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. In Lecture 1 we learned about several different types of 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:

   (a) Predicting a person's body-mass index given their blood-pressure, age and sex? [**1 mark**]

   (b) Discovering hidden patterns in behaviour of shoppers on ebay? [**1 mark**]

   (c) Using a person's genetic information to predict whether they will contract breast cancer or not? [**1 mark**]

   (d) Determining whether an elderly person has had an accident based on information regarding their daily patterns assessed by sensors around their house? [**1 mark**]

2. It is very common to try and predict a persons pre-disposition to a disease based on whether or not they have particular genetic mutations. Imagine that we have are trying to determine if a person is more or less likely to contract heart disease if they have a mutation at the LDLR gene

	No Heart Disease ($H = 0$)	Heart Disease ($H = 1$)
No Mutation ($M = 0$)	0.35	0.30
Mutation ($M = 1$)	0.10	0.25

Table 1: Population joint probabilities of heart disease/LDLR mutation.

(that makes low-density lipoproteins). Table 1 shows the joint probabilities for heart disease status/mutation status in a high-risk population. Please show your working as required.

(a) What is the probability of a person in our population having heart disease, irrespective of their genetic mutation status? [**1 mark**]

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

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

(d) Do you think the LDLR genetic mutation is a good predictor of heart disease? Why or why not? [**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 random variables representing these outcomes.

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

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

   (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**]

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

   (e) What is the probability that $S \leq 4$? [**1 mark**]

   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**]

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

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

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

5. Let $M \sim \text{Bin}(\theta, n)$ be a binomial variable with

$$\mathbb{E}[M] = 6, \;\; \mathbb{V}[M] = 4$$

   What are the values of the θ and n parameters for this binomial distribution? Show you working out. [**2 marks**]

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 **y** using the maximum likelihood estimator for λ. What is the value of the estimate $\hat{\lambda}$ for this data? **[1 mark]**

   (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]**

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

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

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