

BUSINESS SCHOOL

Predictive Analytics and Modelling of Data

**CMSE11428**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Exam Date:  **17th December 2022** |  | From and To:  **13:00 – 15:00** |  | Exam Diet: | **December Exam Diet** |

Please read full instructions before commencing writing

|  |
| --- |
| Exam paper information   * Answer **ALL** questions in **Part 1** * Answer **ONE** of the two questions in **Part 2** |
| Special instructions   * Use **Multiple Choice Question** sheet provided for answering **Part 1** * Use the **Script Book** provided for answering **Part 2** |
| Special items   * Multiple Choice Question sheet   (see instructions on how to fill this out on the next page) |

This examination will be marked anonymously

**INSTRUCTIONS: Multiple Choice Question sheet – Part 1**

Please make sure you read carefully and understand the instructions below and that you comply with these fully.

1. Use only a HB pencil (provided).

2. Multiple choice answer sheet: Using the HB pencil provided, write your name, the name of the Examination and the date in the shaded box on the multiple choice answer sheet. Write your **matriculation number.**

3. Each question contains four possible answers, only **one** of which is correct. Place a firm horizontal pencil mark in the appropriate gap, e.g., if the answer to question 5 was (C).

Q5[A] [B] [C] [D] [E]

4. Faint lines will **not** be read!

Very thin marks made with a sharp, hard pencil may not be read.

If a soft pencil is used there should be no need to press heavily.

This should make rubbing out easy.

Erasures must be completed clean without smudging. An unclean or incomplete erasure may be read as an answer!

5. One mark is given for a correct answer.

No marks will be subtracted for a wrong answer.

No mark is either given or subtracted for:

(a) no attempt

(b) if a line is placed through more than one option

6. The multiple choice answer sheets are machine marked and that mark is final.

**Part 1: Multiple choice questions**

Attempt **ALL** questions

This section has a total of 35 marks, each question is equally weighted and marked out of 5.

Use the **multiple choice answer sheet** to answer questions 1-7

1. What is an example for a linear regression problem? **D**
   1. Classifying machine parts as faulty/not faulty using data from the manufacturing machine
   2. Predicting whether a person will default on their credit
   3. Segmenting a customer database using their sales behaviour
   4. Predicting the number of attendees at a concert based on past attendance numbers
2. You are trying to build a **neural network model** to predict customer sales. A large number of observations have missing values in one predictor. Do you... **C**
   1. Ignore the missing values and fit the neural network, as this method is able to handle missing data well
   2. Remove all observations with missing values in the predictor
   3. Check whether the missing values are due to a structural problem in the data as the number of missing data is large
   4. Replace the missing values with zero
3. In clustering, what is meant by the “complete linkage” approach?
   1. Linking clusters by average inter-cluster dissimilarity
   2. Linking clusters by maximal inter-cluster dissimilarity
   3. Linking all clusters consecutively
   4. Linking clusters by their intra-cluster dissimilarity
4. The least-squares estimate fits the regression line that...
   1. Minimizes the sum of squared error
   2. Minimizes the mean absolute error
   3. Minimizes the mean squared error
   4. Minimizes the percentage error
5. What is **not** an appropriate method to deal with imbalanced classes in an existing sample? C
   1. Downsampling
   2. Upsampling
   3. LOOCV resampling
   4. SMOTE
6. You are evaluating the output from a multivariate regression model with a large number of predictors. What is the most appropriate general goodness-of-fit measure in this case?
   1. Adjusted R^2
   2. R^2
   3. AUC
   4. p-values of the predictors
7. In support vector machines, C functions as a regularization parameter for the model and is often treated as a tuning parameter when fitting the model. What does a small value of C imply for the model fit?
   1. A “tightly” fitted model with low variance
   2. An optimally fitted model that does not overfit to the training data
   3. A “tightly” fitted model with low bias but high variance
   4. A “loosely” fitted model with high bias and low variance

**[Part 1 Total: 35 marks]**

**Part 2: Open-ended discussion questions**

Answer **ONE** out of the two questions (1) **or** (2).

This section has a total of 65 marks.

Use the **Script Book** provided.

1) Logistic regression and decision trees are two supervised machine learning approaches that can be used to predict data.

1. **Describe** these two methods, and

(35 marks)

1. **Compare** them, **discussing** their advantages and potential pitfalls using examples for situations in which each of them would be the more appropriate choice.

(30 marks)

2) Building and training machine learning models comes with pitfalls that can be mitigated with the right techniques.

1. Explain what method you suggest to avoid **overfitting** your model to data, and how and in what situations this approach leads to **resampling** techniques.

(35 marks)

1. Discuss, using a practical example of your choice, in which situations **regularizing** a linear model would be a useful approach.

(30 marks)

**[Part 2 Total: 65 marks]**

**[Exam Total: 100 marks]**

**END OF PAPER**