The University of Melbourne

Semester One 2017 Exam

School: Computing and Information Systems
Subject Number: COMP20008
Subject Title: Elements of Data Processing

Exam Duration: 2 hours

Reading Time: 15 minutes

This paper has 11 pages

Authorised Materials:
No calculators may be used.
Instructions to Invigilators:
Supply students with standard script books.
This exam paper can be taken away by the students after the exam.
This paper may be held by the Baillieu Library.

Instructions to Students:
Answer all 5 questions. The maximum number of marks for this exam is 50. Start each question (but not each part of a question) on a new page.

Formulae

Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Pearson's correlation coefficient: $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Entropy: $H(X) = -\sum_{i=1}^{\#categories} p_i \log_2 p_i$
where p_i is the proportion of points in the ith category.

Conditional entropy: $H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$
where $\mathcal{X}$ is the set of all possible categories for X

Mutual information:
$MI(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

Normalised mutual information:
$NMI(X, Y) = \frac{MI(X,Y)}{min(H(X),H(Y))}$

Accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$ where
TP is number of true positives
TN is number of true negatives
FP is number of false positives
FN is number of false negatives

1. a) Consider the following XML fragment

```
<?xml version="1.0"?>
<degree xmlns="http://compute.com" xmlns:d="http://dgi.com" xmlns:e="http://good.com">
  <subject>
    <name xmlns="http://study.com">COMP20008</name>
    <d:lecture>Lec1</d:lecture>
    <workshop>Work1</workshop>
  </subject>
</degree>
```

i) (1 marks) What is the namespace URI of the element *name*?

ii) (1 mark) What is the namespace URI of the element *degree*?

ii) (1 mark) What is the namespace URI of the element *workshop*?

b) (1 mark) Provide a JSON representation of the following XML fragment.

```
<degree>
  <subject>
    <name>COMP20008</name>
    <lecture>Lec1</lecture>
    <workshop>Work1</workshop>
  </subject>
</degree>
```

c) (3 marks) Explain two advantages of applying principal components analysis to a dataset that has 500 features and 20000 instances. Explain one disadvantage.

d) (1 mark) Provide a real world example that illustrates the concept of "data missing not completely at random".

e) (2 marks) Explain how outlier detection and removal fits in the data wrangling pipeline.

2. a) Below is pseudo code for the VAT algorithm described in lectures. It outputs a permutation P of the objects.

1. Given an N*N disssimilarity matrix **D**
2. Let K={1,…N}, I=J={}
3. Pick the two least similar objects o_a and o_b from **D**
4. P(1)=a; I={a}; J=K-{a}
5. For r = 2, …., N
6. Select (i,j): pair of most similar objects o_i and o_j from **D**
7. Such that i ∈ I, j ∈ J
8. P(r) = j; I = I ∪ {j}; J=J - {j};

sim = 1 / 1+ dis

i) (3 marks) In line 6, explain why the pair of most similar objects o_i and o_j is chosen. Why not the least similar instead? most similar, least distance

ii) (2 marks) Explain how a reordered dissimilarity matrix $\mathbf{D}^*$ can be created using the permutation P.

b) (3 marks) Maxine is having a conversation about data wrangling. She says "The VAT algorithm is nice, but parallel co-ordinates is even more useful for understanding a dataset". Explain one way in which use of parallel-coordinates could reveal similar information to use of VAT. Explain one way in which use of parallel co-ordinates could reveal information that would not be revealed by use of VAT.

c) (2 marks) Using an example, explain the difference between i) user based methods for collaborative filtering and ii) item based methods for collaborative filtering.

3. Consider the following dataset, that records the predictions of a k-nearest neighbor classifier on a test set of 10 instances. There are two classes, X and Y. For each instance, the class that is predicted by the classifier is recorded and the actual (true) class of the instance is recorded.

Instance ID	*Predicted class*	*Actual class*
1	X	X
2	X	Y
3	Y	Y
4	X	X
5	X	Y
6	Y	X
7	X	X
8	Y	Y
9	Y	X
10	Y	Y

a) (1 mark) Would Pearson correlation be suitable to compute the correlation between the *Predicted class* and *Actual class*? Why or why not?

b) (1 mark) Write a formula for the entropy of *Predicted class*.

c) (2 mark) Write a formula for the normalised mutual information between *Predicted class* and *Actual class*.

d) (1 mark) What is the accuracy of the classifier for this test data?

e) (3 marks) Suppose we have some other classifier and use it to make predictions on the same test set of 10 instances. For this other classifier, suppose the normalised mutual information between *Predicted Class* and *Actual class* is reported to be 1 and its accuracy is reported to be 0. Is this possible? Explain your reasoning.

f) (2 marks) What is the purpose of separating a dataset into training and test sets, when evaluating the performance of a classifier?

4. a) Consider the 2-D dataset: {(1,2),(2,2),(3,6),(4,6),(6,6),(12,12)}. We want to use single linkage hierarchical clustering with Euclidean distance. The initial proximity matrix is shown below.

	(1,2)	(2,2)	(3,6)	(4,6)	(6,6)	(12,12)
(1,2)	0	1	$\sqrt{20}$	5	$\sqrt{41}$	$\sqrt{221}$
(2,2)	1	0	$\sqrt{17}$	$\sqrt{20}$	$\sqrt{32}$	$\sqrt{200}$
(3,6)	$\sqrt{20}$	$\sqrt{17}$	0	1	3	$\sqrt{117}$
(4,6)	5	$\sqrt{20}$	1	0	2	10
(6,6)	$\sqrt{41}$	$\sqrt{32}$	3	2	0	$\sqrt{72}$
(12,12)	$\sqrt{221}$	$\sqrt{200}$	$\sqrt{117}$	10	$\sqrt{72}$	0

Suppose the first two iterations yield the new proximity matrix below.

i) (1 mark) Draw the proximity matrix with the "?" cells filled in. Your filled in cell values may use square root symbols.

ii) (1 mark) Based on your answer to i), which clusters will be merged next? Explain your answer.

iii) (1 mark) Draw the (partial) dendrogram obtained after completing the merge in question ii).

	(1,2),(2,2)	(3,6),(4,6)	(6,6)	(12,12)
(1,2),(2,2)	0	?	?	?
(3,6),(4,6)	?	0	?	?
(6,6)	?	?	0	$\sqrt{72}$
(12,12)	?	?	$\sqrt{72}$	0

b) (2 marks) Consider the following two step procedure:

Step 1: We apply the k-means algorithm on a dataset D having n instances, using $k = 4$. This outputs a clustering C.

Step 2: Suppose we then create a new dataset D', by adding two copies of D together. i.e. $D' = D + D$. D' thus has $2n$ instances and each instance from D has 2 copies in D'. We then apply the k-means algorithm on D', with $k = 4$ and using the same initial centroids as in Step 1. This outputs a clustering C'.

Compare the executions of the k-means algorithm in Step 1 and Step 2. Explain how will they be different/similar in terms of final output, number of iterations and time complexity? State any assumptions made.

c) i) (1 mark) One can compute a quality measure for a clustering, known as SSE (sum of squared errors). SSE is the sum of Euclidean distances of objects from their cluster centroids.

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \overline{c_i})^2$$

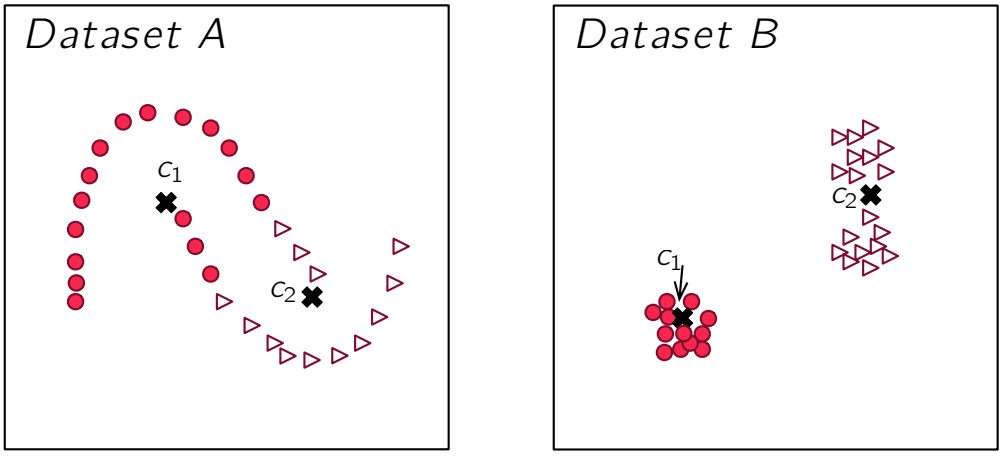
where $\overline{c_i}$ is the centroid of cluster c_i.

As the number of clusters increases, would you expect SSE to increase or decrease? Explain your answer.

ii) (2 marks) The Figure below shows the result of k-means clustering (k=2) with Euclidean distance on two datasets. The solid crosses are the centroids of clusters c_1 and c_2. The solid circle objects are in cluster c_1 and the empty triangle objects are in cluster c_2. Assume both datasets are drawn using the same distance scale.

Assuming s_A = SSE(clusters for dataset A) and s_B = SSE(clusters for dataset B), which of the following equations hold? Why? State any assumptions made.

- $s_a = s_b$
- $s_a < s_b$
- $s_a > s_b$



d) (2 marks) Max is having a conversation about data integration. He says "Using blocking for record linkage between two datasets (dataset A and dataset B) is a bad idea. It is too time consuming to assign the records to blocks. It is much better instead to directly compare the records in A against the records in B without using any blocking step". Argue why Max's statement is incorrect.

5. a) (1 mark) In the context of differential privacy, what is global sensitivity and how does it affect the amount of noise added to the output?

   b) (2 marks) Assume we have a dataset with three features: *Gender*, *Resident* and *Education*. *Gender* is a categorical feature with two possible values (Male, Female) and *Resident* is a categorical feature with two possible values (Yes, No) and *Education* is a categorical feature with possible four values (Highschool, Diploma, Undergraduate, Postgraduate).

   If we are interested in 'count' queries, what is the value of global sensitivity? Explain your answer.

   c) (3 marks) Bloom filters are useful for privacy preserving data linkage with approximate matching. Suppose we are using k hash functions for a bloom filter with length $n = 1000$. Suppose each hash function is biased, such that the probability of mapping its input to an even number is twice the probability of mapping it to an odd number.

   Explain how does the bias of the hash functions affect the probability of having a false positive for a bloom filter membership query? How does the bias of the hash functions affect the probability of having a false negative for a bloom filter membership query? State any assumptions made.

   d) (2 marks) Suppose we are using a 3 party exact matching scheme for privacy preserving data linkage. Alice and Bob each have their own dataset and they are trying to link the datasets together on people's names. They agree on a pre-processing strategy to standardise their data, which converts all upper case letters to lower case letters and converts all nick names to full names (e.g. "jim" to "james", "bob" to "robert", "liz" to "elizabeth"). Explain a possible disadvantage of their data pre-processing strategy with respect to the effectiveness of the privacy preserving data linkage.

e) Consider the following dataset, which is 3-anonymous. The POI feature records a user's last visited point of interest.

Name	Zip Code	Age	Nationality	POI
-	30**	≤ 25	*	Heart Clinic
-	30**	≤ 25	*	University
-	30**	≤ 25	*	University
-	30**	> 40	*	Heart Clinic
-	30**	> 40	*	Church
-	30**	> 40	*	Bar
-	305*	3*	*	Bar
-	305*	3*	*	Bar
-	305*	3*	*	Bar

i) (1 mark) Which features have been used as a quasi-identifier? Explain your answer.

ii) (1 mark) Describe a possible homogeneity privacy attack.

End of Exam