Describe your results.

   (c) Now perform K-means clustering with $K = 4$, and describe your results.

   (d) Using the `scale()` function, perform K-means clustering with $K = 3$ on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in `(a)`? Explain.

3. Hillside is a small charter school in an inner city neighborhood of Newark New Jersey. Feeling pressure from its board to increase student scores on the state standardized test, the school administration recently piloted an intervention program called SIS (Student Intervention for Success) aimed at improving the scores of the lowest-performing students by providing tutoring. SIS provides an intensive help for students, including tutoring, an after-school study skills workshop and peer advising.

   The state test has a maximum score of 25. Students who receive a score of 11 or less are considered to be performing significantly under grade level. For its pilot, Hillside decided to enroll any student who had a 2011 score of 11 or lower in the SIS program at the start of 2012. The 2012 academic year was now over and Hillside administrators wanted to evaluate the results of SIS and report back to the board.

   The file `Hillside-data.csv` contains a sample of 100 Hillside students. Their performance on the past three standardized tests (2010, 2011 and 2012) are reported along with an indicator of whether the student was enrolled in SIS for 2012. Based on these data, answer the following questions:

   (a) Using 2011 as the "before" period and 2012 as the "after" period, perform a difference-in-difference analysis on the change in the average test scores of the SIS students. Based on your DiD estimate, what is the increase in test scores from SIS?

   (b) You suspect the results in part (a) may be overly optimistic because of the effects of regression to the mean. That is, because only the students who performed poorly on the 2011 exam were enrolled in SIS, some increase in their 2012 scores would be expected due simply to regression to the mean. To test this idea, consider the performance of the students between 2010 and 2011. Use the data from 2010 and the data from 2011 to determine whether there was regression to the mean. If so, what is the shrinkage coefficient?

   (c) Using the slope of the regression from part (b) as your shrinkage coefficient, construct a shrinkage estimate of 2012 scores based on the 2011 test results. What is the RMSE of your predictions?

   (d) Lastly to correct your DiD analysis for the shrinkage effect, compute the average of the estimated and actual 2012 scores for both the SIS students and non-SIS students. Considering the estimated 2012 scores as the "before" scores and the actual 2012 scores as the "after" scores, perform another DiD analysis of the SIS program. With this correction for shrinkage, what is your new estimate of the increase in test scores from SIS?

   (e) Briefly comment on why the method in part (a) and the method in part (d) produce different estimates of the effect of the SIS program.