Lab 6

Maria Barouti

Exercise 1 - KNN – Bayesian nonparametric K-nearest neighbor classification

a) Load the data `Auto.rda` and install the package `ISLR`.

b) Create a fuel consumption rating variable named `Economy` that will be treated as categorical based on the following info. For `mpg<=17` mark as `Heavy`. For `mpg>17 & mpg<=22.75` mark as `OK`. For `mpg>22.75 & mpg<=29` mark as `Eco`. For `mpg>29` mark as `Excellent`.

c) Prepare training and testing data for KNN.

d) Call `knn` and use your training and testing data to produce a confusion matrix as well as the classification rate for Y test.

e) Is there a better K? Check the classification rate for all K from 1 to 20. What do you observe?

Exercise 2 - Logistic Regression

a) Load the data `depression_data.csv`. You will see that the data contain many missing responses that are marked as `NA`. Exclude those from your dataset.

b) Fit the logistic regression model. How well does your model predict within the training data? Create a table of true and predicted responses, by classifying a student as having a depression if the probability ofthat exceeds 0.3. How many false positive and negative diagnosis do you get? What is the percentage for the correctly predicted cases? What is the training error rate? Among the students who are really depressed, what fraction are correctly diagnosed?

c) Training and Test data. As we know, prediction error within the training data may be misleading since all responses were known and used to develop our classification rule. To get fair estimate of the correct classification rate, let's (1) Split the data into training and test subsamples; (2) Develop the classification rule based on the training data; (3) Use to classify the test data; (4) Cross-tabulate our prediction with the true classification.

d) Receiver Operating Characteristic (ROC) Curve. Focus on the true positive rate and the false positive rate for different thresholds.
True positive rate = P(predict 1 | true 1) = power
False positive rate = P(predict 1 | true 0) = false alarm