INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN

Revision Aid

|  |  |
| --- | --- |
| **Year** | Year 3 |
| **Semester** | Semester 1 |
| **Date of Examination** |  |
| **Time of Examination** |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Prog Code** | BN302 | **Prog Title** | Bachelor of Science in Computing in Information Technology | **Module Code** | Comp H3018 |
| **Prog Code** | BN013 | **Prog Title** | Bachelor of Science in Computing in Information Technology | **Module Code** | Comp H3018 |
| **Prog Code** | BN104 | **Prog Title** | Bachelor of Science (Honours) in Computing in Information Technology | **Module Code** | Comp H3024 |
| **Prog Code** | BN311 | **Prog Title** | Bachelor of Science in Computing in Information Security and Digital Forensics | **Module Code** | ISDF H3018 |

|  |  |
| --- | --- |
| **Module Title** | Data Mining |

**Internal Examiner(s): *Geraldine Gray***

**External Examiner(s): *Mr Michael Barrett, Dr Tom Lunney***

**Instructions to candidates:**

1. **To ensure that you take the correct examination, please check that the module and programme which you are following is listed in the tables above.**
2. **Question One Section A is COMPULSORY. Candidates should attempt Question One and ANY other two questions in Section B.**

Revision Aid

1. **This paper is worth 100 marks. Question One is worth 40 marks and all other questions are worth 30 marks each.**

**DO NOT TURN OVER THIS PAGE UNTIL YOU ARE TOLD TO DO SO**

The following questions are a revision aid for the Data Mining exam. It is not an exhaustive list of all possible questions that might appear on the exam, but gives an indication of the layout of the paper, and the types of questions to expect.

Reciting lecture note material is NOT sufficient as an answer to questions. **Answers should be in your own words, and when asked for an example, use your own example rather than one from the notes.**

Marking schemes are generally based on 1 mark per point made, provide the point is valid and in your own words.

Note: The order of questions may differ in the actual exam.

**BRING A CALCULATOR**

**SECTION A**

**Question 1. Mandatory Questions. 10 short questions, each worth 4 marks.**

This question will cover the entire course: CRISP-DM; Data Exploration; Data Preparation; Classification and Clustering

Examples of short questions taken from past papers:

1. Define the terms classification and clustering. Provide one application for each of these data mining techniques.
2. Data Mining is an important concept in many industry sectors. Give two examples of applications and briefly outline the term *Data Mining.*
3. *The CRISP-DM* methodology is the de facto standard for data mining projects. Outline each of the stages/processes in CRISP-DM.
4. Explain each of the CRISP-DM processes with regard to their importance of the various data mining project stages.
5. Differentiate between Discrete, and Continuous data. Provide two examples for each category and outline the main characteristics.
6. Explain the terms *Nominal* and O*rdinal* as well as *Interval* and *Ratio* in terms of data types.
7. Explain the terms *binary*, *discrete* and *continuous* in terms of data types.
8. Outline four different data attribute types and provide one example for each type.
9. Visualisation is an important concept in data mining. Briefly explain why. Draw a box plot and explain its characteristics.
10. Explain the characteristics of *Box Plots* and *Scatter Plots*.
11. Explain how the classification process takes place in a data mining project?
12. Explain the role of the test and training datasets.
13. Define the terms underfitting and overfitting. How can these two adverse effects be eliminated?
14. Briefly explain the terms *Precision*, *Accuracy* and *Bias.*
15. Define *Noise* and *Outliers*. Distinguish between *Noise* and *Outliers*
16. *Explain the terms precision and recall, and why you need to consider both.*
17. Compare a decision tree with a neural network in terms of accuracy, training time, the type of pattern that can be found and how easy the output is to interpret.
18. Explain the concept of *Artificial Neural Networks*.
19. k Nearest Neighbour is a classification technique. Outline the reasoning and pitfalls when choosing the correct value for k.
20. Outline why information gain might be calculated when generating a decision tree.
21. Explain the terms pre-pruning and post pruning with respect to training a decision tree.
22. Outline the difference between partitional and hierarchical clustering. Give one example each.
23. What is the *Curse of Dimensionality*? Explain how this can be addressed.
24. Missing values are a common occurrence in data sets. Explain how *missing values* can be handled?
25. Why might it be necessary to discreatise data before mining? Explain how this would be done.
26. Why might it be necessary to normalise/scale data before mining? Explain how this would be done.

**SECTION B**

**Question 2. Data Preparation techniques.**

(Note: these techniques could also be relevant to answers in other questions)

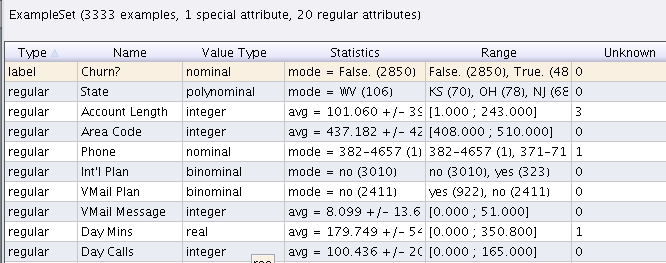


Figure 1. Dataset representing customer phone call patterns

1. Given the meta data displayed in figure 1, recommend THREE pre-processing techniques that would be appropriate to use on this dataset. In your answer, explain both why that technique would be advisable, and what would be the effect of applying the technique.

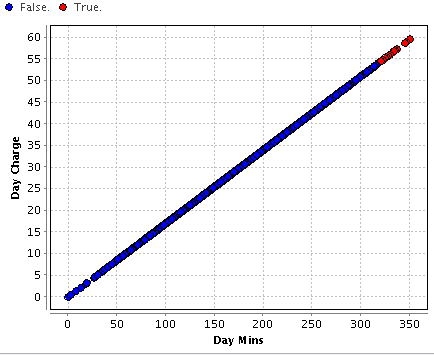
**12 marks**

1. Explain each of the data types listed in the meta data above. **5 marks**
2. Briefly explain the role of the data exploration phase of a data mining process

**3 marks**

1. Interpret the following three plots generated from the dataset in figure 1. **10 marks**

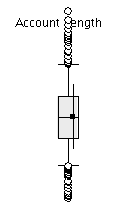
(i) Scatter plot of **day minutes** versus **day charge**, overlaid with the binary class label, **churn** (all true values of the class label are on the top right corner).



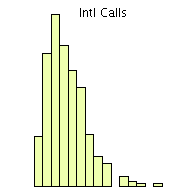
False

True

(ii) Box plot on **account length**



(iii) Histogram of **International calls made**.



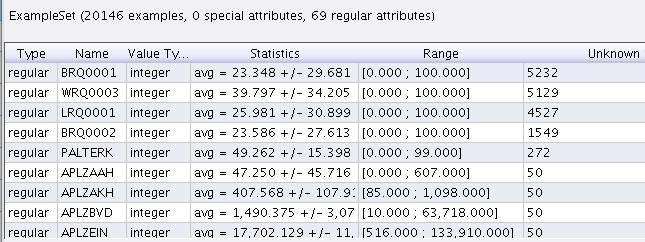


Figure 2

1. Given the meta data displayed in figure 2, recommend THREE pre-processing techniques that would be appropriate to use on this dataset. In your answer, explain both why that technique would be advisable, and what would be the effect of applying the technique.

**12 marks**

1. Explain the entries under **statistics** and **range** in the meta data above. Why is this information useful?

**12 marks**

1. Recommend two ways to visualize the data in the dataset above, explaining what information is portrayed by each.
2. Classify the following attributes as binary, discrete, or continuous. Alsoclassify them as qualitative (nominal or ordinal) or quantitative (interval orratio). Some cases may have more than one interpretation, so briefly indicateyour reasoning if you think there may be some ambiguity.
   1. Time in terms of AM or PM.
   2. Brightness as measured by a light meter.
   3. Brightness as measured by people’s judgments.
   4. Angles as measured in degrees between 0◦ and 360◦.
   5. Bronze, Silver, and Gold medals as awarded at the Olympics.
   6. Height above sea level.
   7. Number of patients in a hospital.

(*7 marks*)

1. Define the terms *Noise* and *Outliers*. Distinguish between *Noise* and *Outliers*.

(*7 marks*)

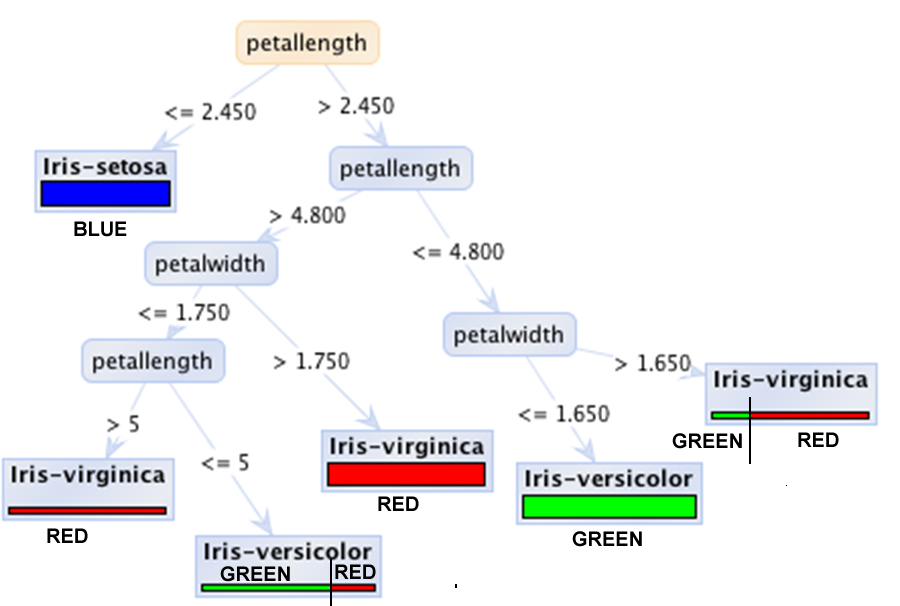
**Question 3. Classification**

1. Outline two methods for generating training and test datasets, and explain which is the better choice.
2. Explain what is meant by model **over-fitting**. Your answer should discuss causes of model over fitting and how the problem can be addressed when building a **decision tree**.
3. Explain how an **impurity measure** can be used to decide on spilt points in a decision tree. **4 marks**
4. Using the data given below, calculate an impurity measure for **student** and **gender**. **8 marks**
5. Based on your results in part (ii) above, advise which attribute should be used as the next split point for the data given.

**2 marks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student** | **Gender** | **Label** |  |
| yes | M | high |
| yes | F | low |
| yes | M | low |
| no | M | high |
| no | F | high |
| no | F | high |
| yes | M | low |

**8 marks**



1. Answer the following questions based on the graphical representation of a decision tree given above:

* What is represented by leaf nodes?
* What is represented by non-leaf nodes?
* What is represented by branches?
* How does the graphical representation illustrate the number of examples incorrectly predicted? Explain your answer.

**5 marks**

1. For each of the following, state how the row of data would be classified by the decision tree above, and how confident you are in the prediction.

|  |  |
| --- | --- |
| **Petal Length** | **Petal Width** |
| 5.2 | 1.7 |
| 4.7 | 1.8 |

**2 marks**

1. Which is the most predictive attribute in the tree above? In your answer, give an overview of how an **impurity measure** selects the most predictive attribute. Illustrate your answer with an example.

**5 marks**

1. Discuss why one might consider **pruning** a decision tree, and what is the benefit of pruning.

**5 marks**

1. Performance evaluation of the predictive capability of a classification algorithm is a large aspect of data mining models. Describe in detail how performance can be evaluated using metrics.

(*8 marks*)

1. Explain how a **nearest neighbour** classifier works.

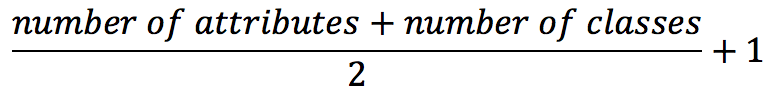
**5 marks**

1. Use the training dataset below to determine whether or not k-NN would predict the correct label for the row of data if k=1 and if k=3.

**Training Data:**

|  |  |  |  |
| --- | --- | --- | --- |
| Number of texts sent per day | Pre-pay or Bill Pay | Uses roaming for more than 20 days a year | **Churn** |
| 5 | Pre | Yes | Yes |
| 3 | Bill | No | Yes |
| 1 | Pre | No | No |
| 0 | Pre | No | No |

|  |  |  |  |
| --- | --- | --- | --- |
| **Test data** |  |  |  |
| Sends more than 5 texts per day | Pre-pay or Bill Pay | Uses roaming for more than 20 days a year | **Churn** |
| 4 | Bill | Yes | Yes |

1. How would you configure a Neural Network for each of the dataset characteristics below? Use the formula below for the hidden layer.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dataset characteristics: | | How many neurons in: | | |
|  | **Number of numeric Attributes** | **Number of classes** | **Input layer** | **Hidden Layer** | **Output Layer** |
| Dataset 1: | 50 | 5 |  |  |  |
| Dataset 2: | 100 | 2 |  |  |  |
| Dataset 3: | 20 | 3 |  |  |  |
| Dataset 4: | 50 | a continuous variable |  |  |  |

1. What weights would be allocated initially? And briefly explain how those weights are adjusted.
2. Given the row of data below, and the start state for the network, explain how the first prediction is calculated. Assume the network is NOT using an activation function.

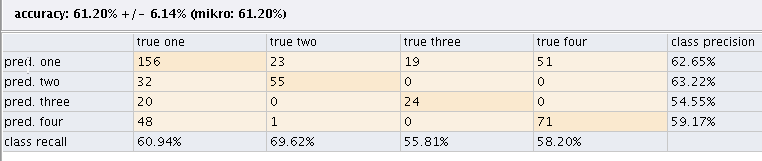
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Neural Network with three inputs, and one output** | | | | | | | | |
| Training Data | |  |  |  |  |  |  |  |
| A | B | C | Y |  |  |  |  |  |
| 0.5 | 0.7 | 0.2 | 1 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Calculating predicted output: | | | |  |  |  |  |  |
|  | Wa | Wb | Wc | Bias | Estimated output |  | Actual | Error |
| 1st epoch | 0.2 | 0.4 | 0.1 | 0.3 |  | | 1 |  |
| Fill in Estimated output and error | | | | | | | |  |
| Will the error cause network weights being increased or decreased? | | | | | | | |  |

1. The following confusion matrices present the classification results that were produced using a decision tree and a neural network in the prediction of churn for an insurance company. Which data mining algorithm performs the best for this particular task? Discuss each evaluation measure you use to assess the performance of each model.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | Decision Tree | PREDICTED CLASS | | | | ACTUAL CLASS |  | + | - | | + | 150 | 40 | | - | 60 | 250 | |  | |  |  |  |  | | --- | --- | --- | --- | | Neural Network | PREDICTED CLASS | | | | ACTUAL CLASS |  | + | - | | + | 250 | 45 | | - | 5 | 200 | |
|  |  |  |

**(*9 marks*)**

1. Using the confusion matrix from part k) above, calculate the overall accuracy for the decision tree model, and the precision and recall for the positive class. Explain the terms **precision** and **recall**. **(6 marks)**

****

1. Interpret the confusion matrix given above. The class label has four possible values, namely ‘one’, ‘two’, ‘three’ and ‘four’. In your answer explain the individual cell entries, the class **precision** entries and the class **recall** entries.

**6 marks**

1. Give the a confusion matrix for the data given below. The class label is CHURN .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of texts sent per day | Pre-pay or Bill Pay | Uses roaming for more than 20 days a year | **Actual**  **Churn** | **Predicted**  **Churn** |
| 5 | Pre | Yes | Yes | Yes |
| 3 | Bill | No | Yes | No |
| 1 | Pre | No | No | No |
| 0 | Pre | No | No | No |

**Question 4. Clustering**

1. Calculate the Manhattan & Euclidean Distance of the points shown in the table below. What needs to be done if the scales of attributes differ?



Note:

(*8 marks*)

1. Calculate the Manhattan & Euclidean distance between the following three data points:

Note:



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute 1** | **Attribute 2** | **Attribute 3** |
| Point a | 10 | 2 | 100 |
| Point b | 15 | 3 | 70 |
| Point c | 12 | 7 | 90 |

**12 marks**

1. Calculate the relative distance between the following three data points:

Note:



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute 1** | **Attribute 2** | **Attribute 3** |
| Point a | 8 | 2 | 50 |
| Point b | 10 | 4 | 200 |
| Point c | 5 | 1 | 250 |

**12 marks**

1. With reference to the data above, explain how a variables range can be used to weight the variable in terms of their influence on distance calculations.

**4 marks**

1. Explain the benefit of **normalising** attributes before calculating distances. Make reference to the data in part (b) above to illustrate you answer.

**4 marks**

1. Using the data given in part c) above, convert each attribute to a range of [0,1] using min-max normalisation.

**6 marks**

1. Define Clustering and list three applications.

**(*5 marks*)**

1. Explain in detail how the **k-means clustering** algorithm works.

**9 marks**

1. Explain the DBScan clustering algorithm.

**(*5 marks*)**

1. Describe a method that can be used to evaluate k-means clusters.

**(*5 marks*)**

1. Explain the difference between **partitioning** and **density based** clustering.

**5 marks**

1. Differentiate between **Partitional** Clustering and **Hierarchical** Clustering.

**(*5 marks*)**

1. Explain the difference between **overlapping** clusters and **complete** clusters.

**(*4 marks*)**

1. List and briefly explain four different types of clusters.

**(*5 marks*)**

1. Explain why the notion of a **cluster** can be ambiguous.

**5 marks**