UNIVERSITY *of* LIMERICK

OLLSCOIL LUIMNIGH

Faculty of Science and Engineering
Department of Computer Science and Information Systems

End-of-Semester Exam

Academic Year:	2017/2018	**Semester:** Autumn
Module Title:	Data Mining and Data Warehousing	**Module Code:** CS4055
Exam Duration:	2 Hours	**Total Marks:** 100
Lecturer:	Dr. N. S. Nikolov	(Equal to 65% of the final grade)

Instructions to Candidates:

This exam consists of 2 sections. Read the instruction for each section at the beginning of the section. Submit the exam paper together with the answer booklet!

SECTION A: Multiple Choice Questions – answer all questions 50 marks

Instructions: Answer all questions in Section A. Each question has exactly one correct answer. Answer each question by circling the correct answer directly in this exam paper. Submit this exam paper together with the answer booklet. Each question is worth **2 marks**.

Please write your ID number below.

Student ID Number: ____

Q1. According to Witten and Frank, data mining is the extraction of implicit, ____________________ and potentially useful information from data.

a. structured
a. visual
b. previously unknown
c. predictive

Q2. Not removing outliers from a dataset can increase the chance for building a/an __________ model.

a. overfitted
b. accurate
c. descriptive
d. deployable

Q3. In a dataset represented as a table of data, the features are the:

a. rows
b. attributes
c. instances
d. outliers

Q4. The confidence in the accuracy of a predictive model is generally higher when the training, validation and test datasets are:

a. identical
b. discretized
c. overlapping
d. not overlapping

Q5. A typical model building algorithm assumes that the instances in the training dataset are

a. sorted in ascending order
b. recursively related to each other
c. independent from each other
d. linearly separable

Q6. Cross-validation is typically used when the training dataset is:

a. relatively small
b. too big to be handled efficiently
c. unstructured
d. noisy

Q7. Classification and numeric prediction are both ___________________ styles of machine learning.

a. clustering
b. declustering
c. supervised
d. unsupervised

Q8. Which one of the following is NOT a classification algorithm?

a. OneR
b. ID3
c. Logistic regression
d. Apriori

Q9. A widely used method for numeric prediction is:

a. Logistic regression
b. Linear regression
c. ID3
d. Naïve Bayes

Q10. Naïve Bayes assumes that all attributes/columns in a table of data are

a. normalised
b. correlated
c. equally important
d. numeric

Q11. The extended version of ID3 for handling numeric data is known as

a. SVM
b. ZeroR
c. C4.5
d. Max-Miner

Q12. The parameter k in the clustering algorithm k-Means is

a. The minimum distance between two adjacent clusters
b. The number of clusters
c. The maximum number of iterations
d. The number of data items in a cluster

Q13. A central concept in instance-based learning is

a. Standard deviation
b. Squared mean error
c. Entropy
d. Distance function

Q14. Association learning algorithms are typically employed for

a. market basket analysis
b. ranking prediction
c. cross-validation
d. matrix factorization

Q15. The support of the association rule **if x=a then y=b** is the number of instances in the dataset for which

a. x=a
b. y=b
c. x=a and y=b
d. x=a or y=b

Q16. Logistic regression assumes that classes are

a. linearly independent
b. linearly separable
c. logistic
d. regressive

Q17. The Perceptron algorithm trains a

a. single-layer neural network
b. deep network
c. random forest
d. list of decision rules

Q18. The prime use of a validation dataset is for

a. faster prediction
b. parameter tuning
c. feature selection
d. removing outliers

Q19. The kappa statistic tells how well a model performs compared to

a. a model with untuned parameters
b. a linear model
c. random prediction
d. Bayesian prediction

Q20. The accuracy of a numeric prediction model can be measured by calculating the

a. True positive rate
b. Squared mean error
c. F-measure
d. Precision and recall

Q21. Assume a classification model does not perform well when applied to a well-designed test dataset. Which one of the following actions is the least appropriate?

a. User cross-validation instead of the test set
b. Attempt an alternative model building algorithm
c. Select a different set of features and re-build the model
d. Remove outliers from the training dataset and re-build the model

Q22. TensorFlow can be described as a

a. Visual analytics dashboard
b. Software library for training deep networks
c. Platform for social network analysis
d. Face recognition algorithm

Q23. Visual Analytics is the science of analytical reasoning supported by:

a. multidimensional scaling
b. highly interactive visual interface
c. graphic design
d. visual clustering

Q24. Convolutional neural networks are best known for their success in

a. text clustering
b. market basket analysis
c. ranking prediction
d. image classification

Q25. One of the biggest challenges in training deep networks is the

a. random neuron problem
b. overlapping of patterns
c. vanishing gradient
d. convolution

END OF SECTION A

SECTION B: Problem Solving – answer all questions 50 marks

Consider the dataset in Table 1 as a training dataset and write the answers to Q26-Q29 in the answer booklet.

Table 1

a	b	class
0.00	0.33	yes
0.07	0.50	yes
0.20	0.00	yes
0.20	0.17	yes
0.33	0.17	yes
0.68	0.25	no
0.68	0.08	no
0.87	1.00	no
0.87	0.83	no
0.93	0.92	no
1.00	0.83	yes
1.00	1.00	yes

Q26. Based on attributes **a** and **b** only partition the 12 instances into four clusters. Explain your decision. **2 marks**

Q27. Assuming bucket size 3, what rule would the OneR algorithm discover for predicting the value of attribute **class**? **16 marks**

Q28. Consider a k-NN learning scheme that uses city-block distance and k=3 for predicting the value of attribute **class**. Would instance **(a=0.94, b=0.9)** be classified as **yes** or **no**? Explain why. **16 marks**

Q29. Assume both attributes **a** and **b** are discretized by representing all values below 0.85 as **low** and all values above 0.85 as **high**. Would Naïve Bayes classify the instance **(low, high)** as **yes** or as **no**? Explain why. **16 marks**

END OF EXAM