FIT5197 2018 S1 Assignment 1

Due date: Friday, Week 6, April 13th 2018

Details

Your solution should be submitted as a single R markdown file via the submission page on "My Assessments" by the deadline at the page. The file needs to run so we can convert it to PDF and view your answers, as well as read your code. Add in metadata at the top of the R markdown file to give your name and student ID and the submission date, as follows:

```
title: "FIT5197 2018 S1 Assignment 1"
author: "Your Full Name, 202343254"
date: "13 April 2018"
```

You can add an additional single PDF file with supplementary material to support any analytic justifications, however all answers should appear in the markdown file. The PDF will be used to assess partial marks in the event that the analysis is wrong so your working needs to be inspected.

This assignment is worth 15% of your final mark for the unit. Each question is worth 20 marks so there is 100 marks in total. It is subject to the standard guidelines regarding academic integrity, late penalties, and special consideration. Some assignments may be selected and the authors interviewed in person to further assess their knowledge of their written submission.

Marking criteria:

- correctness of solution (main)
- readability of explanations (important for partial marks)
- justification of code if interviewed
- clarity of output
- quality of report

Question 1: calculate conditional probability of an event

Tossing a fair die 7 times. Consider the following events, A = "each value appears at least once" and B = "the outcome is alternate in numbers (i.e., no two values are adjacent)". What is $p(A|B)$? Solve this analytically (a formula) and experimentally by simulation, in each case printing the value found. For example the following illustrate the condition $A\&B$:

- 1,3,5,4,3,2,6: +ve, all 1-6 occur and 3s are separated
- 1,3,3,5,4,2,6: -ve, the two 3s are adjacent
- 1,3,5,4,3,2,1: -ve, 6 does not occur
- 1,3,5,2,4,6,6: -ve, the two 6s are adjacent

Question 2: entropy

Given the Boolean dataset "FIT5197_2018_S1_Assignment1_Q2_data.csv" which contains 2 discrete random variables X and Y, and 100 samples with missing values (NA). Do the following:

1. handle NAs by mode imputation, and plot individual variables in a histogram with proper axis labels and title.
2. calculate and report full tables for $p(X)$, $p(Y)$, $p(X,Y)$, $p(X|Y)$, $p(Y|X)$.
3. calculate and report single values for $H(X)$, $H(Y)$, $H(X|Y)$ and $H(Y|X)$

Each sub-question is equal marks.

Question 3: correlations and covariance

Assume X and Y are two independent standard Gaussian random variables, $U = X - Y$ and $V = 2X + 3Y$. What is the correlation and covariance between U and V? Solve the questions analytically, then confirm your analytical results using $1,000,000$ simulation.

Question 4: maximum likelihood estimation of parameters

You are told the following data

$$[4, 3, 2, 4, 6, 3, 4, 0, 5, 6, 4, 4, 4, 5, 3, 3, 4, 5, 4, 5]$$

comes from a Poisson distribution with unknown parameter λ. Solve the MLE of a Poisson parameter analytically and numerically without using existing mle functions in R. Hint: you should first define a log likelihood function. Then use the `optimize()` function in R to find the optimal parameters for the given dataset.

Question 5: central limit theorem

Consider sampling a sequence of 10 i.i.d. random variables from a Poisson distribution with $\lambda = 10$. For this, experimentally justify the central limit theorem using simulation with sample size 100, 1000, 10000 and 100000. You should compute both theoretical and sample mean and sd. In addition, plot each result in a histogram with the theoretical Gaussian curve.

In the first instance (sample size 100), this means you generate 100 such samples of size 10, and compute their means to get 100 means. Then plot the 100 means in a histogram and report the sample mean and sample standard deviation of the 100 values. Compare these with the theoretical values. Then repeat for the other sizes.