Workshop Week 7: COMP20008 2018

Part A

Part A exercises should be done by hand and then will be discussed on whiteboard.

Consider the following hypothetical dataset providing measurements for *Average Steps per day* and *Average Resting Heart Rate*, across a sample of 12 people.

Person ID	**Average Steps per day**	**Average Resting Heart Rate**
1	1000	100
2	2500	105
3	3000	80
4	5000	77
5	6000	74
6	9000	70
7	11000	65
8	14000	63
9	18000	62
10	19000	61
11	19500	60.5
12	22000	55

1. Compute the Pearson correlation between *Average Steps per day* and *Average Resting Heart Rate*. Show your working. How would you interpret this correlation value?

2. Based on the Pearson correlation value, can one conclude that doing more steps per day will cause one's average resting heart rate to decrease? How else might it be interpreted?

3. Discretise the data as follows: Apply 3 bin equal frequency discretisation to *Average Steps per day* and 4 bin equal frequency discretisation to *Average Resting Heart Rate*. Show the values of the discretised features.

4. Using the discretised features, compute the entropies: *H(Average Steps per day)*, *H(Average Resting Heart Rate)*, *H(Average steps per day | Average Resting Heart Rate)*, *H(Average Resting Heart Rate | Average Steps per day)*.

5. Using the above information, compute the mutual information between *Average Steps per day* and *Average Resting Heart Rate*.

Note: In the answers to this workshop, we will include some example Python code for computing mutual information and Pearson correlation, which you might choose to use in Phase 2 or 3 of the project.

Part B

Phase 3 of the project will require you to formulate a question, identify 2 open datasets to help answer this question, as well as conduct some initial wrangling to evaluate feasibility.

The following exercise illustrates a possible simple scenario and is designed to get you thinking about how you would approach Phase 2 under this simple scenario.

Suppose our question is *Are we building enough green spaces in Victoria to ensure a healthy population?*

- **Question 1:** Who would be interested in an answer to this question and why?
- Log in to the Aurin portal `https://aurin.org.au`
- Select Victoria as your region of interest.
- Add the dataset "2015 Local Government Area (LGA) Statistical Profiles". You should select all the attributes to include. This dataset includes information about number of people reporting high blood pressure across different regions in the State. We will use this as a measure of people's health.
- Add the dataset "LGA Visit to green space (once per week)". You should select all the attributes to include. This dataset contains information about number of people who visit local green space each week, across different regions in the State.
- Download each of these datasets as a CSV file.
- **Question 2:** What feature would you use to join these datasets together? How would you approach this in Python?
- **Question 3:** The following are examples of possible initial investigations you could perform for phase 3 on this dataset. Rank them in terms of priority. How many would you have time to do in 6-10 hours wrangling?
  - Scatter plot of blood pressure versus number of visits to green space, across different LGAs.
  - Boxplots for blood pressure feature and number of visits to green space feature.
  - Outlier detection for blood pressure feature and number of visits to green space feature.
  - Clustering
  - Missing value imputation
  - Correlation between blood pressure and number of visits to green space
  - ... anything else?
- **Question 4:** Suppose the correlation between blood pressure versus number of visits to green space turns out to be small. What other features could you examine from the "2015 Local Government Area (LGA) Statistical Profiles" dataset to serve as an indicator of health?
- **Question 5:** What challenges do you think might arise in studying this research question for Phase 3?