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

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).

   Answer:
   Sample average/mean: 467.4876
   Sample standard deviation: 288.3465
   Other reasonable answers: using N as the normalisation in the std. dev. calculation gives 288.0880.

   Doing some reasonable imputation of missing data may give different numbers, which is fine as long as it is clearly stated. Substituting 0 for NaN (for example) would not be reasonable.

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? Answer: regression

   (b) If the rule is used to predict the maximum temperature on the 1st of March, is this performing extrapolation or interpolation? Answer: extrapolation

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! Answer: here is a possible function - note a linebreak on the innermost if statement.

```
function output = sum_to_n(arr, n)
output = [];
for i = 1:length(arr)
    for j = 1:length(arr)
        if(i == j)
            continue;
        end
        if ((arr(i) + arr(j)) == n)
            if(~isempty(output) && (ismember([arr(j), arr(i)], output, '
               rows') || ismember([arr(i), arr(j)], output, 'rows')))
                break;
            else
                output = [output; [arr(i), arr(j)]];
                break;
            end
        end
    end
end
```

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. Answer: 33,34 are clear. Let's not be too hard on people who argue that binary values could be representing categorical, or even integers.

   (b) Which are the two most strongly correlated features? What is the numerical and/or statistical relationship between them? Answer: The first and second columns have a difference of 0.5 for all values, so perfect correlation (1).

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

   Answer: Clearly there are many possible answers, here are a couple of samples:

   - Data column 37 and 38 (and 39) contain the same number of NaNs in their first 8 rows and last 9 rows. When you remove the NaNs you get 784 rows. These columns also range from 15 to 255 which is the bound of 8 bit numbers.
   - Columns 20 and 27 have a clear positive correlation and contain a large number of different values. Col. 20 is on a scale roughly an order of magnitude large than Col. 27, but seems to have few values less than 20. Outlying points are mainly due to unusually large values for Col.27.

   (d) Discuss an interesting feature in the dataset. Answer: Clearly there are many possible answers, here are a couple of samples:

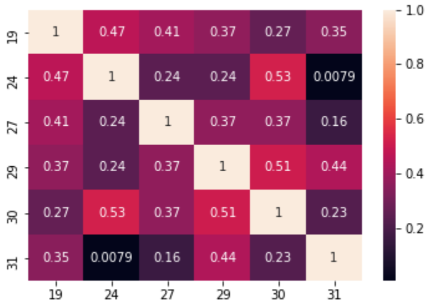
   - Data column 30 is positively skewed. It ranges from 20 to 230, but has a mean of 70.9 and a median of 66.0.

Stat	24	31
Mean	73.0	66.3
Median	70.0	65.0
Min	5.0	5.0
Max	230.0	180.0
Std Dev	30.8	20.9

- Column 27 has many values represented (not just a few discrete values), but has a long right tail. One data point (value around 14) is so far away from the rest that it could be reasonably suspected to be an outlier.

(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?
Answer: column 24 and 31. 0.0079

iii. Discuss the statistical properties of the two features in ii.
Answer: They are both positively skewed but 24 has larger skewness than 31. Some attributes are outlined in tab

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?

Answer:
(a) The quartiles are at 2.8 (25%) and 3.3 (75%). Unfortunately the data only has one decimal place of precision, so there are multiple points at a value 0.1 away - i.e. 2.7 and 3.4.
(b) 246

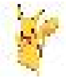
Some helpful Matlab to get these:

[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...

```
Q1 = quantile(sepallength, 0.25)
Q3 = quantile(sepallength, 0.75)
sort(sepallength) %and then inspect the array
```

6. Using the data from question 4.:

   (a) Find the Easter egg in the data. (Hint: think outside the box)
   Answer: As stated in Data column 37, 38 and 39 there are 3 rows that are suspiciously similar. And we all know many images are encoded as uint8s. The hint says box a box is a square and the square root of 784 is 28. So I assumed the rows were RGB values.

   (b) Can you guess what this is a dataset of?
   Answer: Based on the picture it's a pokemon dataset.