FIT2086 Assignment 1

Due Date: Friday, 19/08/2022, 11:55PM

Introduction

There are a total of **five** questions worth $4 + 7 + 7 + 7 + 7 = 32$ 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 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: Please submit a **single** PDF file containing your answers via 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 Friday, 19th of August, and late submissions will incur penalties as per Faculty of I.T. policies.

Question 1 (4 marks)

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:

1. Predicting the amount of bananas that will purchased in the next week? **[1 mark]**
2. Determining whether someone's credit card has been compromised via purchasing history? **[1 mark]**
3. Discovering the genre preferences of Netflix users? **[1 mark]**
4. Using genomic and lifestyle factors to determine the chance of an individual contracting breast cancer in the next year? **[1 mark]**

	$R=0$	$R=1$	$R=2$
$H=0$	?	?	?
$H=1$	?	?	?

Table 1: Empty table of the joint proportions of FC Barcelona winning ($R=2$)/drawing ($R=1$)/losing ($R=0$) a football match when playing at home ($H=1$)/playing away ($H=0$).

Question 2 (7 marks)

It is common in many sporting leagues for teams to alternate playing games at their own venue (i.e., "at home") and at other team's home venues (i.e., "away"). It is usually assumed that teams will play better at home, when they have support from their own fans, than when they play away. Futbol Club Barcelona is a major footballing club in the Spanish Primera División; let us consider the home and away performance for this particular team. This is a (simple) example of sports analytics, an area of data science which is rapidly growing in importance over the last few years. The information regarding the total number of home ($H=1$) and away ($H=0$) wins ($R=2$), as well as home and away draws ($R=1$) and losses ($R=0$) for the 2021/2022 seasons is as follows:

- 12 home games won;
- 2 home games drawn;
- 5 home games lost;
- 9 away games won;
- 8 away games drawn;
- 2 away games lost.

Using this data please answer the following questions; you must provide working/justification.

1. Using the frequencies provided above fill in the entries of Table 1 with the proportions of the times those events occurred, i.e., estimates of the joint probabilities of a win/draw/loss for home/away games (up to 3 decimal places). **[1 mark]**
2. Using these proportions, calculate the marginal probability of Barcelona winning a game, regardless of whether it is played at home or away, i.e., $\mathbb{P}(R=2)$. **[1 mark]**
3. What is the probability that Barcelona will win a game, given that they are playing at home? **[1 mark]**
4. What is the probability that Barcelona will win a game, given that they are playing away? **[1 mark]**
5. Do you believe that Barcelona is more likely to win when at home versus when they play away? **[1 mark]**
6. Imagine that Barcelona will play an away game, then a home game, and then an away game in their next three games. What is the probability that they will not lose two out of three of these games? **[2 marks]**

Question 3 (7 marks)

Imagine that we roll a fair six-sided die and a fair four-sided die (i.e., all sides have the same probability). Let X_1 and Y_1 be the independent random variables representing the outcomes of those events respectively. Let $S = X_1 + 3\,Y_1$ be the sum of the outcome of the roll of the six-sided die and three times the outcome of the roll of the four-sided die. Please answer the following questions with appropriate working/justification.

1. What is the variance of S, i.e., what is $\mathbb{V}[S]$? **[1 mark]**
2. Determine the probability distribution of S, i.e., the probability that $S \in \{4, \ldots, 18\}$. **[1 mark]**
3. What is the expected value of $\sqrt{S}$, i.e., what is $\mathbb{E}\left[\sqrt{S}\right]$? **[1 mark]**
4. Calculate the approximate value of $\mathbb{E}\left[\sqrt{S}\right]$ using the Taylor-series procedure discussed in Lecture 2. **[2 marks]**
5. Imagine that we roll a second fair four-sided die; call the outcome of this roll Y_2. What is the expected value of $(X_1 + 3Y_1 - 2Y_2)^2$, i.e., what is $\mathbb{E}\left[(X_1 + 3Y_1 - 2Y_2)^2\right]$? **[2 marks]**

Question 4 (7 marks)

Imagine that a continuous random variable X defined on the range $[0, 1]$ follows the probability density function

$$p(X = x \,|\, a) = \begin{cases} (a+1)x^a & \text{for } x \in [0,1] \\ 0 & \text{everywhere else} \end{cases} .$$

where $a > 0$ is a parameter that controls the shape of the distribution. Answer the following questions; you must include appropriate working.

1. Plot the probability density function of X when $a = 1/2$ and $a = 2$ for $x \in [0, 1]$. **[2 marks]**
2. Determine the expected value of X, i.e., $\mathbb{E}[X]$. **[1 mark]**
3. Determine the expected value of $1/X$, i.e., $\mathbb{E}[1/X]$. **[1 mark]**
4. Determine the variance of X, i.e., $\mathbb{V}[X]$. **[1 marks]**
5. Determine the median of X. **[2 marks]**

(hint: the answers to Q4.2 through Q4.5 will all be functions of a).

Question 5 (7 marks)

The file `dogbites.1997.csv` contains the daily number of admissions to hospital of people being bitten by dogs for the second half of 1997 [1]. Answer the following questions; you must provide relevant R statements, working or justification as appropriate to obtain full marks.

1. Fit a Poisson distribution to the dog bites data using maximum likelihood. What is the estimated rate, $\hat{\lambda}$, of dog bite incidents in Australia during this period? [**1 marks**]

2. Plug the estimated $\hat{\lambda}$ into the Poisson distribution, and use this to make predictions about future dog bite incidences. Using this model, answer the following questions:

   (a) What is the probability of two or less admissions for a dog-bite in a day? [**1 mark**]

   (b) What are the two most likely number of dog-bite admissions to occur on any given day? [**1 mark**]

   (c) Over a week period (i.e., 7 days), what is the probability of seeing at most 32 dogbites? [**1 mark**]

   (d) What is the probability of seeing three or more dog-bite admissions for at least 12 days in a 14 day period? [**1 mark**]

3. The quality of predictions of a model are only as good as the model is itself representative of the population. Do you believe that the Poisson distribution is an appropriate model for the dog bite data? Plot the observed probabilities of the different number of daily dog-bite incidences against the probabilities for the number of daily dog-bites as predicted by your Poisson model (over the range of 0 to 22 dog-bite admissions), and use this to justify whether or not this model is a good fit to the data. [**2 marks**]

[1] Data source is taken from the Australian Institute of Health and Welfare Database of Australian Hospital Statistics.