Computing @ Curtin University

FINAL ASSESSMENT

End of Semester 2, 2020

COMP5009 Data Mining

This paper is for Bentley Campus students.

This FINAL ASSESSMENT has a total of 100 marks.
Working time: 120 minutes
Reading time: 10 minutes
Download/upload time: 15 minutes

Conditions

- This is a take-home assessment. The test is open-book: you are allowed to access your hand-written notes, lecture slides, textbooks, and printed and electronic materials in your possession.
- The test must be completed by yourself only. No one else should do this test for you.
- You are not allowed to access the Internet during the test for reasons other than to access this test and for emailing the Unit Coordinator for questions and/or technical issues.
- You are prohibited from communicating with people other than the Unit Coordinator during the test.
- You are prohibited from providing information about your work to others during the test.
- You must not post or share this test during and after the test. Please respect the copyright.
- Any attempts to compromise the system are strictly prohibited.
- You must complete and submit the ‘‘Declaration of Originality” form as instructed by your Unit Coordinator/Lecturer for this test.
- **IMPORTANT** Your answers must be your own words. You are not allowed to copy any texts from other sources (including lecture/tutorial materials), even with correct referencing, and present as your own. Copying of texts will be considered **plagiarism**.
- Any breaches of this policy will be considered cheating and appropriate action will be taken as per University policy.

Instructions to Students:

Please read the following instructions carefully before commencing your test. Ignoring of the instructions may make your test invalid.

- This test consists of **7 questions**. Attempt ALL questions.
- You must use the answer file provided. Before commencing your test, you must name your answer file using your surname and student ID, for example `trump_12345678.txt`.
- Record your surname and student ID in the answer file.
- Provide your answer in the space **below** each question heading, e.g.

  ## QUESTION 1, **## QUESTION 2**, etc.

- Do not modify the structure or the question headings. They are important for the processing of your test.
- Do not copy the entire questions into the answer file. You are only required to provide the answer to each question.
- The answer file is a plain-text file. You must not convert it to other types. Use a text editor if possible. If you are using a Word document processor such as Microsoft Word, make sure you save it in plain-text format (*.txt). Do not insert any special characters or objects outside the characters on your keyboard - they can be lost during the processing of your test.
- To avoid loss of your work, make sure you regularly save your file and consider a suitable backup option.
- When finished, upload your answer file to Turnitin as instructed. You are allowed only **once** submission.
- You are responsible for ensuring that the submission is correct and free of errors.

Question 1 (12 Marks)

Data mining is about discovering knowledge. Give four (4) examples of such knowledge discovered from the classification task in the assignment. For each example, briefly explain how you acquired the knowledge. Your answer must be specific to the assignment and should not include examples of knowledge derived from the data preparation step.

Question 2 (10 Marks)

A data mining student is comparing two classifiers A and B on a balanced data set containing two classes. The confusion matrix of each classifier when evaluated on the validation set is given below

CLASSIFIER A	Classified As	
Actual Class	Class 1	Class 2
Class 1	86	14
Class 2	15	85

CLASSIFIER B	Classified As	
Actual Class	Class 1	Class 2
Class 1	95	5
Class 2	24	76

- What is the classification accuracy of each classifier on the validation set?
- Which classifier would you prefer on the basis of the classification performance on the validation set? Explain.

Question 3 (12 Marks)

Consider the example below which shows a training set consisting of 10 samples. Each sample $\mathbf{x}_i$ has three attributes and there are two classes Y and N.

Sample	a_1	a_2	a_3	Class
$\mathbf{x}_1$	T	F	1	Y
$\mathbf{x}_2$	T	T	3	N
$\mathbf{x}_3$	T	F	4	Y
$\mathbf{x}_4$	F	T	4	N
$\mathbf{x}_5$	F	F	5	N
$\mathbf{x}_6$	T	T	5	Y
$\mathbf{x}_7$	F	T	6	N
$\mathbf{x}_8$	T	F	7	Y
$\mathbf{x}_9$	F	F	7	N
$\mathbf{x}_{10}$	F	T	8	Y

Classify a data point $\mathbf{x} = (T, F, 4)$ using Naive Bayes. You must show a full derivation which includes at least the following

- The conditional probabilities/densities $P(a_1|Y)$, $P(a_2|Y)$, $f(a_3|Y)$, $P(a_1|N)$, $P(a_2|N)$, $f(a_3|N)$
- The class prior probabilities $P(Y)$, $P(N)$
- The posterior probabilities $P(Y|\mathbf{x})$, $P(N|\mathbf{x})$
- The final prediction

Use the following simplified Gaussian distribution for the numeric attribute a_3

$$f(x) = \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Use the sample mean and sample standard deviation to estimate the mean and standard deviation of a Gaussian distribution. For a set of n samples $x_1, x_2, \ldots, x_n$, the sample mean and sample standard deviation are respectively

$$\hat{\mu} = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

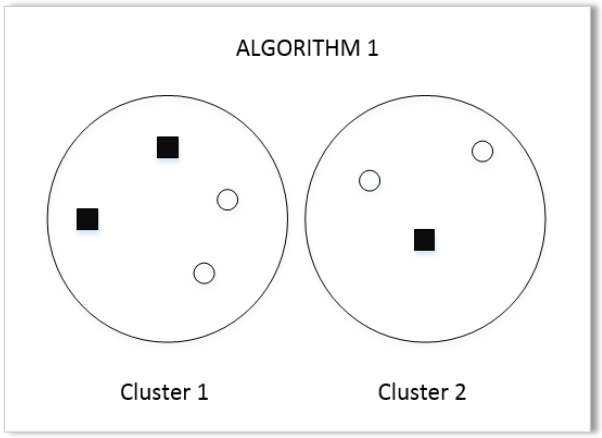
$$\hat{\sigma} = \sqrt{\frac{(x_1 - \hat{\mu})^2 + (x_2 - \hat{\mu})^2 + \ldots + (x_n - \hat{\mu})^2}{n-1}}.$$

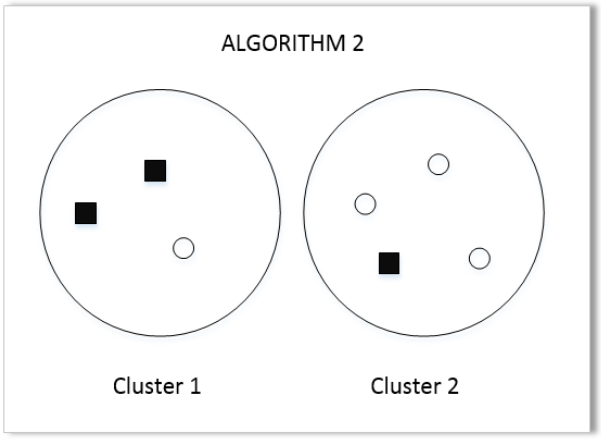
Question 4 (15 Marks)

Two clustering algorithms produce the following results on a data set of samples coming from two classes represented by two different shapes. Which is the better performing algorithm in the following cases

- Case 1: Purity
- Case 2: F_1 measure

Clearly show your derivation. Comment on the results.





You may need to use the following

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

	Same Cluster	Different Clusters
Same Class	TP	FN
Different Classes	FP	TN

Note: Purity refers to the fraction of correctly clustered *samples* over the whole data set, whilst TP, FP, TN, and FN refer to the number of *pairs*. You may express Purity and F_1 measure as a fraction.

Question 5 (15 Marks)

A Data Mining student is clustering the well-known Iris 2D data set using DBSCAN. Each sample in the data set has two numerical attributes representing the petal width and length of the flower. The student has tried the following three sets of DBSCAN hyper-parameters $minpts$ and ε:

- Set 1: $minpts = 6$, $\varepsilon = 0.01$
- Set 2: $minpts = 4$, $\varepsilon = 3$
- Set 3: $minpts = 4$, $\varepsilon = 0.35$

For each of the following clustering results (as shown in Figures 1 to 3), decide which set could have been used to produce it and clearly explain your reasoning. Your argument must be based on the properties of DBSCAN and the values of the data as well as the hyper-parameters.

Note: in the following plots (Figures 1 to 3), the different colours represent different clusters whilst the black circles represent noise points. You should view the plots on a colour monitor.

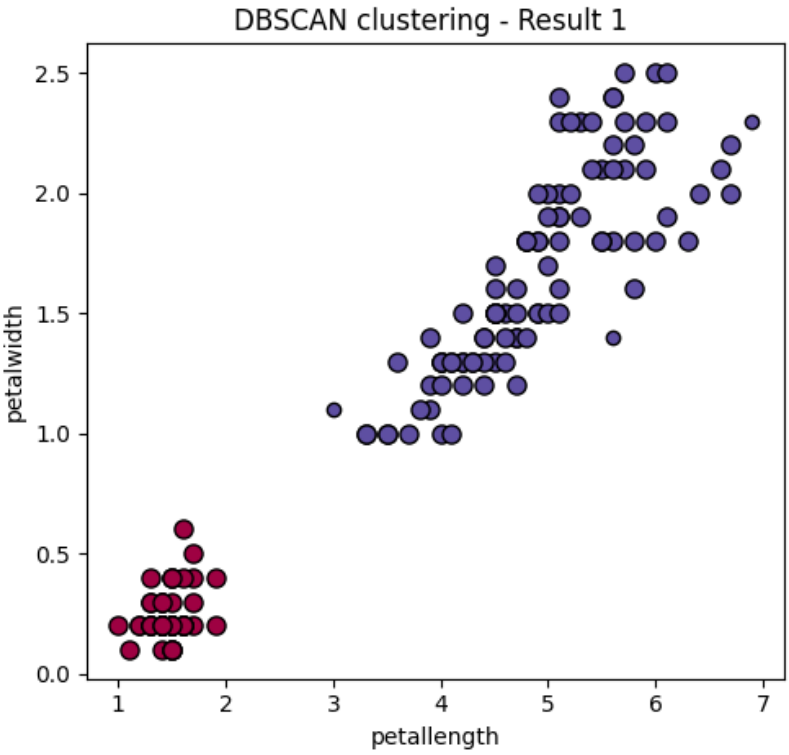


Figure 1: Clustering Result 1

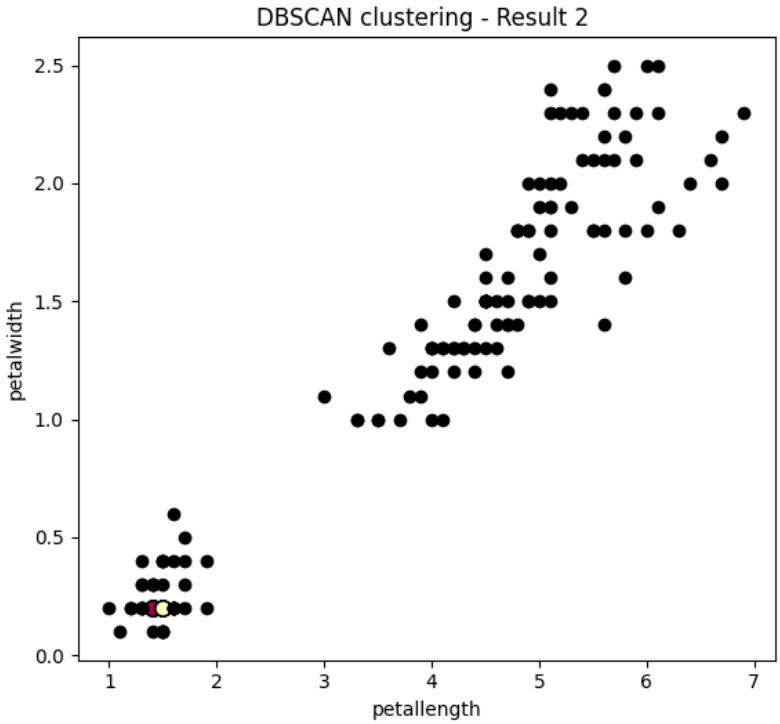


Figure 2: Clustering Result 2

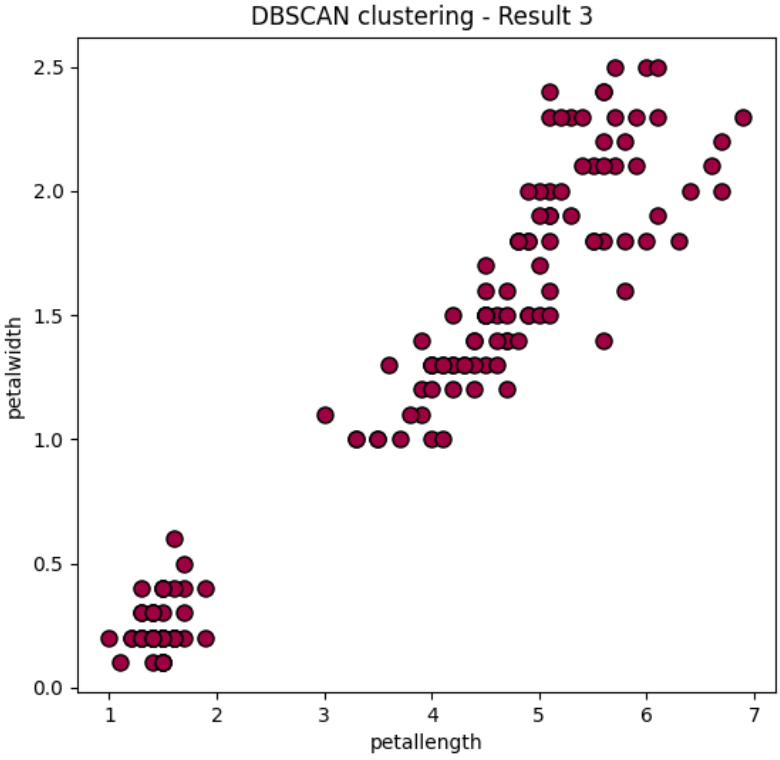
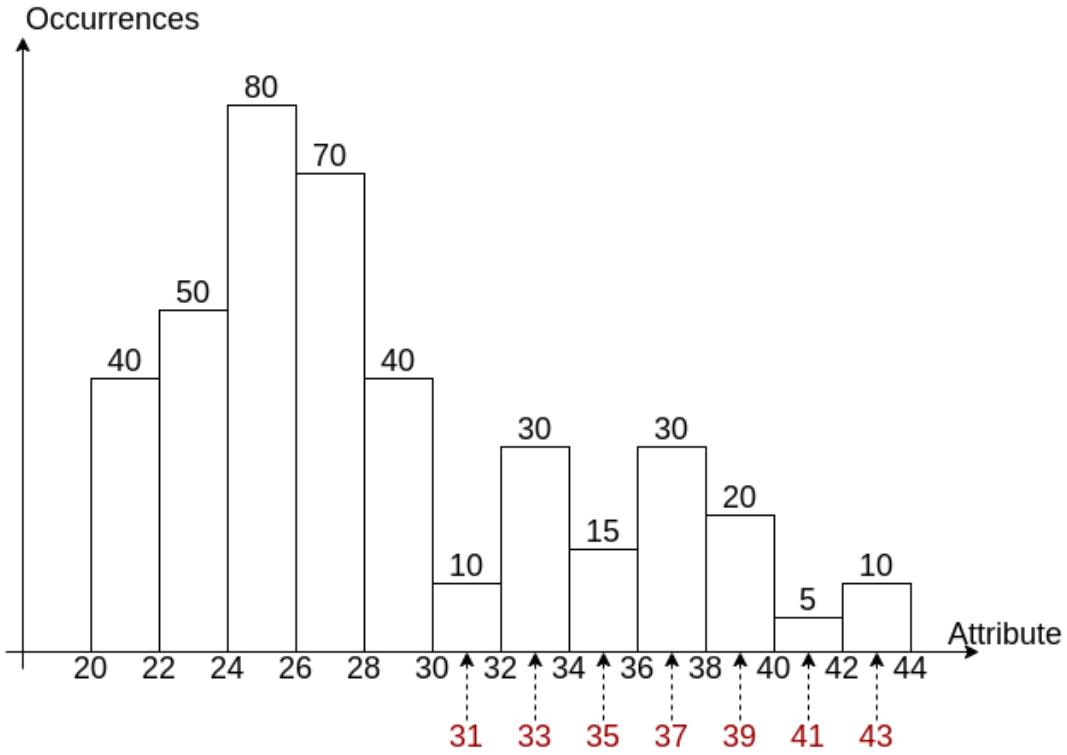


Figure 3: Clustering Result 3

Question 6 (18 Marks)

The histogram of a one-dimensional data set is shown below. The number above each bar represents the occurrences of samples having values in interval of the bar. It is decided that the histogram-based approach is used to detect anomalies, which are defined as the 10% least frequent samples of the whole data set. What would be the prediction (i.e. normal or outlier) for the following data points: 31, 33, 35, 37, 39, 41, 43? Clearly show your derivation.



Question 7 (18 Marks)

Four (4) participants are asked to provide their opinions on six (6) different products. However, one participant in the survey has not provided the rating of product 6. Predict this missing rating using item-based similarity with ratings. Assume that prediction is based on the rating of the most correlated product. Clearly show your full derivation.

	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6
Participant 1	1	2	3	4	5	?
Participant 2	3	4	2	3	4	2
Participant 3	2	3	4	4	3	5
Participant 4	5	4	5	4	1	2

Note: The adjusted cosine similarity between the *normalised* ratings $\bar{U} = (u_1, \ldots, u_s)$ and $\bar{V} = (v_1, \ldots, v_s)$ of a pair of **items** is

$$\mathbf{Cosine}(\bar{U}, \bar{V}) = \frac{u_1 v_1 + \ldots + u_s v_s}{\sqrt{u_1^2 + \ldots + u_s^2}\sqrt{v_1^2 + \ldots + v_s^2}}.$$

END OF FINAL ASSESSMENT