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

Marcus Gallagher

Core Questions

1. Find the (sample) average and (sample) standard deviation of the body mass of tiger snakes, based on the data available at:
   `https://datadryad.org/stash/dataset/doi:10.5061/dryad.14cr5345`
   (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 maximum 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 the maximum temperature on the 1st of March, 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]
[1, 2, 5, 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. Answer the following questions in **LESS THAN** 3 sentences. Use correct and specific statistical language:

   (a) State which features are categorical.

   (b) Which are the two most strongly correlated features? What is the numerical and/or statistical relationship between them?

(c) Discuss an interesting relationship that you observe between a pair of features that are not the ones from the previous question.

(d) Discuss an interesting feature in the dataset.

(e) The following points relate to features 19,24,27,29,30, and 31:

   i. Plot a correlation heatmap between the above features

   ii. Which two features have the lowest correlation? What is the value?

   iii. Discuss the statistical properties of the two features in ii.

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 EITHER[1] (a) the `sepal_width` feature from the Full Iris dataset (150 data points), or (b) the tiger snake data from Q1 above, what is the value of the data point that lies closest (but not exactly on) the boundary of the inter-quartile range?

6. Using the data from question 4.:

   (a) Find the Easter egg in the data. (Hint: think outside the box)

   (b) Can you guess what this is a dataset of?

[1] The option to use the tiger snake data was added because the Iris data is slightly annoying for this question! The tiger snake data could be annoying in a different way however...