FIT2086 Assignment 1

Due Date: Wednesday, 15/8/2018

1 Introduction

There are total of six questions and $4 + 4 + 5 + 4 + 4 + 7 = 28$ marks in this assignment. Please note that working and/or justification must be shown for all questions that require it.

This assignment is worth a total of 10% of your final mark, subject to hurdles and any other matters (e.g., late penalties, special consideration, etc.) as specified in the FIT2086 Unit Guide or elsewhere in the FIT2086 Moodle site (including Faculty of I.T. and Monash University policies).

Students are reminded of the Academic Integrity Awareness Training Tutorial Activity and, in particular, of Monash University's policies on academic integrity. In submitting this assignment, you acknowledge your awareness of Monash University's policies on academic integrity and that work is done and submitted in accordance with these policies.

Submission: No files are to be submitted via e-mail. Correct files are to be submitted to Moodle. Scans of handwritten answers are acceptable but they **must** be clean and legible. Submission must occur before 11:55 PM Wednesday, 15th of August, and late submissions will incur penalties as per Faculty of I.T. policies.

2 Questions

1. In Lecture 1 we learned about several different types of general data science techniques: (i) classification, (ii) scoring (or regression), (iii) anomaly detection, (iv) clustering, (v) recommending systems and (vi) forecasting. For each of the following problems, suggest which of these methods is most appropriate and justify your selection:

   (a) Discovering hidden patterns in the behaviour of listeners on Spotify? **[1 mark]**

   (b) Predicting the life expectancy of a turtle from variables such as its diet and genetic makeup? **[1 mark]**

   (c) Determining whether the Apple share price will go up or down over the next month? **[1 mark]**

   (d) Trying to work out whether an image contains a picture of a human running or walking? **[1 mark]**

2. It is common to try and use data collected on consumers or users of websites to try and predict their interests. Imagine we are running a music streaming service, and have information on genres

	Doesn't Like Rock ($R = 0$)	Likes Rock ($R = 1$)
Doesn't Like Heavy Metal ($M = 0$)	0.7	0.1
Likes Heavy Metal ($M = 1$)	0.05	0.15

Table 1: Observed frequencies of "likes" and "don't likes" for two genres of music in a population of online music streamers.

that a user has previously liked or disliked. We could try to use this information to recommend songs from different genres to the user. Table 1 shows the frequency with which users on our service like and dislike two genres of music, Rock and Heavy Metal.

(a) What is the probability of a person in our population liking Heavy Metal, irrespective of whether they like Rock? **[1 mark]**

(b) What is the probability of liking Heavy Metal given that a person does not like Rock? **[1 mark]**

(c) What is the probability of liking Heavy Metal given a person does like Rock? **[1 mark]**

(d) Do you think liking Rock music is a good predictor of whether a person will like Heavy Metal? Why or why not? **[1 mark]**

3. Imagine that we roll two fair six-sided dice (i.e., all six sides have equal probability). Let X_1 and X_2 be the random variables representing these outcomes.

   Now, imagine we take one of the dice rolls, say X_1, and add a (possibly negative) constant c to the result. If this becomes less than zero, then we set it to zero; denote this by

   $$(X + c)_+ = \max(X + c, 0).$$

   This type of dice roll manipulation occurs frequently in many board and tabletop games.

   (a) What is the expected value of $(X_1 + 1)_+$? **[1 mark]**

   (b) What is the expected value of $(X_1 - 2)_+$? **[1 mark]**

   (c) What is the expected value of $(X_1 - 2)_+ \times (X_2 + 1)_+$? **[1 mark]**

   (d) What is the variance of $(X_1 - 2)_+$? **[1 mark]**

   (e) What is the probability that $(X_1 - 1)_+ > X_2$? **[1 mark]**

   You must show the working/reasoning as to how you obtained these answers

4. Imagine we receive a dataset from a colleague regarding a credit assessment for a bank, and we are asked to model the measurements. What distribution would be appropriate for the following variables (briefly justify your answer):

   (a) Relationship status of the applicant (single or partnered)? **[1 mark]**

   (b) Number of previous times defaulted on loan repayment? **[1 mark]**

   (c) Income in last financial year? **[1 mark]**

   (d) Number of dependents (children, spouse) of applicant? **[1 mark]**

5. Imagine that a continuous random variable X defined on the range $[0, b]$ follows the probability density function

$$p(X = x \mid b) = \begin{cases} \frac{2x}{b^2} & \text{for } x \in [0, b] \\ 0 & \text{everywhere else} \end{cases}.$$

Answer the following questions; you must include working if appropriate.

(a) Determine the expected value of X, i.e., $\mathbb{E}[X]$. **[1 mark]**

(b) Determine the cumulative distribution function for this distribution, i.e., $\mathbb{P}(X \leq x)$. **[1 mark]**

(c) What is the median value of this distribution? **[1 mark]**

(d) Plot the probability density function of X when $b = 2$. **[1 mark]**

6. A clothes store records the heights of a number of customers buying trousers in the previous week. The recorded heights (in metres) were

$$\mathbf{y} = (1.78, 1.65, 1.62, 1.84, 1.75, 1.85, 1.52, 1.55).$$

The stock manager decides to use this information to help determine how much future product to order. To do so, she decides to fit a normal distribution to this data and use it to model the population of future people buying trousers. You must show working/R code as required to obtain full marks.

(a) Fit a normal distribution to the data $\mathbf{y}$ using the maximum likelihood estimator for μ and σ. What are the values of these parameters for this data? **[2 marks]**

(b) Plug these estimates $\hat{\mu}$ and $\hat{\sigma}$ into the normal distribution, and use this to make predictions about future customers. Using this model, answer the following questions:

i. Imagine the store stocks pants suitable for people in the following four height ranges:

$$(< 1.5m),\ (1.5m - 1.65m),\ (1.65m - 1.8m),\ (> 1.8m).$$

What are the estimated proportions of people in the population of future customers that would fall into each of these height ranges? **[2 marks]**

ii. If a new customer walks into the store to buy pants, which height range are they most likely to be in? **[1 mark]**

iii. If the store receives 10 customers in one day, what is the probability that none of them will be $1.65m$ or taller? **[1 mark]**

iv. Imagine the store receives 160 customers per week buying trousers. During a week of sales how many pairs of pants for people between $1.65m$ and $1.8m$ would the store expect to sell? **[1 mark]**