

Faculty of Engineering, the Built Environment
and Information Technology
Department of Computer Science
Data-Mining

Exam duration: 2 days

Exam date: 01 December – 2 December

Total marks: 40 marks

Instructions

1. Read the question paper carefully and answer all the questions below. Write a document containing your answers, which is saved or exported as a PDF file. Include your name, surname and student number on the first page of your submission file. Any file format other than a PDF file is unacceptable for submission, and will not be assessed.
2. Submission of this **Assignment** is electronic. A PDF file containing your answers is the only acceptable format, and must be uploaded on the course ClickUP page via a Turnitin assignment submission named “**Assignments**”, accessible from the same folder from which you downloaded this document. The upload deadline is **Wednesday 2 December 2020 at 23:00**. Half an hour grace period will be given to accommodate technical difficulties, and absolutely no late submissions will be accepted.
3. You may consult any sources you wish, including any additional articles you can find. Include accurate and complete references for all the external sources you refer to. Do not include text taken directly from any sources, even if you quote it correctly. Excessive direct quoting will be penalised. Paraphrase your answers, and give your own opinion. Note that not all sources are reliable or correct, even if they are published in conference proceedings or a journal (in other words, be critical of every source you consult). You may not consult with any other individual with regard to any part of this **Assignment**. Any such external consultation is considered plagiarism, and will not be tolerated.
4. Note that the above point does not imply that everything you mention must come from a source. Many of the questions test your interpretation and understanding of concepts covered in the course content, or related to the course content. Well reasoned arguments will receive credit. Conversely, unsubstantiated or dubious observations from cited sources will not receive credit.
5. By submitting your answers you implicitly agree to the plagiarism policy of the Computer Science Department at the University . This policy is repeated below, for your convenience. If you are found guilty of any plagiarism, disciplinary action will be taken. This may include suspension of your studies at the University of Pretoria. Note that your submission will be tested for plagiarism using the Turnitin system. Your submission will not be permanently uploaded into the Turnitin database.
6. Write clearly and concisely. The duration of the **Assignment** is intended to allow you to carefully consider your answers, and does not imply anything about the length of your submission. If you are asked to discuss multiple points, use bulleted lists to clearly organise your answer. Be sure that the numbering of your answers is correct. If a question asks you to provide a certain number of items (e.g. a certain number of advantages or disadvantages), you will be penalised if you provide more than the required number.

Plagiarism Policy

The Department of Computer Science considers plagiarism to be a serious offence. Disciplinary action will be taken against students who commit plagiarism. Plagiarism includes copying someone else's work without consent, copying a friend's work (even with consent) and copying material from the Internet. Copying will not be tolerated in this module.

Question 1: Introductory Concepts [5 marks]

1. A spatial database can be represented in either a raster or vector format. **Identify** one general complexity associated with performing data mining on a raster format spatial databases, which is not associated with performing data mining on a vector format spatial database. [1]
2. Decision trees and production rules can be used to classify data samples. **Describe** only one general drawback associated with the use of traditional decision trees for classification, which is not associated with the use of traditional production rules for classification. **Explain** both why the drawback is associated with decision trees, and why it is not associated with production rules. [2]
3. One complexity associated with incremental mining is that the model created by the mining process should progressively forget stale and irrelevant patterns. **Identify** only one additional complexity associated with incremental mining. [1]
4. **Suggest** only one possible drawback that could be associated with the re-integration of extracted knowledge into the data source, as part of the KDD (knowledge discovery in databases) process. [1]

Question 2: Data Preparation [5 marks]

1. One simple approach to replacing a missing value for a continuous attribute in a data sample is: [2]
 - Determine the class of the data sample with the missing value.
 - Compute the mean of the missing attribute value over all data samples that also have this class.
 - Replace the missing attribute value with this mean.

Identify only one situation in which this approach to missing value replacement would be inappropriate, and **justify** your answer.
2. One problem associated with data integration is the entity identification problem, in which difficulties may arise when trying to identify which attributes are equivalent to one another in two or more data sets that need to be combined. Metadata describing the nature of attributes in the various data sets may assist in the identification process. Imagine that two data sets need to be integrated and the entity identification problem arises, but no metadata is available. **Suggest** only one approach that could be used to resolve the entity identification problem by using only the data sets themselves (i.e. with no external assistance). [1]
3. Feature construction refers to the construction of new attributes from existing attributes in a data set. Imagine that a data set contains general demographic data for the population of an entire country, and has an attribute named **pregnancies**, which indicates how many times an individual has been pregnant. Furthermore, a data analyst decides to construct a new categorical attribute named **sex**. The analyst writes a script that sets the **sex** attribute value for an individual to **female** when the number of pregnancies for the individual is one or more, and sets the value to **male** when the number of pregnancies is zero. **Suggest** only two general drawbacks that are associated with this approach. [2]

Question 3: Exploratory Data Analysis [5 marks]

1. A very commonly used visual dimension in data visualisations is colour. **Identify** only two drawbacks that are specifically associated with using colour as a visual dimension. [2]
2. Exploratory data analysis focuses on data visualisation. **Briefly explain** how a scientific visualisation could be converted into a data visualisation for exploratory data analysis purposes. [2]
3. An area chart is a data visualisation technique. **Identify** only one drawback associated with area charts. [1]

Question 4: Data Clustering [5 marks]

1. **Identify** only one complexity associated with divisive hierarchical clustering algorithms, which is not associated with agglomerative hierarchical clustering algorithms. [1]
2. Consider the k -means and k -medoids clustering algorithms. Unseen data samples are data samples that were not used by the clustering algorithm to construct clusters, but still have the same underlying characteristics as the data samples used by the algorithm. We say that a clustering algorithm generalises well if the clusters produced by the algorithm correctly represent any unseen data samples. **Identify** whether the k -means algorithm or the k -medoids algorithm is likely to generalise better, and **justify** your answer. [2]
3. Consider the cluster quantisation error quality measure. **Identify** only one situation in which the cluster quantisation error will produce an inaccurate assessment of cluster quality, and **justify** your answer. [2]

Question 5: Decision Trees [5 marks]

1. Most decision tree building algorithms use a greedy approach. This means that once an attribute test has been chosen, the algorithm cannot backtrack to change the attribute test. Answer the following questions:
 - (a) **Briefly describe** only one advantage of this greedy approach. [1]
 - (b) **Briefly describe** only one disadvantage of this greedy approach. [1]
2. One advantage of decision tree pruning is that it can result in more compact trees that are more understandable to human data analysts. **Identify** another possible advantage associated with decision tree pruning, and **briefly explain** why this advantage exists. [2]
3. The C4.5 algorithm can convert decision trees into rule sets. One of the final steps in this conversion process groups rules into groups (where each group contains all the rules predicting the same class), and orders these groups in ascending order of the number of false positives generated by the group. **Briefly explain** why this ordering is sensible. [1]

Question 6: Rule Induction [5 marks]

1. The AQR algorithm repeatedly selects a positive example called a seed. A seed is used to generate complexes by repeatedly choosing a negative example (i.e. an example that is incorrectly covered by a complex, because its class does not match the class of the seed). The algorithm then specialises the complexes to exclude the negative example. AQR chooses the negative example that is closest to the seed. **Briefly explain** why this choice of a negative example is a good strategy. [2]
2. When the CN2 algorithm specialises complexes using a continuous attribute, the attribute's range is divided into equal sized sub-ranges. The size of sub-ranges is a single user-defined parameter for the algorithm. A set of selectors is then constructed, where each selector tests whether the attribute's value either exceeds or does not exceed a value at a sub-range boundary. **Identify** only two drawbacks to this approach. [2]
3. The CN2 algorithm prunes the worst complexes from the STAR set at the end of each series of complex specialisations. The amount of pruning that takes place is controlled by a user-specified limit on the size of the STAR set. A smaller size limit will result in more pruning, while a larger size limit will result in less pruning. **Briefly explain** how the CN2 algorithm's search for a best complex will be affected by choosing a smaller size limit on the STAR set. [1]

Question 7: Self-Organising Maps [5 marks]

1. The learning rate and neighbourhood radius of a self-organising map should both decrease as training continues. **Briefly explain** what the result will be if these values do not decrease in this way. [1]
2. **Identify** only one disadvantage associated with weight-centric neuron labelling for self-organising maps. [1]
3. In the context of self-organising maps, unsupervised example-based neuron labelling can be seen as an unsupervised version of example-centric neuron labelling. **Identify** two reasons that unsupervised example-based neuron labelling is much more complex for a human data analyst to perform than example-centric neuron labelling. [2]
4. **Briefly describe** only one drawback associated with the SIG* algorithm. [1]

Question 8: Ant Algorithms [5 marks]

1. In the AntMiner algorithm a quality function computes a numeric value that represents the quality of a rule, as follows:

$$Q = \frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}$$

where TP is the number of true positives for the rule, FP is the number of false positives for the rule, TN is the number of true negatives for the rule, and FN is the number of false negatives for the rule, all computed over the entire training set. Answer the following questions:

- (a) **Briefly explain** what the fraction $\frac{TP}{TP+FN}$ represents. [1]
 - (b) **Briefly explain** what the fraction $\frac{TN}{FP+TN}$ represents. [1]
2. The AntMiner algorithm has several similarities to a decision tree building algorithm, but tends to deal with attribute interactions better. **Briefly explain** why AntMiner deals with attribute interactions better. [2]
 3. In the ant-based clustering algorithm proposed by Deneubourg *et al*, the probability of an unladen ant to pick up an item is calculated as: [1]

$$P_p = \left(\frac{k_1}{k_1 + f} \right)^2$$

where k_1 is a constant and f is the perceived fraction of items in the ant's neighbourhood. **Briefly explain** what the effect is when the value of k_1 is increased.