FIT2086 Assignment 1

Due Date: Friday, 18/8/2017

1 Introduction

There are total of six questions and $4 + 4 + 5 + 4 + 2 + 5 = 24$ 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 both Moodle and Turnitin. Please read these submission instructions carefully and take care to submit the correct files in the correct places. Submission must occur before 11:55 PM Friday, 18th of August, and late submissions will incur penalties as per Faculty of I.T. policies.

2 Questions

1. Please award 1/2 mark if the answer is not justified with a very brief bit of text similar to the below.
   (a) **A**: Regression – the outcome is continuous.
   (b) **A**: Either clustering to discover patterns, or recommendation systems.
   (c) **A**: Classification – we are guessing a categorical outcome.
   (d) **A**: Anomaly detection as we are analysing repeated patterns to see if something different occurs.
2. Please use common-sense when marking this answer in regards to rounding.
   (a) What is the probability of a person in our population having heart disease, irrespective of their genetic mutation status? **[1 mark]**

A: Marginal probability $P(H = 1)$; using the sum rule we have

$$\begin{aligned} P(H = 1) &= P(H = 1, M = 0) + P(H = 1, M = 1) \\ &= 0.3 + 0.25 = 0.55 \end{aligned}$$

(b) What is the probability of contracting heart disease given a person has no mutation? **[1 mark]**

A: Conditional probability rule:

$$\begin{aligned} P(H = 1 \,|\, M = 0) &= P(H = 1, M = 0)/P(M = 0) \\ &= P(H = 1, M = 0)/(P(H = 0, M = 0) + P(H = 1, M = 0)) \\ &= 0.3/(0.35 + 0.3) = 0.4615 \end{aligned}$$

(c) What is the probability of contracting heart disease given a person has a mutation? **[1 mark]**

A: Conditiona probability rule:

$$\begin{aligned} P(H = 1 \,|\, M = 1) &= P(H = 1, M = 1)/P(M = 1) \\ &= P(H = 1, M = 1)/(P(H = 0, M = 1) + P(H = 1, M = 1)) \\ &= 0.25/(0.1 + 0.25) = 0.7143 \end{aligned}$$

(d) Do you think the LDLR genetic mutation is a good predictor of heart disease? Why or why not? **[1 mark]**

A: The LDLR genetic mutation is a decent predictor of heart disease as it increases likelihood of heart disease by 1.5 times. However, some students might identify this increase and believe it is not particular big or significant, so an answer that says it is OK but not a good predictor should still be acceptable. It is not a bad predictor, however.

3. Sufficient explanation of the student's answer is required; it is insufficient to simply give a number as an answer.

(a) What is the variance of a single dice roll X_1? **[1 mark]**

A: Answer 1: Variance is equal to $\mathbb{E}\left[(X_1 - \mathbb{E}\left[X_1\right])^2\right]$ so noting that

$$\mathbb{E}\left[X_1\right] = (1/6)(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

and then

$$\mathbb{E}\left[(X_1 - \mathbb{E}\left[X_1\right])^2\right] = (1/6)((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2) = 2.9167$$

Answer 2: Variance is equal to $\mathbb{E}\left[X_1^2\right] - \mathbb{E}\left[X_1\right]^2$; noting that

$$\mathbb{E}\left[X_1^2\right] = (1/6)(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = 15.1667$$

we have $\mathbb{V}\left[X_1\right] = 15.1667 - 3.5^2 = 2.9167$.

(b) What is the probability of seeing $X_1 = 2$ and $X_2 = 5$? **[1 mark]**

A: The two dice rolls are independent so

$$P(X_1 = 2, X_2 = 5) = P(X_1 = 2)\, P(X_2 = 5) = (1/6) \cdot (1/6) = 1/36.$$

(c) Now we add the two values of the two dice rolls together to get the new random variable $S = X_1 + X_2$. What is the variance is of S? **[1 mark]**

A: Again, the two dice rolls are independent so we use

$$\mathbb{V}\left[X_1 + X_2\right] = \mathbb{V}\left[X_1\right] + \mathbb{V}\left[X_2\right] = 2.9167 + 2.9167 = 5.8334$$

(d) What is the probability that $S = 11$? **[1 mark]**

A: Note that $S = 11$ only for $(X_1 = 5, X_2 = 6)$ and $(X_1 = 6, X_2 = 5)$. Each of these has probability $1/36$ so $P(S = 11) = 2/36$.

(e) What is the probability that $S \leq 4$? **[1 mark]**
A: The combinations of X_1, X_2 that satisfies $S \leq 4$ are:

$$\begin{aligned}
&(X_1 = 1, X_2 = 1),\\
&(X_1 = 1, X_2 = 2),\\
&(X_1 = 1, X_2 = 3),\\
&(X_1 = 2, X_2 = 1),\\
&(X_1 = 2, X_2 = 2),\\
&(X_1 = 3, X_1 = 1).
\end{aligned}$$

i.e., there are 6 combinations so the probability is $6 \cdot (1/36) = 6/36 = 1/6$.

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

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

(a) Sex of the patient (male or female)? **[1 mark]**

A: The data is binary, so the Bernoulli distribution, or potentially the binomial is a good fit.

(b) Age of the patient? **[1 mark]**

A: The age of the patient is a continuous variable, or a continuous variable rounded to whole numbers. The normal distribution is appropriate.

(c) The patient's weight at their initial visit? **[1 mark]**

A: A patient's weight is a continuous variable and can be modelled by a normal distribution.

(d) Number of children that the patient has? **[1 mark]**

A: The number of children is a discrete, numeric variable with overall small values (likely less than 10) that are counts of events (i.e., child births). A Poisson distribution is appropriate. A binomial is not appropriate as we are not counting successes out of a finite number of trials.

5. Please award one mark each for correctly solving for θ and n. Sufficient working out to show the student approached the problem correctly is necessary.

A: To answer the question, first recall the formulas for theoretical mean and variance of a Binomial distribution with success probability θ and number of trials n:

$$\mathbb{E}[M] = n\theta, \;\; \mathbb{V}[M] = n\theta(1-\theta)$$

Given the mean and variance as in our question, we have a simultaneous equation to solve. One solution is as follows: divide $\mathbb{V}[M]$ by $\mathbb{E}[M]$ to get

$$(1-\theta) = 4/6$$

which we can solve to obtain $\theta = 1 - 4/6 = 1/3$. Plugging this into the formula for $\mathbb{V}[M]$ we get

$$n(1/3)(1-1/3) = 4$$

which we can solve to obtain $n = 4/(1/3)/(1-1/3) = 4/(2/9) = 18$.

6. A lecturer records the number of student attendances at her weekly consultation period for the first eight weeks of semester. She believes that the rate of attendances does not vary from week to week, and therefore a Poisson model may be justified. The recorded number of attendances were:

$$\mathbf{y} = (2, 1, 2, 4, 3, 4, 1, 2)$$

(a) Fit a Poisson distribution to the data $\mathbf{y}$ using the maximum likelihood estimator for λ. What is the value of the estimate $\hat{\lambda}$ for this data? **[1 mark]**

A: The estimate is

$$\hat{\lambda} = \frac{2+1+2+4+3+4+1+2}{8} = 2.375$$

(b) Plug this estimate of $\hat{\lambda}$ into the Poisson distribution, and use this to make predictions about future consultation periods (provide any working as well as the line of R code to calculate probabilities in addition to the values of the estimated probabilites).

i. What is the probability of seeing at most one student during a consultation period? **[1 mark]**

A: This is $P(X = 0 \,|\, \lambda = 2.375) + P(X = 1 \,|\, \lambda = 2.375)$

$$\texttt{dpois}(\texttt{x} = 0, \texttt{lambda} = 2.375) + \texttt{dpois}(\texttt{x} = 1, \texttt{lambda} = 2.375) \approx 0.3139$$

ii. What is the probability of seeing three or more students during a consultation in two consecutive weeks? **[1 mark]**

A: Unfortunately this question is a bit ambigious on second reading. The original intent was to ask "what is the probability that we will see 3 or more students in consultation for two consecutive weeks", i.e., 3 or more one week and 3 or more the next week. To answer this, note that the probability that we will see 3 or more students in one week and in the next week, i.e., $P(X_i \geq 3 \,|\, \lambda = 2.375) \cdot P(X_{i+1} \geq 3 \,|\, \lambda = 2.375) = (1 - P(X \leq 2 \,|\, \lambda = 2.375))^2$

$$(1 - \texttt{ppois}(\texttt{q} = 2, \texttt{lambda} = 2.375))^2 \approx 0.1795$$

However, a second interpretation could be "what is the probability of seeing a total of 3 or more students over a two-week period", in which case we note that $X_i + X_{i+1} \sim \text{Poi}(2 \times 2.375)$, and we have

$$1 - \texttt{ppois}(2, 2.375 * 2) \approx 0.8526$$

Given the ambiguity, I would give marks for either interpretation, and a half-mark if they only manage to calculate $P(X_i \geq 3 \,|\, \lambda = 2.375)$, i.e., probability of seeing 3 or more in a single week.

iii. What number of students is the lecturer most likely to see in consultation during a single consultation period? **[1 mark]**

A: This is *not* the expected value (as many students will think); it is the value of X that is most likely under our fitted model, which is $X = 2$, i.e.,

$$\arg\max_x \{P(X = x \,|\, \lambda = 2.375)\} = 2.$$

So the most likely number of students the lecturer will see is 2.

iv. Over the entire semester of 12 weeks, how many students in total would the lecturer expect to see? **[1 mark]**

A: The expected value of $\mathbb{E}[X] = 2.375$ under our fitted model, so

$$\mathbb{E}\left[\sum_{i=1}^{12} X_i\right] = 12\,\mathbb{E}[X_i] = 28.5.$$

Note: don't remove any remarks if the students round the last quantity to 28 or 29; give a half-mark if the student rounds $\mathbb{E}[X] = 2.375$ to 2 before multiplying by 12.