FIT2086 Assignment 1

Due Date: Sunday, 30/08/2020, 11:55PM

1 Introduction

There are a total of five questions worth $4 + 6 + 5 + 7 + 8 = 30$ 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. You must ensure your submission contains answers to the questions in the order they appear in the assignment. Submission must occur before 11:55 PM Sunday, 30th 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/applications: (i) risk prediction, (ii) recommendation systems, (iii) forecasting, (iv) anomaly detection, (v) image recognition systems. For each of the following problems, suggest which of these application types the problem belongs to and justify your selection:

   (a) Discovering the hidden patterns in the behaviour of Netflix users? **[1 mark]**

   (b) Estimating the number of people at the platform of a train station from CCTV footage? **[1 mark]**

   (c) Predicting the number of new cases of COVID-19 over the coming month? **[1 mark]**

   (d) By using the recent usage habits/actions of the users of a mobile app, predicting whether or not these users will continue using the app in the near future? **[1 mark]**

2. An electro-cardiograph (ECG) is used for detecting the presence of a wide range of heart diseases such as heart valve disease, heart arrhythmias, etc. Despite this, an ECG does not provide

	$H=0$	$H=1$
$T=0$	?	?
$T=1$	?	?

Table 1: Empty table of the proportions of having ($H=1$)/not having ($H=0$) a heart attack given that an individual had normal ($T=0$)/abnormal ($T=1$) ECG result six months prior.

evidence for the presence of all diseases that can cause a heart attack. Even if a individual's ECG test does not show any sign of heart disease there is still a chance that the individual will suffer a heart attack at a later stage.

The file `heart.disease.csv` contains data on a number of patients. For each patient (row) the file contains the results of their ECG, indicating whether the ECG was normal ($T=0$) or abnormal ($T=1$), as well as whether the patient subsequently suffered a heart attack six months after the test ($H=1$) or not ($H=0$). Using this data please answer the following questions; you must provide working/justification.

(a) Using the data in `heart.disease.csv`, write R code to find the frequency with which a individual had/did not have a heart attack given that their ECG results were normal/abnormal. Using these frequencies, fill in the entries of Table 1 with the proportions of the times those events occured, i.e., estimates of the joint probabilities of a normal/abnormal ECG followed by a heart attack/no heart attack in six months time. **[2 marks]**

(b) Using these proportions, calculate the marginal probability of having a heart-attack irrespective of whether ECG results were normal or abnormal, i.e., $\mathbb{P}(H=1)$. **[1 mark]**

(c) What is the probability that a person will have a heart attack given that their ECG results were abnormal? **[1 mark]**

(d) What is the probability that a person will have a heart attack given that their ECG results were normal? **[1 mark]**

(e) Do you believe that the ECG test is a good predictor of possible heart attack? You must justify your answer. **[1 mark]**

3. Imagine that we roll a fair six-sided die (i.e., all six sides have equal probability) and toss a fair coin. Let X_1 and Y_1 be the independent random variables representing the outcomes of those events respectively (the outcome of the coin is represented numerically as head: $Y_1=1$, tail: $Y_1=0$.). Let $S=X_1+3Y_1$ be the sum of the outcome of the roll and three times the outcome of the coin toss. Please answer the following questions with appropriate working/justification.

(a) What is the variance of S, i.e., what is $\mathbb{V}[S]$? **[1 mark]**

(b) Determine the probability distribution of S, i.e., the probability that $S=\{1,\ldots,9\}$. **[1 mark]**

(c) What is the expected value of $\sqrt{S}$, i.e., what is $\mathbb{E}\left[\sqrt{S}\right]$? **[1 mark]**

(d) Calculate the approximate value of $\mathbb{E}\left[\sqrt{S}\right]$ using the Taylor-series procedure discussed in Lecture 2. **[1 mark]**

(e) Imagine that we roll a second fair six-sided die; call the outcome of this roll X_2. What is the expected value of $(X_1+3Y_1+X_2)^2$, i.e., what is $\mathbb{E}\left[(X_1+3Y_1+X_2)^2\right]$? **[1 mark]**

4. Imagine that a continuous random variable X defined on the range $(-s, s)$ follows the probability density function

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

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

(a) Plot the probability density function of X when $s = 1$ and $s = 2$ for $x \in (-3, 3)$. **[2 marks]**

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

(c) Determine the variance of X, i.e., $\mathbb{V}[X]$ *(hint: it will be a function of s).* **[2 marks]**

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

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

5. The file `heights.hk.csv` contains the records of the heights of 18 year old children based on a survey that was conducted in 1993 in Hong Kong[1]. Imagine that during the same year, an amateur football team in Hong Kong is planning to select players from their local school district and the coach wants to determine how likely it is that they can find players of different heights to fill various positions on the field. They decide to fit a normal distribution to this data and use it to model the height of 18 year old people in the city.

(a) Fit a normal distribution to the height data using the maximum likelihood estimator for μ and the unbiased estimator of variance for σ^2. What are the values of these parameters for this data? **[2 marks]**

(b) Plug these estimates $\hat{\mu}$ and $\hat{\sigma}^2$ into the normal distribution, and use these to answer the following questions posed by the coach:

i. Imagine that the coach groups the players according to following 4 height ranges:

$$(< 1.65m),\ (1.65m - 1.75m),\ (1.75m - 1.85m),\ (> 1.85m).$$

What are the proportions of candidate players that would fall into each of these height ranges as predicted by your Gaussian model? **[2 marks]**

ii. If a new candidate player registers for the selection trial, which height range are they most likely to be in? **[1 mark]**

iii. A football team consists of 18 players (11 on the field and 7 on the bench). The coach would like to have at least 4 players in their team that are taller than $1.80m$ for use in defence and attack. Assuming that the height of the players and their chances of being selected to the team are independent, what is the probability that the coach will have at least 4 players taller than $1.80m$ available in their team? **[1 mark]**

(c) Is the normal distribution an appropriate model for this height data? Plot the observed probabilities of the different heights (using a histogram) against the probability density predicted by your Gaussian (normal distribution) model. Justify whether or not you believe the distribution appears to be a good fit to this data. **[2 marks]**

[1] Data source: Maternal and Child Health Centres (MCHC)