DATA7703 - Machine Learning for Data Scientists
Assignment 1

Due: 4:00pm Friday 26th August, 2022

This assignment is worth 15% of the total marks for the course. Submit answers to following questions in a single pdf file. Your code should be well-formatted and in machine-readable format. Part marks may be given for any working shown. State any assumptions you need to make to answer a question.

Questions

1. The file provided (`Snakes.xlsx`) was obtained from:

   `https://datadryad.org/stash/dataset/doi:10.5061/dryad.14cr5345`

   It contains data relating to tiger snakes. In particular, Sheet 1 of the file contains recordings of the mass (BODY MASS) and length (SVL) of a sample of adult snakes. Note that there are some missing values in this data.

   (a) [2 marks] Perform linear regression, deleting any rows that contain missing values. State the equation for your linear regression model and the sum of squared error of the model.

   (b) [2 marks] Repeat (a) but instead use a quadratic regressor. Compare the error of this model with the model from (a).

   (c) [2 marks] Filling in missing values in data is known as *imputation*. A simple way of doing this is to calculate the sample mean of the values for a given column and substitute this for the missing values. Perform mean imputation on the adult snakes data, followed by linear regression. State the equation for your linear regression model and the sum of squared error of the model.

- **(d)** [2 marks] Produce a plot showing the (imputed) data and your models from (a), (b) and (c) above.

2. (6 marks) Read Section 5.8 of Alpaydin and then exercise 7 in 5.10. Using the approach described, fit a 2-D quadratic regression model to the dataset `reg2d.csv`, which has the inputs as the first two columns and the target function values in the third column. What are the coefficient values for your linear model? Submit a listing of your python code and the results/outputs it produces for this question.

3. The `palmerpenguins` dataset is described here:

   `https://allisonhorst.github.io/palmerpenguins/`

   On the course website you will find this data in `penguins_size.csv`. Consider the classification problem of trying to identify the species of penguin using two of the features, namely the Body Mass (g) and Flipper Length (mm). Submit a listing of your python code and the results/outputs it produces for this question.

   - **(a)** [1 mark] Carry out a strategy for handling any missing data.
   - **(b)** [2 marks] Produce a scatterplot of the data in the feature space, colouring points according to the class label (i.e. the penguin species). Add a legend to this plot.
   - **(c)** [1 mark] Split the data randomly into a training set (70%) and test set (30%).
   - **(d)** [2 marks] Apply the k-Nearest Neighbor algorithm to your data. Experiment with different k values and plot a graph of the test and training set errors as a function of k.
   - **(e)** [2 marks] Apply decision trees to your data. Report the training and test set errors that result. Display the final decision tree model.