The University of Melbourne

Semester One 2016 Exam

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

Exam Duration: 2 hours

Reading Time: 15 minutes

This paper has 6 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: $H(\mathbf{X}) = -\sum_{i=1}^{\#bins} p_i \log_2 p_i$
where p_i is the proportion of points in the ith bin.

Joint entropy: $H(\mathbf{X}, \mathbf{Y}) = -\sum_{i=1}^{\#binsX} \sum_{j=1}^{\#binsY} p_{ij} \log_2 p_{ij}$
where p_{ij} is the proportion of points in the $\{i, j\}$th bin.

Mutual information:
$MI(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{\#binsX} \sum_{j=1}^{\#binsY} p_{ij} \log_2 \frac{p_{ij}}{p_i p_j}$

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

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

e{2}nt*ht

1. a) (1 mark) Write a regular expression to specify the following: *Two occurrences of the letter 'e', followed immediately by one 'n', followed immediately by at least one 't', followed immediately by zero or more occurrences of 'h'*

b) (2 marks) Using CouchDB as an example, explain why it is useful to have revision numbers for records and how it assists with horizontal sharding in Map-Reduce.

c) (2 marks) Two students, Bob and Alice, are having a conversation about data wrangling. Alice argues that "UNIX is fantastic for data wrangling, because it provides pipes, as well as standard input and output redirection facilities". Provide concrete examples to support Alice's claim.

d) Bob and Alice are having a debate about imputation methods for missing values. Bob argues that imputation using the mean of a feature works best. Alice argues that imputation using matrix factorization is better. To decide who is correct, they select a dataset with 10 features $F_1, \ldots, F_{10}$ and 100 instances $x_1, \ldots, x_{100}$. For each instance, they randomly make one of its feature values to be missing (yielding 100 missing values overall). They then separately apply each of the two imputation methods and assess which performs better in recovering the missing values.

i) (2 marks) What metric would be appropriate here for assessing performance of an imputation method? Provide a formula for your metric and justify it.

ii) (3 marks) Bob and Alice are joined by a third student, Barbara. Barbara says "You are both wrong. If an instance has any missing values, it is better to discard it entirely rather than trying to perform imputations." Describe three scenarios which support Barbara's reasoning.

(i).

Mean square error

$$\frac{1}{100}\sum_{i=1}^{100}\left(true\_value(xi) - impute\_value(xi)\right)^2$$

computationally expensive

(ii). ① when missing value is categorical. It is hard to do imputation. ✓

✓ ②. when the missing value has a higher percentage in the feature, then do imputation is meaningless → have more noise

where ① each instance is either complete

② or has mostly missing value

③ when the dataset is large, that it contains sufficient information even discard some examples

④ imputation is likely to

cause problem or contaminate
later analysis
(unreliability).

2. a) Richard is a data wrangler. He does a survey and constructs a dataset recording average time/day spent studying and average grade for a population of 1000 students:

Student Name	*Average time per day studying*	*Average Grade*
...	...	...

i) (3 marks) Richard computes the Pearson correlation coefficient between *Average time per day studying* and *Average grade* and obtains a value of 0.85. He concludes that more time spent studying causes a student's grade to increase. Explain the limitations with this reasoning and suggest two alternative explanations for the 0.85 result.

① use Pearson correlation mean he assume linear relationship.

limitation ② Pearson correlation cannot give causality

ii) (2 marks) Richard separately discretises the two features *Average time per day spent studying* and *Average grade*, each into 2 bins. He then computes the normalised mutual information between these two features and obtains a value of 0.1, which seems surprisingly low to him. Suggest two reasons that might explain the mismatch between the normalised mutual information value of 0.1 and the Pearson Correlation coefficient of 0.85. Explain any assumptions made.

b) (1 mark) When building a classification model, it is recommended to divide a dataset D into a training portion and a testing portion, construct the classification model based on the training portion and then evaluate its accuracy using the testing portion. Suppose instead we construct the classification model based on all of D and then evaluate its performance (accuracy) using all of D. What is the disadvantage?

c) (4 marks) Given a dataset with two classes, A and B, suppose the root node of a decision tree has 50 instances of class A and 150 instances of class B. Consider a candidate split of this root node into two children, the first with (25 class A and 25 class B), the second with (25 class A and 125 class B). Write a formula to measure the utility of this split using the entropy criterion. Explain how this formula helps measure split utility.

3. Consider the following dataset D which describes 3 people:

Age	Weight
20	40
30	50
25	25

a) (2 marks) Using a Euclidean distance function, construct the dissimilarity matrix for D. Show all working. (You may leave any square root terms unsimplified).

b) (2 marks) A dissimilarity matrix can be visualised as a heat map. Explain why it can be useful to reorder a dissimilarity matrix, using a technique such as the VAT algorithm. What information can this provide and why?

c) (2 marks) Would there be a benefit of applying principal components analysis to D to assist in visualisation? Explain.

d) (2 marks) Instead of using the VAT algorithm, one can apply parallel co-ordinates to visualise a dataset. Name one advantage of using parallel co-ordinates instead of VAT to visualise data. Name one disadvantage of using parallel co-ordinates instead of VAT to visualise data.

e) (2 marks) The 2-party privacy preserving data linkage protocol is vulnerable to small differences in the input fields being linked. Explain why this is the case.

4. For each of the following concepts, briefly explain i) its meaning, ii) why the concept is important.

a) (2 marks) Contextual outlier (in the context of data cleaning)

b) (2 marks) User based methods (in the context of collaborative filtering)

c) (2 marks) Map-Reduce (in the context of distributed systems)

d) (2 marks) Blocking (in the context of record linkage)

e) (2 marks) NoSQL (in the context of databases)

5. a) (4 marks) Explain the difference between clustering based outlier detection and k-nearest neighbor based outlier detection. What are their relative advantages/disadvantages?

   b) (4 marks) The Head of University X has a new idea: the University will require each of its students to wear a lightweight ring, having an embedded sensor with GPS technology that measures a student's location each 10 minutes and sends the (latitude, longitude) co-ordinates back to a University central server. The University plans to use this information to improve its service delivery to students.

   Describe two privacy risks of this scheme for students. For each, propose a method for reducing the privacy risk, whilst maintaining acceptable utility. Explain any assumptions made.

   c) (2 marks) Consider the typical 3 phase data science pipeline

   - Wrangle the data (80% effort)
   - Analyse the data (< 20% effort)
   - Present, deploy and communicate results (< 20% effort)

   Explain two reasons why the wrangling phase requires so much of the total effort.

End of Exam