THE UNIVERSITY OF MELBOURNE

Student number	

Semester 2 Practical assignment, 2022

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Submission deadline: 12:00 pm Friday 21 October

This assignment consists of 3 pages (including this page) with 3 questions and 40 total marks

Instructions to Students

Writing

- This assignment is worth 10% of your total mark.
- The deadline is **12:00 pm Friday 21 October (AEDT)**. No late submission is allowed. The total score is 40.
- Give answers to 5 decimal places.
- Submit your answer and R code through the LMS.
- Write your answers to the questions in one file. Submit you R code in a separated file.

Question 1 (10 marks)

Given a starting point $x_1 \in [0,1]$, a number $r \in [0,4]$ and a number $n = 2,3,\cdots$ of iterations, define the sequence $x_1,\cdots,x_n$ by $x_j = rx_{j-1}(1-x_{j-1}), j = 2,\cdots,n$.

(a) Write a function in R with arguments x_1, r, n and i that

(i) computes $x_1,\cdots,x_n$

(ii) prints the last i values $x_{n-i+1},\cdots x_n$

(iii) produces a plot of $x_1,\cdots,x_n$ versus $1,\cdots,n$.

The plot should have the horizontal axis labelled `j` and the vertical axis labelled `xj`.

(b) Run your function with the values $x_1 = 1/\pi, 0.5$ and $3/\pi; r = 0.5; n = 4000; i = 10$. Briefly describe the values printed out. Comment on the plots.

(c) Repeat part (b) with $r = 3$.

Question 2 (10 marks)

Let π be the probability of success for a Bernoulli trial. Let X be the number of failures obtained to obtain r successes. X is said to follow a negative binomial distribution with parameters r and π.

$$\Pr(X=i) = \binom{i+r-1}{r-1}\pi^r(1-\pi)^i, \quad i = 0,1,2,\ldots$$

(a) Construct an algorithm to sample from the above distribution and implement it in R for $\pi = 0.6$ and $r = 3$. You should write the algorithm in your answer.

(b) Generate $100,000$ samples and compute the sample mean.

(c) Report the estimated $\Pr(X=i)$ from your sampled values for $i = 0,\ldots,10$

Question 3 (20 marks)

The file *heart_failure.csv* contains 299 observations and 13 variables. This data comes from Chicco and Jurman (2020). The variables in the dataset are:

- `age`: Age of the patient (years)
- `anaemia` : Decrease of red blood cells or hemoglobin (binary)
- `creatinine_phosphokinase`: Level of the CPK enzyme in the blood (mcg/L)
- `diabetes` : If the patient has diabetes (binary). 1 means the patient has diabetes
- `ejection_fraction` : Percentage of blood leaving the heart at each contraction (percentage)
- `high_blood_pressure`: If a patient has hypertension (binary)
- `platelets` : Platelets in the blood (kiloplatelets/mL)
- `serum_creatinine`: Level of creatinine in the blood (mg/dL)
- `serum_sodium`: Level of sodium in the blood (mEq/L)
- `sex` : Woman or man (binary). Man is 1
- `smoking`: If the patient smokes or not (binary)
- `time`: Follow-up period (days)
- `DEATH_EVENT` If the patient deceased during the follow-up period (binary). 1 means the patient deceased.

(a) Fit a binomial regression model (logistic regression in this case) to predict mortality by heart failure using all the features in this data set. Print the estimated coefficients of the model. What is the residual deviance and degrees of freedom of this model?

(b) We would like to treat `diabetes` as a factor. Fit a second model that allows the effect of creatinine level to depend on whether the patient has diabetes or not. What is the residual deviance and degrees of freedom of this model?

(c) What is the test statistic and p-value for the null hypothesis of a common slope versus the alternative of different slopes for `serum_creatinine`? What should be the preferred model and why?

(d) Perform stepwise selection with the preferred model in part (c) and print the selected model. Which variables are selected?

(e) What is the predicted probability of not surviving of a female patient age = 50, level of the CPK enzyme is 250 mcg/L, 30% of blood leaving the heart at each contraction , level of blood creatinine is 2 mg/dL , level of blood sodium is 130 mEq/L , with follow-up of 90 days, has hypertension, does not smoke, has diabetes and anaemia, and platelets level 262000 kiloplatelets/mL.

(f) Report the false negative rate for the final model, using the cutoff 0.7. Take negative outcome to be the event that the patient deceased.

End of Assignment — Total Available Marks = 40