**Data Mining Research & Practices – 2018 Midterm**

1. Suppose that the data mining task is to cluster the following 8 points (with (x,y) representing location) into 2 clusters. **N1(4,2), N2(4,4), N3(6,6), N4(3,0), N5(13,8), N6(9,6), N7(8,12), N8(1,2).** The distance function is ***Manhattan*** distance. Suppose initially we assign **N1** and **N6** as the center of each cluster, respectively. Use the k-means algorithm to show  
   (a) The two cluster centers after the first round execution.

(b) The final two clusters and centers.

1. (a) Use the similarity matrix in the following table to perform **single link** and **complete link** hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| P1 | **1** | **0.55** | **0.40** | **0.85** | **0.45** |
| P2 | **0.55** | **1** | **0.70** | **0.80** | **0.90** |
| P3 | **0.40** | **0.70** | **1** | **0.95** | **0.75** |
| P4 | **0.85** | **0.80** | **0.95** | **1** | **0.60** |
| P5 | **0.45** | **0.90** | **0.75** | **0.60** | **1** |

(b) Explain the idea of “complete link” for hierarchical clustering.

1. Suppose that the ID3 algorithm is used to construct a decision tree to decide whether the consumer is suitable to buy mobile App. Table 1 contains ten different 10 records. The target classification is “Yes” or “No” for buying mobile App.

**(，，)**

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Platform** | **Income** | **Buy App?** |
| A | iOS | High | Yes |
| B | iOS | High | Yes |
| C | Android | Low | No |
| D | iOS | Middle | Yes |
| E | Android | High | No |
| F | Other | Low | No |
| G | iOS | Middle | Yes |
| H | Android | Middle | Yes |
| I | Other | Middle | No |
| J | iOS | High | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| I value | | | |
| I(1,2) | 0.92 | I(2,3) | 0.97 |
| I(1,3) | 0.81 | I(2,4) | 0.92 |
| I(1,5) | 0.65 | I(2,5) | 0.86 |

(a) Which attribute will be selected as the first test attribute? Please use the **information gain** measure as the attribute selection measure.

(b) Show the final decision tree returned by the ID3 algorithm. You need to clearly indicate the class label of each leaf node.

(c) Predict the class label of an unlabeled data sample with ***Platform = “iOS” and Income = “Low”*** according to the constructed decision tree.

(d) Predict the class label of an unknown data sample with the values ***Platform =”*** ***Android” and Income = “High”***, using the **Naive Bayesian classification**.

There are two classifiers (test drugs) C1 and C2 for heart disease. Suppose that there are 10% people having heart disease in a city.

(a) For heart disease patients, there are 80% positive by C1. For patients who don’t have heart disease, there are 60% negative by C1. Compute the***Precision****,* ***Recall,*** *and* ***Accuracy*** of the classifier C1.

(b) For patients who don’t have heart disease, there are 70% negative by C2. Suppose that the ***Precision*** of the classifier C2 is 25%. What are the ***Recall*** and ***Accuracy*** of the classifier C2?

1. Explain the idea of bootstrap.
2. Explain the Random Sampling with Holdout method.
3. Explain the idea of Boosting approach and how to adjust the weights of training samples.
4. Explain the differences between the bagging and boosting approaches.
5. Suppose that leaves(T) denotes the set of leaf nodes in a regression tree T. Let *f* denote a leaf node in *leaves*(T) ; let *Cf*denote the set of data points in a leaf node *f*; and let *Yi* be the value (target variable) of a data point *i* in *Cf*. Use above symbols (*leaves*(T), *f*, *Cf*. *i*, *Yi* ) to answer the following questions. Use examples or diagrams to aid your explanations.
6. Explain how the impurity is measured for the regression tree in **CART**.
7. Derive the equation for measuring the impurity of the regression tree in **CART**.
8. Briefly explain the differences in building a classification tree and a regression tree.
9. Explain how to determine the split of a node for numerical attribute in constructing a classification tree. Use examples or diagrams to aid your explanations.
10. Explain how each node is determined in constructing Random Forest.
11. Explain how to obtain the classification result by using Random forest with K trees. Explain how to derive the numerical prediction result by using Random forest with K trees.
12. Explain the idea of collaborative filtering for binary rating matrix with implicit rating (1: purchased, 0: non-purchased). Use examples or diagrams to aid your explanations.
    1. How to derive the similarity measures between users.
    2. How to derive the recommendation scores on items for the target user based on his/her neighbors?
13. There are two Classes, C1 and C2. The total number of documents in the training set is 100, and the number of documents belonging to C1 is 60. The following table shows the probability of P(Xi | Cj). Given a document D1 that contains some terms shown in the table.
14. Please use the Naive Bayesian document classification method to determine which Class does D1 belong to.
15. Give a simple example to explain how to calculate P(Xi|Cj)?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table** | | |  | **Document D1** | |
| Term(Xi) | P(Xi|C1) | P(Xi|C2) |  | Document term | Frequency |
| X1 | 1/16 | 3/32 |  | X1 | 0 |
| X2 | 1/4 | 1/16 |  | X2 | 2 |
| X3 | 1/8 | 1/16 |  | X3 | 2 |
| X4 | 1/16 | 1/8 |  | X4 | 0 |
| X5 | 1/8 | 1/4 |  | X5 | 1 |
| X6 | 3/16 | 1/32 |  | X6 | 0 |
| X7 | 1/16 | 1/4 |  | X7 | 0 |
| X8 | 1/8 | 1/8 |  | X8 | 1 |

1. Briefly explain the main idea of Probabilistic Model (Multinomial model) for document classification.
2. Briefly explain the difference between the assumption of unigram language model and Bigram language model?

9.

(a) Assume that the probability distribution of D1 is 1/999 for each class *Ci*, *i* = 1 to 999, and the probability distribution of D2 is 3/16, 1/16, 1/8, 2/8, 3/16, 1/8, 1/16 for each class, respectively. Which one (D1 or D2) has higher value of entropy? Which one has higher impurity? Explain why.

(b) Explain the basic concept of ***Info(D)*** (entropy of D).

(c) Explain the basic concept of ***Gini Index***.

10.

(a) Explain why we want to minimize the following equation.

(b) Explain how the ALS (Alternating-Least-Squares) approach handles the implicit feedback (presence or absence of an event). You need to use an example to explain the values of the rating matrix and the usage of cost function (confidence) in ALS.

11. A database has five transactions. Let min\_sup=55% and min\_conf=85%.

Table 1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TID | DATE | ITEMS\_BOUGHT | | | | | | |  |
| T1 | 2015/10/10 | A | B | C |  | E | F |  | H |
| T2 | 2015/11/01 |  |  | C | D | E | F |  |  |
| T3 | 2016/01/01 | A |  | C |  | E | F | G |  |
| T4 | 2016/01/08 | A | B |  |  |  |  |  | H |
| T5 | 2016/01/20 | A | B | C |  | E | F | G | H |
| T6 | 2016/02/11 | A | B |  | D |  |  |  | H |
| T7 | 2016/03/01 |  |  | C | D | E |  | G |  |

1. List all of the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where X is a variable representing customers and itemi denotes variables representing items (e.g., “A”, “B”, etc.): 
2. Establish the ***FP-tree*** and find out “*Conditional Pattern Base”,* *“Conditional FP-tree”* and “*Frequent Patterns Generated”* for **item F.**
3. Explain the idea of Gradient Descent. You should draw a diagram and use an example to aid your explanations.