THE UNIVERSITY OF WARWICK

Department: Computer Science

Module Code: CS910

Module Title: Foundations of Data Analytics

Exam Paper Code: CS9100_B

Duration: 2 hours

Exam Paper Type: 24-hour window

STUDENT INSTRUCTIONS

1. Read all instructions carefully. We recommend you read through the entire paper at least once before writing.
2. There are SEVEN questions. Answer BOTH questions from Section A, and choose THREE questions from Section B.
3. You should not submit answers to more than the required number of questions.
4. You should handwrite your answers either with paper and pen or using an electronic device with a stylus (unless you have special arrangements for exams which allow the use of a computer). Start each question on a new page and clearly mark each page with the page number, your student id and the question number. Handwritten notes must be scanned or photographed and all individual solutions collated into a single PDF with pages in the correct order.
5. Please ensure that all your handwritten answers are written legibly, preferably in dark blue or black ink. If you use a pencil ensure that it is not too faint to be captured by a scan or photograph.
6. Please check for legibility before uploading. It is your responsibility to ensure your work can be read.
7. Add your student number to all uploaded files.
8. You are permitted to access module materials, notes, resources, references and the Internet during the online assessment.
9. You must not communicate with any other candidate during the assessment period or seek assistance from anyone else in completing your answers. The Computer Science Department expects the conduct of all students taking this assessment to conform to the stated requirements. Measures will be in operation to check for possible misconduct. These will include the use of similarity detection tools and the right to require live interviews with selected students following the assessment.
10. By starting this assessment, you are declaring yourself fit to undertake it. You are expected to make a reasonable attempt at the assessment by answering the questions in the paper.

IMPORTANT INFORMATION

- We strongly recommend you use Google Chrome or Mozilla Firefox to access the Alternative Exams Portal.
- You are granted an additional 45 minutes beyond the stated duration of this assessment to allow for downloading/uploading your assessment, your files and any technical delays.
- Students with approved Alternative Exam Arrangements (Reasonable Adjustments) that permit extra time and/or rest breaks will have this time added on to the stated duration.
- You must have completed and uploaded the assessment before the 24-hour assessment window closes.
- Late submissions are not accepted.
- If you are unable to submit your assessment, you must submit Mitigating Circumstances immediately, attaching supporting evidence and your assessment. The Mitigating Circumstances Panel will consider the case and make a recommendation based on the evidence to the Board of Examiners.

SUPPORT DURING THE ASSESSMENT

Operational Support:

- Use the Alternative Exams Portal to seek advice immediately if during the assessment period:
  - you cannot access the online assessment
  - you believe you have been given access to the wrong online assessment

Technical Support:

- If you experience any technical difficulties with the Alternative Exam Portal please contact *helpdesk@warwick.ac.uk*.
- Technical support will be available between 09:00 and 17:00 BST for each examination (excluding Sunday).

Academic Support:

- If you have an academic query, contact the invigilator (using the 'Contact an Invigilator' tool in AEP) to raise your issue. Please be aware that two-way communication in AEP is not currently possible.
- Academic support will normally be provided for the duration of the examination (i.e. for a 2 hour exam starting at 09:00 BST, academic support will normally be provided between 09:00 and 11:45 BST). Academic support beyond this time is at the discretion of the department.

Other Support:

- If you can not complete your assessment for the following reasons submit Mitigating Circumstances immediately:
  - you lose your internet connection
  - your device fails
  - you become unwell and are unable to continue
  - you are affected by circumstances beyond your control

Continued . . .

THE UNIVERSITY OF WARWICK

Examination: Summer 2021

Foundations of Data Analytics

Section A Solve **BOTH** questions.

1. [20 marks]

   Consider a data set (X, Y), where X is the feature representing "Years-of-Education" and Y is the class/label representing "Salary". Assume the following set of possible values: $X \in \{0, 1, \ldots, 20\}$ and $Y \in \{0, 1, \ldots, 10000\}$ (a value $Y = y$ means a salary of y thousand pounds). Assume further that we are given the true distribution $\mathbb{P}(x, y)$ of the data (X, Y).

   (a) Which of the above assumptions about the data (X, Y) is the least realistic? [3]

   (b) Say that we want to test the performance of some classification algorithm using the following loss model: $L(\widehat{y}, y) = \widehat{y} - y$, where $\widehat{y}$ is the predicted value and y is the true value. Is this a reasonable loss model? Justify your answer. [5]

   (c) Consider some feature x. What is the best corresponding label $\widehat{y}$ for the loss model $L_1(\widehat{y}, y) = \widehat{y} - y$? What about for the loss model $L_2(\widehat{y}, y) = y - \widehat{y}$? [5]

   (d) Under which loss model would $\mathsf{E}[Y]$ be the best corresponding label for some feature x? [7]

   Solution: *Comprehension – requires student to show understanding of concepts*

   (a) Assuming that the true distribution of the data is given is the least realistic.

   (b) No. Say that, for some feature x the true value is $y = 30$. For the possible predicted values $\widehat{y}_1 = 40$ and $\widehat{y}_2 = 20$, the loss model yields vastly different numbers, i.e., 10 and -10, although both $\widehat{y}_1$ and $\widehat{y}_2$ are within the same margin from the true value y.

   (c)
   $$\begin{aligned} Loss(\widehat{y}) &= \mathsf{E}[L_1(\widehat{y}, Y) \mid X = x] \\ &= \mathsf{E}[\widehat{y} - Y \mid X = x] \\ &= \widehat{y} - \mathsf{E}[Y \mid X = x] \,, \end{aligned}$$
   and hence the best label is $\min Y$, i.e, 0 (i.e., always predict a salary of 0). Under the other loss model L_2, the best label would be $\max Y$, i.e., 10000.

   (d)
   $$L(\widehat{y}, y) = 1_{\widehat{y} \neq \mathsf{E}[Y]} \,,$$
   because the loss would be 0 only when $\widehat{y} = \mathsf{E}[Y]$.

2. [20 marks]

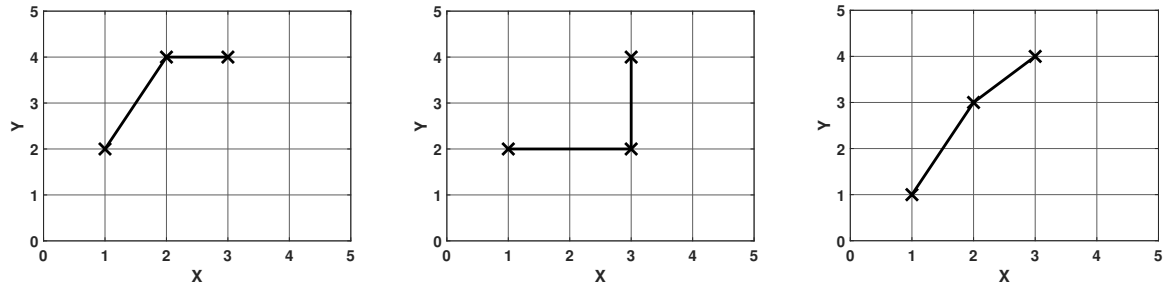

Figure 1: Q-Q plots

(a) Consider the data (X, Y), where X and Y are random variables with $\mathbb{P}(X = i) = p_i$ for $i = 1, 2, 3$ and $\mathbb{P}(Y = j) = q_j$ for $j = 2, 3, 4$. Let $S = \{0, 0.5, 1\}$ be a set of probabilities. For each of the Q-Q plots from Figure 1, determine the parameters p_i's and q_j's, to generate the quantiles of that particular plot (represented by an 'x') and corresponding to the probabilities from S. For each case, your answer should look like:

$$X : \begin{pmatrix} 1 & 2 & 3 \\ p_1 & p_2 & p_3 \end{pmatrix}, \; Y : \begin{pmatrix} 2 & 3 & 4 \\ q_1 & q_2 & q_3 \end{pmatrix} .$$

[12]

(b) Someone has access to the number of inhabitants in $n = 40000$ UK cities, i.e., $x_1, x_2, \ldots, x_n$; the corresponding average is μ and the standard deviation is σ. Describe a procedure to determine whether the underlying distribution is heavy-tailed.[8]

Solution: *Comprehension – requires student to show understanding of concepts*

(a)

$$\text{For (a)} \quad X : \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \; Y : \begin{pmatrix} 2 & 3 & 4 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\text{For (b)} \quad X : \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{pmatrix}, \; Y : \begin{pmatrix} 2 & 3 & 4 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

No solution is possible for (c) because 1 cannot be a quantile of Y, as 1 is strictly smaller than the smallest value of Y.

(b) One approach is to plot the empirical distribution of the data on a log-log scale; if the plot is approximately linear then there is enough visual evidence that the data is heavy-tailed.

Section B Choose **EXACTLY THREE** questions.

3. [20 marks]

Consider a data set $(X, Y) = \{(x_i, y_i) : i = 1, \ldots, 2n\}$, where $n > 0$; X is the feature and Y is the class/label. You build a prediction model on the first half of the data (i.e., the training part) and test it on the remaining half. Let

$$f(n) := E_1(n) - E_2(n) \ ,$$

where $E_1(n)$ and $E_2(n)$ are the errors on the training and testing parts, respectively.

(a) Describe the monotonicity of $f(n)$ in the parameter n; e.g., is the function $f(n)$ necessarily increasing or decreasing? Justify your answer. [15]

(b) Describe the behavior of $E_2(n)$ as a function of n. [5]

Solution: *Comprehension – requires student to show understanding of concepts*

(a) $f(n)$ is neither increasing nor decreasing. Consider the $0/1$ loss/error model, $n = 1$, the data $(X, Y) = \{(1, 11), (2, 12)\}$, and the model $y = x + 10$. Then $f(1) = 0$.

Consider now $n = 2$, the data $(X, Y) = \{(1, 11), \cancel{(2, 13),} (3, 13), (4, 14)\}$, and the model $y = x + 10$. Then $f(2) = 1$. However, if the training and testing parts are swapped, then $f(2) = -1$.

(b) $E_2(n)$ converges to the true error $\mathsf{E}\left[L(\widehat{Y}, Y)\right]$ for some loss model L.

4. [20 marks]

   (a) In some exam, you are given a multiple-choice problem with n choices, out of which exactly one is correct. There is a probability of p that you have already seen the problem, in which case you know the solution. In the case when you have not seen the problem, you pick one of the choices uniformly at random. What is the probability that you have seen the problem, provided that you answered correctly? [10]

   (b) Three roommates have to decide which one of them should make the coffee today! They agree to simultaneously roll a six-sided die each, until one of them gets a number which is twice the sum of the other two; that person must make the coffee. How many plays do they need on average? [10]

Solution: *Comprehension – requires student to show understanding of concepts*

(a) Denote by K (you knew the problem), C (correct answer), N (you didn't know the problem). We are asked for

$$\mathbb{P}(K|C) = \mathbb{P}(C|K)\frac{\mathbb{P}(K)}{\mathbb{P}(C)} = \frac{p}{\mathbb{P}(C|K)\mathbb{P}(K) + \mathbb{P}(C|N)\mathbb{P}(N)} = \frac{p}{p + \frac{1}{n}(1-p)} .$$

(b) Denote the three numbers by X, Y, Z. Then

$$\mathbb{P}(X = 2(Y+Z)) = \mathbb{P}((4,1,1),(6,1,2),(6,2,1)) = \frac{3}{216} = \frac{1}{72} .$$

The success probability is hence $p := \frac{3}{72} = \frac{1}{24}$. We are asked the expected value of a Geometric random variable with parameter p, which is $\frac{1}{p}$. Thus, the roommates need 24 plays on average.

5. [20 marks]

   Consider the training data set $(X, Y) = \{(x_1, y_1), (x_2, y_2)\}$, where x_i's and y_i's are some natural numbers.

   (a) Give some values for the x_i's and y_i's such that a feature $x = 3$ will be classified identically by both Linear Regression and Naïve Bayes Classifier. [6]

   (b) Give some values for the x_i's and y_i's such that a feature $x = 3$ will be classified differently by Linear Regression and Naïve Bayes Classifier. [6]

   (c) (Independent of (a) and (b)) A lecturer wants to predict what scores will students achieve in the next exam. The students' features are:

   - height (in cm)
   - number of Tweeter followers
   - eyes' colour (represented numerically)
   - number of lectures attended
   - coursework grades

   The lecturer has a training data set of 100 students previously taking the exam and is keen on using multilinear regression as a prediction model. Initially, the lecturer chooses one of the five features at random and computes the training error. Then, the lecturer keeps adding additional features, as long as the training error is smaller or equal to the previous training error with one less feature. With how many features will the lecturer's multilinear regression model end up with? Justify your answer. [8]

Solution: *Comprehension – requires student to show understanding of concepts*

(a) $\{(3, 3), (3, 3)\}$. The linear regression model would be $y = f(x) = x$, and thus $f(3) = 3$; Naïve Bayes would also output 3, as 3 is the only class/label in the training set.

(b) $\{(1, 1), (2, 2)\}$. The linear regression model would be $y = f(x) = x$, and thus $f(3) = 3$. However, Naïve Bayes would either output 1 or 2, as those are the only labels in the training set.

(c) Five, because adding a feature can only decrease the training error.

6. [20 marks]

Consider the following data set

$$(X_1, X_2, Y) = \begin{pmatrix} 3 & 6 & 1 \\ 1 & 5 & 0 \\ 4 & 8 & 1 \\ 2 & 7 & 0 \end{pmatrix},$$

where Y is the label. You intend to build an ID3 decision tree classifier using only split rules of the form $X_1 \geq a$, $X_1 \leq b$, $X_2 \geq c$, or $X_2 \leq d$, for some numbers a, b, c, d.

(a) What is the entropy at the root of the tree? [3]

(b) What is the split rule at the root of the tree? [7]

(c) How many more splits would ID3 perform, after the split rule from (b)? [3]

(d) What would be the worst split rule at the root of the tree, i.e., minimizing the information gain? [7]

Justify all your answers.

Solution: *Comprehension – requires student to show understanding of concepts*

(a)

$$-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

(b) Sorting on X_1 and X_2 we have the sub-data sets

$$(X_1, Y) = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \\ 4 & 1 \end{pmatrix} \text{ and } (X_2, Y) = \begin{pmatrix} 5 & 0 \\ 6 & 1 \\ 7 & 0 \\ 8 & 1 \end{pmatrix}$$

The best rule is $X_1 \leq 2$ because that would result in an information gain of 1.

(c) None, because the entropies on both branches become 0.

(d) The worst rule is $X_2 \leq 6$ because that would result in an information gain of 0.

7. [20 marks]

Consider the following *online* version of the K-means clustering algorithm, framed in the 2-dimensional Euclidean space: Initially, K centroids are randomly generated. New data points are received in an online manner (i.e., one after another); each new point is allocated to the cluster with the closest centroid, relative to the Euclidean distance, and then the centroid of that particular cluster is recomputed as the average between 2 points (the previous centroid and the new point).

(a) What is the key advantage of the online K-means in comparison with the original?[5]

(b) Give an instance of 4 points for which the original K-means yields a better performance than the online K-means. [8]

(c) Give an instance of 4 points for which the online K-means yields a better performance than the original K-means. [7]

Solution: *Comprehension – requires student to show understanding of concepts*

(a) The online K-means can yield substantial savings in the memory requirement. Unlike the original version which stores all the points in memory, the online version only keeps $K + 1$ points.

(b) Consider the points $\{(0, 0), (0, 1)\}$ (in some imaginary cluster A) and the points $\{(10, 0), (10, 1)\}$ (in some imaginary cluster B). Assume that the original K-means initially picks $(0, 0)$ and $(10, 0)$, whereas online K-means picks $(0, 0)$ and $(0, 1)$. The original K-means yields the correct clustering, as opposed to the online K-means.

(c) Similarly as in (b), except that the initial picks are reversed.