DA5020 – Assignment 11

This assignment provides you with an opportunity to implement the kNN algorithm and identify suitable values of k. In this exercise you will build a k-nearest neighbor classifier to predict the onset of diabetes using the Pima Indians Diabetes Database. The dataset contains the following explanatory variables: number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function and age. The response variable is Outcome, which indicates whether or not the patient has diabetes.

Question 1 — (5 points)

Load the **diabetes** dataset "diabetes.csv", inspect the data and gather any relevant summary statistics.

Question 2 — (5 points)

Normalize the explanatory variables using min-max normalization.

Question 3 — (5 points)

Split the data into a training set and a test set i.e. perform an 80/20 split; 80% of the data should be designated as the training data and 20% as the test data.

Question 4 — (25 points)

Create a function called knn_predict(). The function should accept the following as input: the training set, the test set and the value of k. For example knn_predict(**train.data**, **test.data**, **k**).

- Implement the logic for the k-nn algorithm from scratch (without using any libraries). There is an example in the lecture series on Canvas. The goal of your k-nn algorithm is to predict the ***Outcome*** (i.e. whether or not the patient has diabetes) using the ***explanatory variables***.
- The function should return a list/vector of predictions for all observations in the test set.

Question 5 — (10 points)

Demonstrate that the knn_predict() function works and use it to make predictions for the **test set.** You can determine a suitable value of k for your demonstration. After which, analyze the results that were returned from the function using a confusion matrix. Explain the results. Note: refer to the 'Useful Resources' section for more information on building a confusion matrix in R.

Question 6 — (+5 bonus points)

Repeat question 5 and perform an experiment using different values of **k**. Ensure that you try at least 5 different values of **k** and display the confusion matrix from each attempt. Which value of k produced the most accurate predictions?

Useful Resources

- Pima Indians Diabetes Database
- Normalizing Data with R
- Confusion Matrix in R
- Model Evaluation Techniques for Classification models
- kNN Algorithm using R

Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.A11.FirstName.LastName.Rmd and your PDF/HTML DA5020.A11.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.

- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.

- Not submitting a knitted PDF or HTML will result in reduction of 30 points.

- Not submitting the .Rmd file (or both) will result in a score of 0.