UNIVERSITY OF ABERDEEN SESSION 2023/2024

Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University

Examination in JC3503 Data Mining and Visualisation

August 2024

---

# Part A (25 marks)

*Answer ALL questions. The marks for each question are shown in brackets.*

---

1. Consider the following table which shows 10 instances of data for binary attributes A, B and C of an entity that are classified as either + or −.

| **A** | **B** | **C** | **Class** |
|---|---|---|---|
| 0 | 0 | 0 | + |
| 0 | 0 | 1 | − |
| 0 | 1 | 1 | − |
| 0 | 1 | 1 | − |
| 0 | 0 | 1 | + |
| 1 | 0 | 1 | + |
| 1 | 0 | 1 | − |
| 1 | 0 | 1 | − |
| 1 | 1 | 1 | + |
| 1 | 0 | 1 | + |

You have been asked to construct a decision tree that captures the classification knowledge from the above training data. **Note: there is a table with some useful calculations on the next page that will help you with this**.

(a) Compute the Information Gain for the A attribute selected as the root node. **[4 mark]**

(b) Compute the Information Gain for the B attribute selected as the root node. **[3 mark]**

(c) Compute the Information Gain for the C attribute selected as the root node. **[3 mark]**

(d) Based on your answers for 1(a)..(c), which attribute would you split on at the root of the tree? **[1 mark]**

| Useful Fractions | | |
|---|---|---|
| $1/3 = 0.33$ | $2/3 = 0.67$ | $3/7 = 0.43$ |
| $4/7 = 0.57$ | $4/9 = 0.44$ | $5/9 = 0.56$ |

| Useful Multiplication | | |
|---|---|---|
| $0.3 * 0.92 = 0.28$ | $0.33 * -1.60 = -0.53$ | $0.4 * -1.32 = -0.53$ |
| $0.43 * -1.22 = -0.52$ | $0.44 * -1.18 = -0.52$ | $0.56 * -0.84 = -0.47$ |
| $0.57 * -0.81 = -0.46$ | $0.6 * -0.74 = -0.44$ | $0.67 * -0.58 = -0.39$ |
| $0.7 * 0.98 = 0.69$ | $0.9 * 0.99 = 0.89$ | |

| Useful $Log_2$ Calculations | | |
|---|---|---|
| $log_2(0.33) = -1.60$ | $log_2(0.4) = -1.32$ | $log_2(0.43) = -1.22$ |
| $log_2(0.44) = -1.18$ | $log_2(0.5) = -1$ | $log_2(0.56) = -0.84$ |
| $log_2(0.57) = -0.81$ | $log_2(0.6) = -0.74$ | $log_2(0.67) = -0.58$ |

2. Statistically we define four levels of measurement for attribute values of data: **Nominal**, **Ordinal**, **Interval**, and **Ratio**.
   Classify the following attribute values into these four levels of measurement:

   (a) Animal classification: {Bird, Mammal, Reptile} **[1 mark]**

   (b) Temperature feel: {Cold, Warm, Hot} **[1 mark]**

   (c) Height **[1 mark]**

   (d) Assignment grade: {A, B, C, D, F } **[1 mark]**

3. Consider the following table, which shows 3 documents, which are classified as either relating to electronics (E), or fruit (F).

| Document ID | Words in document | Class |
|---|---|---|
| 1 | apple mac iPad apple | E |
| 2 | apple iPhone mac | E |
| 3 | apple pear orange pear | F |

   Based on the data given in the above table:

   (a) Calculate the prior probability of a document occurring in each class, i.e. P(E) and P(F) **[2 mark]**

(b) For each word in the document vocabulary, calculate the conditional word probability given a class label, i.e. P(apple | E), P(apple | F), etc. **[6 mark]**

(c) Using Naïve Bayes classification and the information you have already computed from 3.(b), what formulas would you use to determine the class label of a new document with the words: {pear apple pear mac}? **Note** that you don't have to compute these formulas. **[2 mark]**

| **Useful Fractions** | | |
|---|---|---|
| $1/4 = 0.25$ | $2/4 = 0.5$ | $3/4 = 0.75$ |
| $1/7 = 0.14$ | $2/7 = 0.29$ | $3/7 = 0.43$ |
| $4/7 = 0.57$ | $5/7 = 0.71$ | $6/7 = 0.86$ |

---

# Part B (25 marks)

*Answer ALL questions. Each part is worth 25 marks; the marks for each question are shown in brackets.*

---

4. Consider the following samples:

2, 3, 1, 2, 1, 3

(a) Please calculate the mean for the above samples. **[1 mark]**

(b) Please calculate the standard deviation for the above samples. **Note**: To compute the standard deviation, one of the square root calculations given below will be useful: **[2 mark]**

| **Useful Square Root Calculations** | | |
|---|---|---|
| $\sqrt{3/5} = 0.77$ | $\sqrt{4/6} = 0.82$ | $\sqrt{5/6} = 0.91$ |
| $\sqrt{4/5} = 0.89$ | $\sqrt{3/6} = 0.71$ | $\sqrt{2/5} = 0.63$ |

(c) Using the answers you calculated in 2.(a), fill in the formulas you would use to calculate the z-score (standard score) for each of the above samples. **Note**: you do not need to *solve* the formulas; you can stop when you reach the point of needing to do any complex calculations. **[3 mark]**

5. Consider the following transactions involving five items. Imagine that you have been asked to produce association rules for the items using Apriori algorithm:

| **transaction_ID** | **items_bought** |
|---|---|
| 1 | newspaper,beer,pen,water |
| 2 | beer,magazine,pen |
| 3 | newspaper,beer,pen,water |
| 4 | newspaper,magazine,pen,water |

(a) Using a minimum support of 0.75, generate the frequent itemsets for the above data showing clearly the application of Apriori principle in pruning infrequent itemsets. **[5 mark]**

(b) Using a minimum confidence of 0.75, generate the association rules generated from the frequent itemsets computed in 5.(a) showing clearly the application of Apriori principle in pruning low confidence rules. **[4 mark]**

6. Given the following proximity matrix for data points a–e, you use the agglomerative hierarchical clustering algorithm to cluster the data.

| | $a$ | $b$ | $c$ | $d$ | $e$ |
|---|---|---|---|---|---|
| $a$ | 1.00 | 0.80 | 0.90 | 0.65 | 0.20 |
| $b$ | 0.80 | 1.00 | 0.70 | 0.60 | 0.50 |
| $c$ | 0.90 | 0.70 | 1.00 | 0.40 | 0.30 |
| $d$ | 0.65 | 0.60 | 0.40 | 1.00 | 0.35 |
| $e$ | 0.20 | 0.50 | 0.30 | 0.35 | 1.00 |

**Note**: Please use $sim(i, j)$ to represent similarity between $i$ and $j$, where $i$ and $j$ are points or clusters. For instance, $sim(a, b) = 0.90$ and $sim(ab, d) = 0.65$, where ab is a cluster containing Points a and b.

(a) **Draw a dendrogram (tree diagram)** for the algorithm using the **MIN** (Single Link) inter-cluster similarity measure. Please also give detailed steps of your calculation. **[5 mark]**

(b) **Draw a dendrogram (tree diagram)** for the algorithm using the **MAX** (Complete Linkage) inter-cluster similarity measure. Please also give detailed steps of your calculation. **[5 mark]**

END OF EXAM PAPER