FIT2086 Assignment 1

Due Date: Wednesday, 28/8/2019

1 Introduction

There are total of five questions and $4+7+6+7+7=30$ 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 Wednesday, 28th 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 energy usage of the customers of an electricity company? **[1 mark]**

   (b) Estimating whether or not a person will default on a loan ? **[1 mark]**

   (c) Predicting the amount of rainfall over the coming month? **[1 mark]**

   (d) Finding sub-types of cancers from tumour characteristics? **[1 mark]**

2. Sports analytics (i.e., the application of data science techniques to competitive sports) is a rapidly growing area of data science. In this question we will look at some very basic analytics applied to the outcomes of consecutive games of Australian Rules Football (AFL). The file

	$W_t = 0$	$W_t = 1$
$W_{t-1} = 0$	?	?
$W_{t-1} = 1$	?	?

Table 1: Empty table of proportion of wins/losses for Port Adelaide (W_t) given whether they won/loss their previous game (W_{t-1}).

`port.adelaide.csv` contains a record of the outcomes of games of AFL played by the Port Adelaide (PA) football club in the seasons 1998, 1999, 2000 and the first two rounds of 2001. The data is sequential, in the sense that each recorded binary variable records a win (1) or a loss (0) in the order in which the games were played.

A simple question regarding this type of data might be regarding the existence of (de)motivativing effects on a team if they have won/lost their previous game. Let W_t denote the outcome a game and W_{t-1} denote the outcome of the game played in the previous round. Answer the following questions; you must provide working/justification.

(a) Using the data in `port.adelaide.csv`, write some R code to find the frequency with which PA won/lost a game after it won/lost its previous game. Using these frequencies, fill in the entries of Table 1 with the proportions of times these events occured, i.e., estimates of the joint probabilities of a loss being followed by a loss, a loss being followed by a win, a win being followed by a loss and a win being followed by a win. **[2 marks]**

(b) Using these proportions, calculate the marginal probability of PA winning a game irrespective of whether they won or lost their previous game, i.e., $\mathbb{P}(W_t = 1)$. **[1 mark]**

(c) What is the probability that PA will win a game given that they won their previous game? **[1 mark]**

(d) What is the probability that PA will win a game given that they lost their previous game? **[1 mark]**

(e) Do you think winning/losing the previous game had an effect on the PA players in their next game? That is, do you think the events W_{t-1} and W_t are independent or not? Justify your answer. **[1 mark]**

(f) Calculate the probability of PA losing their next two games given that they won their previous game. **[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 independent random variables representing these outcomes. Let $S = X_1 + X_2$ be the sum of the two rolls. 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 $S = \{2, \ldots, 12\}$. **[2 marks]**

(c) Write a simple one-line formula describing this probability distribution. **[1 mark]**

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

(e) Imagine we roll a third dice, X_3. What is the expected value of $(X_1 + X_2 + X_3)^2$, i.e., what is $\mathbb{E}\left[(X_1 + X_2 + X_3)^2\right]$? **[1 mark]**

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

$$p(X = x \mid s) = \begin{cases} \frac{3}{4s}\left(1 - \left(\frac{x}{s}\right)^2\right) & \text{for } x \in (-s, s) \\ 0 & \text{everywhere else} \end{cases} .$$

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

(a) Plot the probability density function of X when $s = 1$ and $s = 2$. **[2 marks]**

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

(c) Determine the variance of X, i.e., $\mathbb{V}[X]$ (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 `dogbites.total.csv` contains the daily number of admissions to hospital of people being bitten by dogs from 13th of June, 1997 through to 30th of June, 1998[1]. Answer the following questions; you must provide relevant R statements, working or justification as appropriate to obtain full marks.

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

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

i. What is the probability of at most one admission for a dog-bite in a day? **[1 mark]**

ii. What number of dog-bite admissions is most likely to occur on a given day? **[1 mark]**

iii. Over a four week period (i.e., 28 days), how many dog-bite admissions would the hospital system expect to see? **[1 mark]**

iv. What is the probability of seeing six or more dog-bite admissions for at least 8 of the days in a 28 day period? **[1 mark]**

(c) Is the Poisson distribution an appropriate model for the dog bites data? Plot the observed probabilities of the different number of daily dog-bites against the frequencies predicted by your Poisson model. If you believe the distribution is not a good fit, justify why and discuss possible reasons it might not be an appropriate model for this phenomena? **[2 marks]**

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