UNIVERSITY OF LONDON

ST3189 ZB

BSc DEGREES AND GRADUATE DIPLOMAS IN ECONOMICS, MANAGEMENT, FINANCE AND THE SOCIAL SCIENCES, THE DIPLOMA IN ECONOMICS AND SOCIAL SCIENCES AND THE CERTIFICATE IN EDUCATION IN SOCIAL SCIENCES

Machine Learning

Wednesday 15 May 2019: 10.00 – 12.00

Time allowed: 2 hours

DO NOT TURN OVER UNTIL TOLD TO BEGIN

Candidates should answer all **FOUR** questions. All questions carry equal marks. **Candidates are strongly advised to divide their time accordingly.**

Graph paper is provided at the end of this question paper. If used, it must be detached and fastened securely inside the answer book.

A handheld calculator may be used when answering questions on this paper and it must comply in all respects with the specification given with your Admission Notice. The make and type of machine must be clearly stated on the front cover of the answer book.

A table of common distributions is included as an appendix on the final page of this paper.

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

1. (a) Indicate whether the following statements are true or false, providing an explanation in one sentence.

i. The training error of 1-nearest neighbour is always 0. [3 marks]

ii. A decision tree generated with binary splitting has always $m+1$ terminal nodes, where m is the number of internal nodes. [3 marks]

iii. In K-fold cross-validation each data point belongs to exactly one test fold, so the test folds are independent. Then the error estimates of the separate folds are also independent. Equivalently, given that the data in test folds i and j are independent, e_i and e_j, the error estimates on test folds i and j are also independent. [3 marks]

(b) Consider minimising the following penalized 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 λ is a non-negative tuning parameter.

i. For $p_\lambda(|\theta|) = \frac{\lambda}{2}\theta^2$, show that (1) is $\widehat{\theta} = \frac{z}{1+\lambda}$. [4 marks]

ii. Take $p_\lambda(|\theta|) = \frac{\lambda^2}{2}I(|\theta| \neq 0)$, where $I(\cdot)$ denotes an indicator function, which is one if the statement in $(\cdot)$ is true and 0 otherwise. Show that (1) is $\widehat{\theta} = zI(|z| > \lambda)$. [5 marks]

(c) Consider the k-nearest neighbours approach using Euclidean distance on the dataset shown in Figure 1.

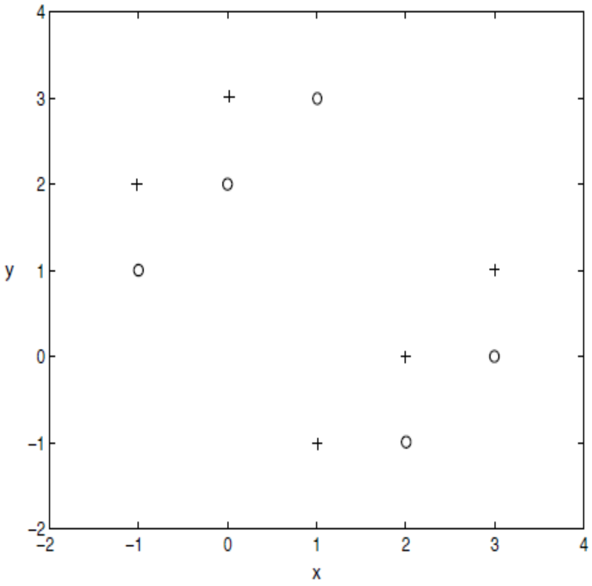

Figure 1: Each point belongs to one of two classes: “ + ” and “ ∘ ”.

i. What is the *Leave-One-Out Cross Validation* (LOOCV) misclassification error when using 1-nearest neighbour? [3 marks]

ii. Which of the following values of k leads to the lowest LOOCV misclassification error: 3, 5, 9? What is the error for that k? [4 marks]

2. Let $x = (x_1, \dots, x_n)$ be a random sample from the Gamma$(2, \theta)$ distribution, with probability density function given by

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

Note that $E(x_i) = 2/\theta$.

(a) Assign the Gamma(α, β) prior to θ and find the corresponding posterior distribution. [7 marks]

(b) Find a Bayes estimator for θ based on the posterior in part 2(a). [5 marks]

(c) Discuss how you can calculate a 95% credible interval. Provide the interpretation of such an interval [6 marks]

(d) Assume that $n = 5 \sum_{i=1}^{n} x_i = 29.6$. Find the Bayes factor for the hypotheses $H_0 : \theta = 0.5$ and $H_1 : \theta = 1$ and provide an interpretation of it. Note that you can answer in terms of logarithms or exponentials and there is no need to give the final number. [7 marks]

3. (a) Consider the simple linear regression model for data $(X_i, Y_i)_{i=1}^n$ that can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon,$$

where ϵ_i's are independent, normally distributed with zero mean and variance σ^2. Show that the least squares line will always pass from the point $(\bar{X}, \bar{Y})$. [6 marks]

(b) Describe what is meant by overfit in the context of linear regression and provide an example where overfit can occur. What steps can be taken to guard against overfit? [6 marks]

(c) The tree in Figure 2 provides a regression tree based on a dataset regarding child car seat sales (in thousands of units) obtained from 400 stores. The variables appearing in the regression tree are *Age:* average age of the population in the area of the store, *income:* community income level (in thousands of dollars).

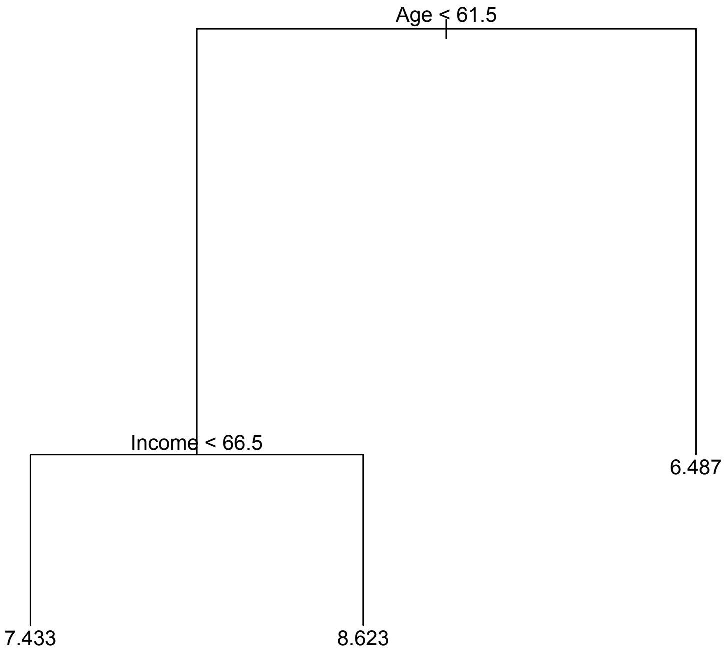

Figure 2: Regression tree for Question 3 (b) i.

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 2. [7 marks]

4. (a) Suppose we are given the following dataset

$$\{(0,4),(1,3),(2,6),(3,3),(5,5),(5,6),(9,9)\}.$$

We apply k-means, with $k = 3$, to cluster the dataset using the Manhattan distance for computing the distances between centroids and pairs in the dataset. The Manhattan distance between (x_1, y_1) and (x_2, y_2) is defined as $|x_1 - x_2| + |y_1 - y_2|$. Suppose that the initial clusters of k-means are C_1, C_2 and C_3 with

- $\{(1,3),(3,3),(5,6)\}$,
- $\{(0,4),(2,6)\}$,
- $\{(5,5),(9,9)\}$,

and that the centroids for C_1, C_2 and C_3 are $(3,4)$, $(1,5)$ and $(7,7)$ respectively. If we then run k-means for one single iteration, what are the new clusters and what are their centroids? [8 marks]

(b) Let $x_1, \ldots, x_n$ be n observations and $\{C_1, \ldots, C_K\}$ be a partition of $\{1, \ldots, n\}$. Let n_k be the number of observations in cluster C_k for $k = 1, \ldots, K$. For each cluster C_k, define $\bar{x}_k = \frac{1}{n_k}\sum_{i \in C_k} x_i$ to be the within-cluster mean and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ to be the overall mean. Let

$$T = \sum_{k=1}^{K}\sum_{i \in C_k}(x_i - \bar{x})^2 \text{ to be the total deviance to the overall mean,}$$

$$W = \sum_{k=1}^{K}\sum_{i \in C_k}(x_i - \bar{x}_k)^2 \text{ to be the within-cluster deviance to the cluster mean,}$$

$$B = \sum_{k=1}^{K}\sum_{i \in C_k}(\bar{x}_k - \bar{x})^2 \text{ to be the between-cluster deviance.}$$

i. Verify that $T = W + B$. [5 marks]

ii. Explain how T and B change during the course of the K-means algorithm. [4 marks]

(c) Consider a three-class classification problem with $Y = 1$ (Class 1) or $Y = 2$ (Class 2) or $Y = 3$ (Class 3). Let $\mu_1 = 0, \mu_2 = 1, \mu_{31} = 0.5, \mu_{32} = 0.5$ and $\sigma^2 = 1$. Suppose that $X|\{Y = 1\}$ and $X|\{Y = 2\}$ are from normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. Suppose that $X|\{Y = 3\}$ is from a mixture of two normal distributions with 50% chance being from $N(\mu_{31}, \sigma^2)$ and the other 50% chance being from $N(\mu_{32}, \sigma^2)$. In addition, $P(Y = 1) = P(Y = 2) = P(Y = 3) = 1/3$. Classify the point $x_0 = 0.3$ based on the posterior probabilities. Note that you can answer in terms of logarithms or exponentials and there is no need to give the final numbers. [8 marks]

END OF PAPER

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 $x = 0,\ 1, \ldots,\ n$.
- $E(X) = n\theta, \quad \text{Var}(X) = n\theta(1-\theta)$.

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 $x = 0,\ 1, \ldots$.
- $E(X) = \frac{r(1-\theta)}{\theta}, \quad \text{Var}(X) = \frac{r(1-\theta)}{\theta^2}$.

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 $x = 0,\ 1, \ldots$.
- $E(X) = \lambda, \quad \text{Var}(X) = \lambda$.

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 $-\infty < x < \infty$.
- $E(X) = \mu, \quad \text{Var}(X) = \sigma^2$.

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

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

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