Data Mining INFS4203/7203 Mock Final-Term Exam: An Example Answer

Q1. [6 marks]

a. **[2 marks] Briefly describe the meaning of association rule in association rule mining**
b. **[2 marks] Briefly describe the meaning of strong association rule in association rule mining**
c. **[2 marks] Given the frequent k-itemsets, list the steps to generate the frequent (k+1)-itemsets from frequent k-itemsets.**

Answer:

a. An association rule is an implication of the form A-> B, where A and B are non-empty itemsets and the intersection of A and B is empty.
b. A rule A->B is called strong association rule if the support for the union of A and B is no smaller than the minimum support threshold, and the confidence of A->B is no smaller than the minimum confidence threshold.
c. Candidate generation, candidate pruning, support counting, candidate elimination

Q2. [10 marks]

a. **[2 marks] Briefly describe the objective of clustering**
b. **[4 marks] Use natural language to describe the k-means algorithm**
c. **[2 marks] List two limitations of the k-means algorithm**
d. **[2 marks] Briefly discuss how to select "k" in the k-means algorithm**

Answer:

a. Partition a set of objects into groups according to their properties, such that objects within one group are similar and objects from different groups are dissimilar.
b. 1.Select k points as the initial centroids;
2.Repeat the following two steps until the cluster assignment do not change:
3.Form k clusters by assigning each point to the closest centroid,
4.Update the centroid of each cluster by the mean of the points within the cluster

c. can not handle spiral data; selecting an appropriate k is difficult; be easily impacted by initialization…

d. we use elbow method: try different k, and plot the mean SSE versus k; select the corresponding k at the "Elbow" of the curve.

Q3. [8 marks]

a. **[2 marks] Briefly describe the objective of anomaly detection**
b. **[4 marks] Briefly describe the cluster-based technique for anomaly detection**

c. **[2 marks] List one advantage and one limitation of the cluster-based technique for anomaly detection**

Answer:

a. Detect the data point that deviates from what is standard, normal or expected
b. In cluster-based anomaly detection, clustering is conducted first on the data, and objects are evaluated with respect to how strong they belong to each cluster. If a point does not strongly belong to any cluster, this point is an anomaly. The strong can be measured either by the data point's distance to any centroid, or the density around the point.
c. Advantage: can borrow the merit from the development of clustering technique; limitation: outlier may distort the cluster

Q4. [7 marks] You are working on a dog recognition classifier. Assume that "Dog" is the positive class (y=1) and "Not Dog" is the negative class (y=0). Additionally, an Algorithm "A" is used to classify the test set. The following table show the corresponding classification results obtained by the algorithm, where the left column is the ground truth class and the right column is the predicted class. Answer the following questions.

Ground Truth	Predicted
Dog	Dog
Not Dog	Dog
Not Dog	Not Dog
Dog	Dog
Dog	Not Dog
Dog	Not Dog
Not Dog	Not Dog
Not Dog	Dog
Not Dog	Not Dog
Dog	Dog

a. **[1 mark] Evaluate the algorithm A under the metric accuracy (round to 2 decimal places).**
b. **[1 mark] Evaluate the algorithm A under the metric precision (round to 2 decimal places).**
c. **[1 mark] Evaluate the algorithm A under the metric recall (round to 2 decimal places).**
d. **[1 mark] Evaluate the algorithm A under the metric F1 measurement (round to 2 decimal places).**
e. **[1 mark] Given another algorithm B, whose accuracy is 0.60, precision is 0.4 and recall is 0.8, compare A and B. Discuss which algorithm you would choose and why.**
f. **[2 marks] Briefly discuss how to evaluate the performance of a classification algorithm if the test set is not available.**

Answer:

a. Accuracy = 0.6
b. Precision = 0.6
c. Recall = 0.6
d. F1 = 0.6
e. F1 of B is 0.53. So, A is better because A has better F1 and equal accuracy compared to B.
f. We use cross-validation to evaluate. For example, if we use 5 cross-validation, we randomly divide the data into 5 folders, with almost equal number of instances in each folder. Each time, we use one folder as test data and all the other folders as training data. In this way, we do training and test 5 times, and the average performance is used to evaluate the algorithm.

More explanation:

Confusion matrix

	Predicted P	Predicted N
GT P	3	2
GT N	2	3

Q5. [8 marks]

We have some data about whether people go to watch a football game. The data includes three attributes: whether the game is on a weekend or not, whether the person has friends to watch the game together or not, and whether there is any football star in the game or not:

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
2	Yes	Yes	No	No
3	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes
5	No	Yes	Yes	Yes
6	No	Yes	No	Yes
7	Yes	Yes	No	No
8	Yes	Yes	Yes	Yes
9	Yes	No	No	Yes
10	Yes	No	No	No

11	No	No	No	Yes

a. **[5 marks] Construct a decision tree from the provided dataset to predict whether people will watch the football game given three attributes "Weekend", "Friends" and "Star there", using the GINI Index-based splitting criterion.**
b. **[2 marks] Briefly discuss the advantage of using Gain Ration-based splitting criterion compared to Information Gain-based splitting criterion.**
c. **[2 marks] Briefly describe the two pruning methods in decision tree.**

Answer:

a.

Choice 1:

If Weekend = Yes, continue
- If Star = Yes, Yes
- If Star = No, continue
  - If Friends = Yes, No
  - If Friends = No, Yes or No

If Weekend = No, continue

Choice 1.1: Split with Friends
- If Friends = Yes, continue
  - If Star = Yes, Yes
  - If Star = No, Yes or No
- If Friends = No, Yes

Choice 1.2: Split with Star
- If Star = Yes, Yes
- If Star = No, Continue
  - If Friends = Yes, Yes or No
  - If Friends = No, Yes

Choice 2:

If Star = Yes, Yes

If Star = No, continue

Choice 2.1: Split with Weekend
- If Weekend = Yes, continue
  - If Friends = Yes, No
  - If Friends = No, Yes or No
- If Weekend = No, continue
  - If Friends = Yes, Yes or No
  - If Friends = No, Yes

Choice 2.2: Split with Friends
- If Friends = Yes, continue,
  - If Weekend = Yes, No
  - If Weekend = No, Yes or No

If Friends = No, continue,

If Weekend = Yes, Yes or No

If Weekend = No, Yes

b.

The decision tree using information gain as purity measurement will select the attribute with large number of distinct values to split. However, selecting such an attribute may result in poor generalization performance. The decision tree using gain ratio as purity measurement will select the attribute with a better balance of the number of distinct values and information gain.

c.

Pre-pruning: halt tree construction in the construction phase. Do not split a node if this would result in poorer generalization

Post-pruning: remove branches after the tree construction phase if removing would result in better generalization

More explanation:

GINI index for Weekend: 0.4481
GINI index for Friends: 0.4848
GINI index for Star there: 0.4481

We can either start with Weekend or Star:

Choice 1:
If we start with Weekend,
If Weekend = Yes

Tid	Weekend	Friends	Star there	Watch?
2	Yes	Yes	No	No
3	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes
7	Yes	Yes	No	No
8	Yes	Yes	Yes	Yes
9	Yes	No	No	Yes
10	Yes	No	No	No

Gini index for Friends: 0.4857
Gini index for Star: 0.4048

We choose Star as the second split:

If Star = Yes

Tid	Weekend	Friends	Star there	Watch?
3	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes
8	Yes	Yes	Yes	Yes

We do not need further splitting, and the prediction is Yes

If Star = No

Tid	Weekend	Friends	Star there	Watch?
2	Yes	Yes	No	No
7	Yes	Yes	No	No
9	Yes	No	No	Yes
10	Yes	No	No	No

If Friends = Yes, No
If Friends = No, Yes or No

If Weekend = No

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
5	No	Yes	Yes	Yes
6	No	Yes	No	Yes
11	No	No	No	Yes

Friends:0.3333
Star: 0.3333
For the second splitting, we choose either one to slit.
Choice 1.1: If Friends = Yes

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
5	No	Yes	Yes	Yes
6	No	Yes	No	Yes

If Star = Yes, Yes
If Star = No, Yes or No

If Friends = No

Tid	Weekend	Friends	Star there	Watch?
11	No	No	No	Yes

Yes

Choice 1.2: If Star = Yes, Yes

If Star = No, If Friends = Yes, Yes or No; If Friends = No, Yes

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
6	No	Yes	No	Yes
11	No	No	No	Yes

==========

Choice 2: We start with Star

If Star = Yes,

Tid	Weekend	Friends	Star there	Watch?
3	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes
5	No	Yes	Yes	Yes
8	Yes	Yes	Yes	Yes

We can only split with Weekend:
If weekend = Yes, Yes
If weekend = No, Yes

If Star = No

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No

2	Yes	Yes	No	No
6	No	Yes	No	Yes
7	Yes	Yes	No	No
9	Yes	No	No	Yes
10	Yes	No	No	No
11	No	No	No	Yes

Weekend: 0.4048

Friends: 0.4048

Choice 2.1: Split with Weekend

If weekend == Yes

Tid	Weekend	Friends	Star there	Watch?
2	Yes	Yes	No	No
7	Yes	Yes	No	No
9	Yes	No	No	Yes
10	Yes	No	No	No

If Friends = Yes, No

If Friends = No, Yes or No

If weekend = No

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
6	No	Yes	No	Yes
11	No	No	No	Yes

If friend = Yes, Yes or No

If friend = No, Yes

Choice 2.2: Split with Friends

If Friends = Yes

Tid	Weekend	Friends	Star there	Watch?
1	No	Yes	No	No
2	Yes	Yes	No	No
6	No	Yes	No	Yes
7	Yes	Yes	No	No

If Weekend = Yes, No

If Weekend = No, Yes or No

If Friends = No

Tid	Weekend	Friends	Star there	Watch?
9	Yes	No	No	Yes
10	Yes	No	No	No
11	No	No	No	Yes

If Weekend = Yes, Yes or No

If Weekend = No, Yes

Q6. [6 marks]

Given a data set below, answer the following questions.

Tid	Home Owner	Marital Status	Annual Income	Class
1	**Yes**	**Single**	**125K**	**No**
2	**No**	**Married**	**120K**	**No**
3	**No**	**Single**	**70K**	**No**
4	**Yes**	**Married**	**120K**	**No**
5	**No**	**Divorced**	**95K**	**Yes**
6	**No**	**Married**	**60K**	**No**
7	**Yes**	**Divorced**	**120K**	**No**
8	**No**	**Married**	**85K**	**Yes**
9	**No**	**Married**	**75K**	**No**
10	**No**	**Single**	**120K**	**Yes**

a. [2 marks] What is the "Attribute conditional independence" assumption for Naïve Bayesian classifier?

b. [4 marks] Predict the class label of a given test record X: (Home Owner: Yes, Marital Status: Single, Annual Income: 80K) using the Naïve Bayesian Classifier.

Answer:

a. All features are conditionally independent from each other given the class information.

b.

P(X|Yes) is in proportion to P(Yes)P(HO = Yes|Yes) P(MS = Single|Yes) P(AI = 80K|Yes) = (3+1)/(10+2)*(0+1)/(3+2)*(1+1)/(3+3)*0.0120 = 0.0003

P(X|No) is in proportion to P(No)P(HO = Yes|No) P(MS = Single|No) P(AI = 80K|No)= (7+1)/(10+2)*(3+1)/(7+2)*(2+1)/(7+3)*0.0114 = 0.0010
So it is predicted as "No" as P(X|Yes)< P(X|No)

More explanation:
P(HO= Yes |Yes) = 0, so we need to use Laplacian correction
If Class = Yes, AI = [85, 95, 120] with mean 100 and variance 325 so we have
P(AI = 80K|Yes) = 0.0120
If Class = No, AI = [125, 120, 70, 120, 60, 120, 75] with mean 98 and variance 823, so we have
P(AI = 80K|No) = 0.0114

Q7. [5 marks]

Given a two-dimensional data set below, answer the following questions.

(x, y)	Class
(0, 5)	+
(2, 3)	+
(4, 3)	+
(5, 5)	-
(6 ,7)	-
(8, 5)	-
(3, 7)	+
(4, 10)	-
(7, 9)	+

a. [1 mark] What is the basic idea of a k-NN method?

b. [2 marks] Classify a new data point p = (4, 5) using Manhattan distance and majority vote using 3- nearest neighbours.

c. [2 marks] Briefly discuss when we need to normalize the attribute values in k-NN classification and list the names of two normalization methods.

Answer:

a. Find from the historical record those as similar as possible to the new record.
b. The Manhattan distance from (4, 5) to all the 9 points are 4, 4, 2, 1, 4, 4, 3, 5, 7. The 3-nearest neighbours are (4, 3) +, (5, 5) -, (3, 7) +. By majority voting, (4,5) is classified as +
c. When different attributes have different degrees of magnitude, we need to normalize the values of the attributes.
   List two:
   Max-min normalization
   Standardization.