ETM 58d Business Analytics
Homework 2, due April 15th, 2019

Instructions: Please solve the following exercises using R (http://www.r-project.org/). Also, you are requested to use certain R packages for this particular homework. You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress.

Do not share your code (except the one in your progress journals)! As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Tasks

Task 1

This task is to understand if we can obtain significant information regarding the game outcomes using the odd data from multiple bookmakers.

a) Select a least 5 bookmakers to check if over/under 2.5 game result can be explained by the odds for different types of bets. This can be achieved by training a classification model but suppose we would like to perform an analysis in an unsupervised way. For each game, you need to end up with a feature vector of the odds (i.e. home odd for bookmaker x, away odd for bookmaker x, tie odd for bookmaker x, over2.5 odd for bookmaker x, under 2.5 off for bookmaker x, both teams to score/YES for bookmaker x and etc.). Suppose we perform a principal component analysis (PCA) on this feature set to end up with a lower dimensional representation (i.e. 2D or 3D). Comment on your PCA results (i.e. eigenvectors, variance covered and etc.). Suppose we plot the new coordinates where the points are color-coded based on their under/over 2.5 status. Is there any interesting information? Comment on your findings.
b) Follow the similar strategy by applying multidimensional scaling (MDS). Since you are given a data matrix. You need to transform this information to distance matrix. Use Manhattan and Euclidean distance to perform the same task in part (a). What are your conclusions? What are the differences between the MDS on Euclidean distance and Manhattan distance?
c) Compare your PCA results with MDS.

Task 2

Build a logistic regression model to predict if a game ends as over or under. Comment on your results. Recall that you need to leave some data out for testing for fair evaluation of your approach (performance on the training data do not provide idea about the generalization performance of a classifier). You can make use of the starter codes provided in the link https://github.com/BU-IE-582/fall18-instructor. These codes are helpful in structuring your models for your homework and project. Note that it is an example code that builds a lasso logistic regression for match result (three class classification problem). So you need to modify this code to build a logistic regression (without penalties) and the target value is over/under instead of (home,tie,away).