UNIVERSITY OF SURREY ©

Faculty of Engineering and Physical Sciences

Department of Computer Science

MSc Programme in Data Science

Module COMM054

Data Science Principals and Practices

FHEQ Level 7 Examination

Time allowed: Two Hours Semester 1, 2019/20

Answer all **four** questions.

Q1 carries 40 marks, **Q2**, **Q3 and Q4** carry 20 marks each.

Where appropriate, the mark carried by an individual part of a question is indicated in squared brackets [ ]. Calculators are allowed.

Additional Materials: None

Question 1 Multiple-choice questions (Each question carries 2 marks, 40 marks in total):

1) Image data are
   a) Unstructured data
   b) Semi-structured data
   c) Structured data
   d) None of the above.

2) Under which situations should a full factorial design be considered?
   a) When the number of factors is not more than three
   b) When there are strong interactions between the factors
   c) When all factors significantly contribute to the output
   d) All of above.

3) Which of the following statements is true?
   a) Central composite design works efficiently when there is a large number of factors
   b) Latin hypercube sampling can produce a more accurate mean of a distribution than Monte Carlo sampling with a small number of samples
   c) Taguchi Method works well when there are strong interactions between all factors
   d) None of above.

4) Which of the following statements about density based approaches to outlier detection is wrong?
   a) Density based approaches are able to detect local outliers
   b) Density based approaches assume that the density around a normal data object is similar to the density around its neighbours
   c) A data point is considered to be an outlier if the density around it is considerably different to the density around its neighbours
   d) None of the above.

5) Noise in data cannot be reduced by
   a) Regression
   b) Binning
   c) Clustering
   d) Statistical tests.

6) Which of the following is NOT a means for quality control in crowdsourcing:
   a) Allow redundancy
   b) Ask Gold-standard questions
   c) Use qualification tests
   d) Provide monetary rewards.

7) Every learning algorithm has three essential components. What are they?
   a) Confusion Matrix, Greedy Search, Decision Tree
   b) Representation, Evaluation Criteria, Optimization Algorithm
   c) Mean Square Error, Optimization Algorithm, Linear Model
   d) Confusion Matrix, Mean Square Error, Accuracy.

8) Under-fitting and over-fitting are pitfalls in data-mining. What are their symptoms?
   a) Under-fitting: Poor performance on both train and test data. Over-fitting: Excellent performance on train data, poor performance on test data.
   b) Under-fitting: Excellent performance on train data, poor performance on test data. Over-fitting: Poor performance on both train and test data.
   c) Under-fitting: Poor performance on both train and test data. Over-fitting: Excellent performance on test data, poor performance on train data.
   d) Under-fitting: Excellent performance on test data, poor performance on train data. Over-fitting: Poor performance on both train and test data.

SEE NEXT PAGE

9) If you perform model-selection or regularization parameter tuning by optimising performance on your training data, what is the expected outcome?

a) A well trained model.
b) Over-fitting.
c) Under-fitting.
d) Program crash.

10) Regression models typically use mean-squared error (MSE) as the optimization criterion. What is one potential drawback of this?

a) It is slow to compute.
b) It needs more data points than dimensions.
c) It is not robust to many dimensions.
d) It is not robust to outliers.

11) What is the main aim of pruning in Decision Tree learning?

a) To improve the training speed.
b) To improve the testing speed.
c) To reduce the memory requirement.
d) To reduce overfitting.

12) Which of the following is NOT a reason to do feature selection?

a) To reduce the processing requirement of a subsequent computation.
b) To reduce the memory requirements of a subsequent computation.
c) To get rid of irrelevant dimensions.
d) To get rid of outlier instances that cause overfitting.

13) Which of the following are NOT reasons to do clustering

a) To gain domain knowledge by discovering natural groups in the data
b) To model normal data before performing outlier/anomaly detection.
c) To find similar data points to a particular data point of interest.
d) To learn a predictive model for category given labelled training data.

14) The statement that “$P(A|B) = P(B|A)$ whenever A and B are independent events” is:

a) Always True
b) Never True
c) Not enough information: we need to know if A and B are disjoint events
d) Not enough information: we need to know if the events are equally likely.

15) Events A and B are independent and it is known that P(A|B) = 0.2 and P(B|A) = 0.5. The probability P(A ∪ B) is:

a) 0.7
b) 0.6
c) 0.4
c) 0.1

16) The arrival of buses at the Stag Hill campus can be modelled with a Poisson process having intensity $\lambda = 8$ (buses per hour). Given that no bus arrived in the last half an hour, compute the probability that in the next half an hour 3 buses will arrive:

a) $\frac{e^{-4}\,8^4}{4!}$
b) $\frac{e^{-4}\,4^3}{3!}$
c) $\frac{e^{-8}\,8^3}{3!}$
d) $\frac{e^{-8}\,3^4}{4!}$

17) Which of the following points would Bayesian and frequentists disagree on?
a) The use of a non-Gaussian noise model in probabilistic regression.
b) The use of probabilistic modelling for regression.
c) The use of prior distributions on the parameters in a probabilistic model.
d) The idea of assuming a probability distribution over models.

18) A result is called “statistically significant” whenever
a) The null hypothesis is true.
b) The alternative hypothesis is true.
c) The p-value is less or equal to the significance level.
d) The p-value is larger than the significance level.

19) Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting:
a) Increase the amount of training data.
b) Improve the optimisation algorithm being used for error minimisation.
c) Decrease the model complexity.
d) Reduce the noise in the training data.

20) You are reviewing papers for the NeurlPS (a top machine learning conference), and you see submissions with the following claims. Which one would you consider accepting:
a) My method achieves a training error lower than all previous methods!
b) When the regularisation parameter is chosen so as to minimise test error, my method achieves a test error lower than all previous methods!
c) When the regularisation parameter is chosen so as to minimise cross-validation error, my method achieves a test error lower than all previous methods!
d) When the regularisation parameter is chosen so as to minimise cross-validation error, my method achieves a cross-validation error lower than all previous methods!

Question 2: The following question is about design of experiments in data collection.

2.1 How many experiments (treatments) are required for a 4-level 4-factor full factorial design? Explain why. [5 marks]

2.2 How many experiments (treatments) are required for a 3-level 3-factor central composite design? Explain why. [5 marks]

2.3 Explain the main differences between Monte Carlo sampling and Latin hypercube sampling. How many samples are needed for 5-level 5-factor problem when the Latin hypercube sampling is used?

[5 marks]

2.4 Explain under which situations the factorial design methods, Taguchi's method, or sampling methods should be used. [5 marks]

Question 3 Answer the following questions regarding supervised learning.

3.1 When assessing a supervised learning method, the resulting accuracy of prediction on test data is obviously very important. List five additional factors that are also useful to assess the learning procedure to decide its suitability for a particular application. [5 marks]

3.2 Label the following applications according to their suitability for being treated as a classification or regression problem in supervised learning, or neither. [5 marks]

a) Predicting whether a customer will defect to a competitor.
b) Predicting tomorrow's temperature.
c) Expected profit on a financial transaction.
d) Predictive text on a mobile phone.
e) Person recognition at an access gate.
f) Compressing the number of bytes required to store an image.
g) Steering system on a self-driving car.
h) Road pedestrian detector in a self-driving car.
i) Deciding suitable sizes for a line of t-shirts given height and weight of the population.
j) Airline flight price prediction.

3.3 Linear regression and logistic regression are two methods in supervised learning. Briefly outline their similarities and differences. [5 marks]

3.4 Explain the role of the objective function in machine learning. Illustrate your answer by describing the objective function of your favourite learning algorithm. [5 marks]

Question 4 Suppose that $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}$ is sampled from a uniform distribution on the interval $(0, \theta)$, where the parameter $\theta > 0$ is unknown.
4.1 Write down the likelihood function of θ with respect to $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}$. [10 marks]

4.2 Find the maximum likelihood estimation (MLE) of θ with respect to $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}$? [10 marks]

END OF PAPER