UNIVERSITY OF LONDON

ST3189

Summer 2020 online assessment guidance

ST3189 Machine Learning

The assessment will be an **open-book take-home online assessment within a 24-hour window**. The requirements for this assessment remain the same as the originally planned closed-book exam, with an expected time/effort of 2 hours.

Candidates should answer all **FOUR** questions. All questions carry equal marks.

You should complete this assessment using **pen and paper**. Please use **BLACK ink only**.

Handwritten work then needs to be scanned, converted to PDF and then uploaded to the VLE as **ONE individual file** including the coversheet. Each scanned sheet should have your **candidate number** written clearly at the top. Please **do not write your name anywhere** on any sheet.

The paper will be available at **12.00 midday (BST) on Monday 27 July 2020**.

You have **until 12.00 midday (BST) on Tuesday 28 July 2020** to upload your file into the VLE submission portal. However, you are advised not to leave your submission to the last minute. *We will deduct 5 marks if your submission is up to one hour late, 10 marks if your submission is more than one hour late but less than two hours late (etc.).*

An appendix with properties of common distributions is provided at the end of the paper.

Workings should be submitted for all questions requiring calculations. Any necessary assumptions introduced in answering a question are to be stated.

You may use *any* calculator for any appropriate calculations, but you may not use any computer software to obtain solutions. Credit will only be given if all workings are shown.

If you think there is any information missing or any error in any question, then you should indicate this but proceed to answer the question stating any assumptions you have made.

The assessment has been designed with a duration of 24 hours to provide a more flexible window in which to complete the assessment and to appropriately test the course learning outcomes. As an open-book assessment, the expected amount of effort required to complete all questions and upload your answers during this window is no more than 2 hours. Organise your time well and avoid working all night.

UL20/0544

You are assured that there will be no benefit in you going beyond the expected 2 hours of effort. Your assessment has been carefully designed to help you show what you have learned in the hours allocated.

This is an open book assessment and as such you may have access to additional materials including but not limited to subject guides and any recommended reading. But the work you submit is expected to be 100% your own. Therefore, unless instructed otherwise, you must not collaborate or confer with anyone during the assessment. The University of London will carry out checks to ensure the academic integrity of your work. Many students that break the University of London's assessment regulations did not intend to cheat but did not properly understand the University of London's regulations on referencing and plagiarism. The University of London considers all forms of plagiarism, whether deliberate or otherwise, a very serious matter and can apply severe penalties that might impact on your award. The University of London 2019-20 Procedure for the Consideration of Allegations of Assessment offences is available online at:

https://london.ac.uk/sites/default/files/governance/assessment-offence-procedure-year-2019-2020.pdf

The University of London's Rules for Taking Online Timed Assessments have been included in an update to the University of London General Regulations and are available at:

https://london.ac.uk/sites/default/files/regulations/progregs-general-2019-2020.pdf

Answer **all** parts of the following questions.
An **appendix** with properties of common distributions is provided at the end.

1. (a) Indicate whether the following statements are true or false. Briefly justify your answers.

   i. Other things equal, a classifier trained on less training data is more likely to overfit. **[3 marks]**

   ii. Consider the lasso approach to perform variable selection when regressing S&P 500 index daily returns on the daily returns of a number of selected stocks. One can apply the basic 5-fold or 10-fold cross-validation method to choose the optimal tuning parameter, $\widehat{\lambda}$, and consistently get good predictive performance. **[3 marks]**

   iii. To perform variable selection, one can use the information criterion approach to select the model size, including AIC and BIC. Compared with AIC, BIC tends to select a larger model size when the sample size n becomes sufficiently large. **[3 marks]**

   (b) Consider minimising the following penalised sum of squared errors

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \tag{1}$$

   over $\theta \in \mathcal{R}$, where $\mathcal{R}$ is the real line and $\lambda \geq 0$ is the tuning parameter.

   i. Take $p_\lambda(|\theta|) = \lambda|\theta|$, show that the solution to (1) is

$$\widehat{\theta} = \text{sign}(z)(|z| - \lambda)I(|z| > \lambda),$$

   where $\text{sign}(z)$ denotes the sign of z and $I(\cdot)$ denotes an indicator function, which is 1 if the statement in $(\cdot)$ is true and 0 otherwise. **[6 marks]**

   ii. Take $p_\lambda(|\theta|) = \lambda_1\theta^2 + \lambda_2|\theta|$, where $\lambda_1, \lambda_2 \geq 0$ are two tuning parameters. What is the solution to equation (1)?
   (*Hint: Rewrite the optimisation problem in terms of the form in equation* (1) *and then apply the result in part i.*) **[6 marks]**

   (c) The table below is the training data with $n = 6$ observations of a 3-dimensional input $\mathbf{x} = (x_1, x_2, x_3)^T$ and a qualitative output y (the colour is blue or red).

i	x_1	x_2	x_3	y
1	0	3	0	red
2	2	0	0	red
3	0	1	3	red
4	0	1	2	blue
5	-1	0	1	blue
6	1	1	1	blue

   i. Compute the Euclidean distance between each observation in the training data and the test point $\mathbf{x}_* = (0, 0, 0)^T$. **[2 marks]**

   ii. If we use k-nearest neighbours with $k = 3$, what is the predicted colour corresponding to the test point $\mathbf{x}_*$? **[2 marks]**

2. Consider a sample $x = (x_1, \ldots, x_n)$ of independent random variables, identically distributed from the Rayleigh(θ) distribution. The probability density function for each x_i is

$$f(x_i|\theta) = \theta x_i \exp(-0.5\theta x_i^2), \quad x_i > 0,$$

where $\theta > 0$ is a unknown parameter.

(a) Assign the Gamma(α, β) prior on θ, for $\alpha, \beta > 0$ and derive the corresponding posterior distribution. **[5 marks]**

(b) Derive the Jeffreys prior for θ. Use it to obtain the corresponding posterior distribution. **[7 marks]**

(c) Provide a normal approximation to the posterior distribution of θ that converges to the true posterior as the sample size increases. **[4 marks]**

(d) Assuming a quadratic error loss function, find the Bayes estimator for each of the posteriors in parts (a), (b) and (c). **[4 marks]**

(e) Let y represent a future observation from the same model. Find the posterior predictive distribution of y for one of the posteriors in parts (a) or (b). **[5 marks]**

3. (a) Given a random data sample consisting of n observations, consider a bootstrap sample generated from it.

i. Describe how the bootstrap sample is obtained and argue that the probability that the j-th observation is not in the bootstrap sample is $\left(1 - \frac{1}{n}\right)^n$.

ii. Provide the probability that the j-th observation is in the bootstrap sample for $n = 5$.

iii. Name a potential use for the observations that are not in the bootstrap sample.

iv. Suppose that we want to predict Y for a particular value of the predictor X based on a machine learning method. Describe how we can estimate the standard deviation of our prediction.

[12 marks]

(b) The tree in Figure 1 (on the next page) provides a regression tree based on a dataset regarding fuel consumption (in miles per gallon) for 392 cars. The variables appearing in the regression tree are *cylinders:* number of cylinders (between 4 and 8) and *horsepower:* engine horsepower.

i. Provide an interpretation of this tree. **[6 marks]**

ii. Create a diagram that represents the partition of the predictors' space according to the tree of Figure 1. **[7 marks]**

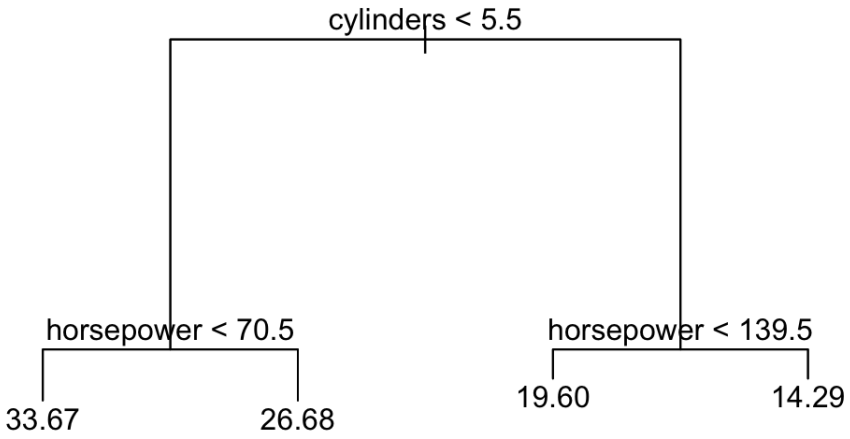

Figure 1: Regression tree for Question 3 (b)

4. (a) Some of the most commonly-used types of linkage in hierarchical clustering are:

   - *Single linkage*: distance between clusters is the *minimum* distance between any pair of points from two clusters.
   - *Complete linkage*: distance between clusters is the *maximum* distance between any pair of points from two clusters.
   - *Average linkage*: distance between clusters is the *average* distance between any pair of points from two clusters.

   i. Which of the three types of linkage described above would most likely result in clusters most similar to those given by k-means?

   [4 marks]

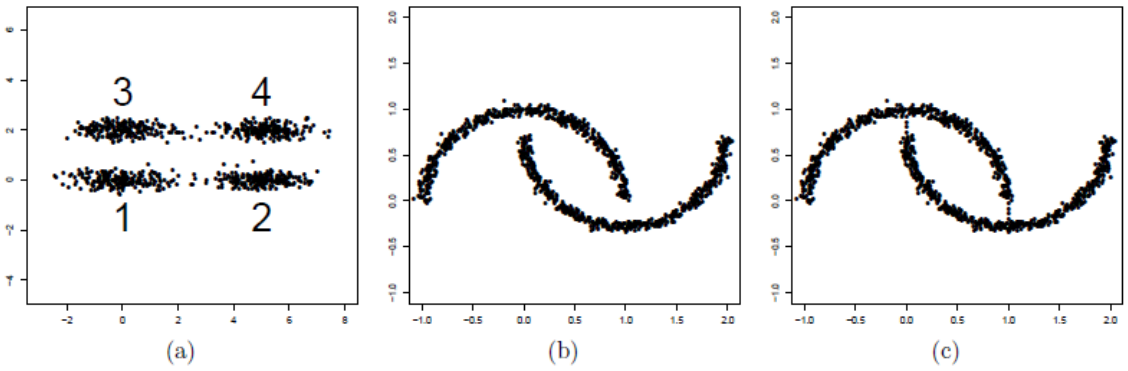

Figure 2: For Question 4 (a)

ii. Consider the data in Figure 2(a). What would be the result if we extract two clusters from the tree given by hierarchical clustering on this dataset using single linkage? Describe your answer in terms of the labels 1–4 given to the four 'clumps' in the data. Do the same for complete linkage and average linkage. **[5 marks]**

iii. Which of the three types of linkage (if any) would successfully separate the two 'half-moons' in Figure 2(b)? Which about Figure 2(c)? Briefly explain your answer. **[6 marks]**

(b) Consider a binary classification problem, $Y = 1$ (class 1) and $Y = 2$ (class 2). Suppose that $X|Y = 1$ is from a normal distribution $N(2, 1^2)$ and $X|Y = 2$ is from the Normal distribution $N(1, 0.5^2)$. In addition, $P(Y = 1) = 1/3$ and $P(Y = 2) = 2/3$. Derive the Bayes decision rule that classifies regions to classes 1 and 2 for this problem. **[10 marks]**

Appendix: Table of Common Distributions

Binomial(n, θ)**:** number of successes in n independent Bernoulli trials with probability of success θ.

- $f(x|\theta) = P(x|\theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$ for $x = 0,\ 1, \ldots,\ n$.
- $E(X) = n\theta$, $\text{Var}(X) = n\theta(1-\theta)$.

NegBin(r, θ)**:** number of successes before r^{th} failures in repeated independent Bernoulli trials.

- $f(x|\theta) = P(x|\theta) = \binom{x+r-1}{x}\theta^x(1-\theta)^r$ for $x = 0,\ 1, \ldots$.
- $E(X) = \frac{r(1-\theta)}{\theta}$, $\text{Var}(X) = \frac{r(1-\theta)}{\theta^2}$.

Poisson(λ)**:** often used for the number of events which occur in an interval of time.

- $f(x|\lambda) = P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0,\ 1, \ldots$.
- $E(X) = \lambda$, $\text{Var}(X) = \lambda$.

Normal N(μ, σ^2)**:** characterized by first two moments.

- $f(x) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $-\infty < x < \infty$.
- $E(X) = \mu$, $\text{Var}(X) = \sigma^2$.

Beta(α, β)**:** characterized by parameters $\alpha > 0$ and $\beta > 0$.

- $f(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ for $0 \le x \le 1$, $B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- $E(X) = \frac{\alpha}{\alpha+\beta}$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$.

Gamma(α, β)**:** characterized by parameters $\alpha > 0$ and $\beta > 0$.

- $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$ for $0 \le x < \infty$, $\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}dy$.
- $E(X) = \frac{\alpha}{\beta}$, $\text{Var}(X) = \frac{\alpha}{\beta^2}$.

IGamma(α, β)**:** characterized by parameters $\alpha > 0$ and β. If $X \sim \text{Gamma}(\alpha, \beta)$, then $1/X \sim \text{IGamma}(\alpha, \beta)$.

- $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right)$ for $0 \le x < \infty$.
- $E(X) = \frac{\beta}{\alpha-1}$, $\text{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$. for positive integer n.

END OF PAPER