FIT2086 Assignment 1

Due Date: Monday, 21/08/2023, 11:55PM

Introduction

There are a total of **five** questions worth $4 + 6 + 8 + 7 + 7 = 32$ marks in this assignment. Please note that working and/or justification must be shown for all questions to obtain full marks.

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. You are not permitted to use generative AI (such as, but not limited to, ChatGPT) in any capacity when answering this assignment.

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 Monday, 21st 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, (vi) generative AI systems. For each of the following problems, suggest which of these application types the problem belongs to and justify your selection:

1. Creating an image of a gorilla riding a motorcyle? **[1 mark]**
2. Discovering the shopping preferences of customers at Coles? **[1 mark]**
3. Predicting the amount of panadol that will purchased in the next week? **[1 mark]**
4. Using genomic and lifestyle factors to determine the chance of an individual contracting diabetes in the next two years? **[1 mark]**

	$H = 0$			$H = 1$	
	$W = 0$	$W = 1$		$W = 0$	$W = 1$
$P = 0$	0.176	0.235	$P = 0$	0.117	0.058
$P = 1$	0.060	0.117	$P = 1$	0.059	0.178

Table 1: Table of the joint probabilities of a football team winning ($W = 1$) or losing ($W = 0$) when playing at home ($H = 1$) or away ($H = 0$) and whether they won ($P = 1$) or lost ($P = 0$) their previous game.

Question 2 (6 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. It is also commonly assumed that teams gain confidence from wins, and lose confidence from losses, so that they are more likely to win after a previous win and lose after a previous loss. Imagine we are working as a data analyst for a football team, and have collected data from the performance of the 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.

Imagine we have obtained information on a large number of previous games the team has played, along with information on whether they were played at home or not. Specifically, we have: whether the team won ($W = 1$) or lost ($W = 0$) their game; whether the game was played at home ($H = 1$) or away ($H = 0$); and whether the previous game that they played was a win ($P = 1$) or a loss ($P = 0$). Using this data we have calculated the joint probabilities $\mathbb{P}(W, H, P)$ of these three events. These are shown in Table 1. Please use these probabilities to answer the following questions with appropriate working/justification/explanation:

1. What is the marginal probability of the team winning a game, i.e., what is $\mathbb{P}(W = 1)$? **[1 mark]**

2. What is the probability that the team will win a game, given that they lost their previous game, regardless of whether they are playing home or away? **[1 mark]**

3. What is the probability that the team will win a game, given that they won their previous game, regardless of whether they are playing home or away? **[1 mark]**

4. Do you believe that the team is more likely to win after winning their previous game than if they lost their previous game? **[1 mark]**

5. Imagine that the team has won their previous game. They will next play at home and then away. What is the probability that they will go on to win one or more of the next two games? **[2 marks]**

Question 3 (8 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 $S = 2\,X_1 - Y_1$ be two times the outcome of the roll of the six-sided die minus the outcome of the roll of the four-sided die. Please answer the following questions with appropriate working/justification.

1. What is the expected value of S, i.e., what is $\mathbb{E}\left[S\right]$? **[1 mark]**
2. What is the variance of S, i.e., what is $\mathbb{V}\left[S\right]$? **[1 mark]**
3. Determine the probability distribution of S, i.e., the probability that $S \in \{-2, \ldots, 11\}$. **[1 mark]**
4. What is the expected value of S^3, i.e., what is $\mathbb{E}\left[S^3\right]$? **[1 mark]**
5. Calculate the approximate value of $\mathbb{E}\left[S^3\right]$ using the Taylor-series procedure discussed in Lecture 2. **[2 marks]**
6. Imagine that we roll a second fair four-sided die; call the outcome of this roll Y_2. What is the expected value of $(2\,X_1 - Y_1 + 2\,Y_2)^2$, i.e., what is $\mathbb{E}\left[(2\,X_1 - Y_1 + 2\,Y_2)^2\right]$? **[2 marks]**

Question 4 (7 marks)

Imagine that a continuous random variable X defined on the range $(0, s)$ follows the probability density function

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

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

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

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

Question 5 (7 marks)

The file `covid.2023.csv` contains the number of days until recovery for patients with COVID-19 recorded from January through to May, 2020, before COVID vaccines were available. This data is useful as it gives an indication into how long it may take for an individual to recover if infected with a novel virus in the future, and can be used to estimate the load of hospitals and health care systems. Answer the following questions; you must provide relevant working/code or justification as appropriate to obtain full marks.

1. Fit a Poisson distribution to the COVID recovery data using maximum likelihood. What is the estimated value of this parameter for this data? **[1 mark]**

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

   (a) What is the probability of a patient recovering in 10 or less days? **[1 mark]**

   (b) What are the three most likely number of days it takes a patient to recover? **[1 mark]**

   (c) Imagine that five individuals have COVID. What is the probability that these five individuals take a combined total of between 60 to 80 days (inclusive) between them to recover? **[1 mark]**

   (d) What is the probability that three or more of these five patients will recover on or after day 14? **[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 COVID recovery data? Create a plot with the number of days to recovery (over the range 0 to 40 days) on the x-axis, that shows (on the y-axis): (i) the proportions of that many days-to-recovery in the data set; and (ii) the probabilities for the number of days-to-recovery as predicted by your Poisson model. Use this to justify/argue whether or not you believe this model is a good fit to the data. **[2 marks]**