COMP4702/COMP7703/DATA7703 - Machine Learning
Homework 1 - Introduction and Exploratory Data Analysis

Marcus Gallagher

Due: 12:00pm Tuesday 12th March, 2019

Core Questions

1. Find the (sample) average and (sample) standard deviation of the 'hw1q2.csv' dataset provided on the Course blackboard site (Correct to 4 decimal places).

2. Imagine we record the maximum temperature in Brisbane for the month of February, but we forget to make the recording on the 6th and the 16th (y_6 and y_{16}). We decide to predict the temperature on the missing days according to the following rule:

$$y_t = \frac{1}{2}(y_{t-1} + y_{t-2})$$

   (a) Is this performing classification or regression?

   (b) If the rule is used to predict y_{t+1}, is this performing extrapolation or interpolation?

3. Write a function, `sum_to_n()`, which takes an unordered array of unique integers and an integer, n, and returns all unique pairs which sum to n.

   Examples:

arr	n	output
[1, 2, 3, 4]	5	[1, 4; 2, 3]
[1, 4, 5, 3, 2]	6	[1, 5; 4, 2; 3, 3]
[1, 2, 5, 2, 6, 3]	7	[1, 6; 2, 5]

   Supply your code (Matlab or python) for this question. Important: you must write this code yourself!

4. Perform some exploratory data analysis on the `hw1mystery.csv` dataset, provided in this folder. The Data relates to a sport, it is your job to investigate further. Answer the following questions (no more than 2 sentences per question; if you don't know the answer, make a guess!):

   (a) Over how many years was the data gathered? (round to the nearest whole year)

   (b) What do you think attribute B represents? Why?

   (c) What do you think attribute N represents? Why?

   (d) What do you think attribute G represents? Why?

   (e) What do you think attribute K represents? Why?

   (f) What sport do you think this dataset has been taken from? Why?

Extension Question

5. Non-parametric statistics are commonly used in machine learning. They are useful to describe data that does not necessarily follow a known distribution. Find and read an explanation of a **box-whiskers plot**. Using the `sepal_length` feature only from the Full Iris dataset (150 data points), how many data points lie outside the inter-quartile range?