**Data Mining Research & Practices – 2016 Midterm Homework**

1. (a) (8%) Explain how k-means clustering is executed in Map-Reduce by using four Map tasks and three Reduce tasks to cluster data into four clusters. You should use examples (partial data) and draw a diagram to aid your explanations. Clearly indicate the key values that are shuffled to the Reduce tasks.
2. (4%) Explain the map method and reduce method for k-means clustering. You also need to clearly indicate the input and output of the two methods.
3. (a) (6%) Briefly explain the Hadoop Distributed File System (HDFS). You should draw a diagram to aid your explanations. You need to explain the functions of namenode and datanode.

(b) (6%) Assume that the input file is split into six partitions F1, F2, F3, F4, F5 and F6. F2 and F6 are stored in datanode D1. F1 and F4 are stored in datanode D2. F3 and F5 are stored in datanode D3. Six map tasks M1, M2, M3, M4, M5 and M6 are started to process the input file partitions F1, F2, F3, F4, F5 and F6 respectively. There are three reduce tasks R1, R2 and R3. Six workers W1 ~ W6 are started to run the map tasks and reduce tasks. Draw a diagram to show a cluster of computer nodes (machines) and indicate the following – the datanodes with input file partitions, the map tasks, reduce tasks, the workers, the job tracker and the namenode. Clearly indicate the tasks that are assigned to the workers. Assume that the job tracker decides on where to run each map task based on the concept of locality.

1. (8%) Suppose that the data mining task is to cluster the following 8 points (with (x,y) representing location) into 2 clusters. **N1(3,5), N2(6,6), N3(11,8), N4(4,9), N5(8,8), N6(2,2), N7(7,6), N8(3,4).** The distance function is ***Manhattan*** distance. Suppose initially we assign **N4** and **N5** 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) (8 %) 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** | **0.60** | **0.80** | **0.70** | **0.90** |
| P2 | **0.60** | **1** | **0.55** | **0.36** | **0.95** |
| P3 | **0.80** | **0.55** | **1** | **0.85** | **0.42** |
| P4 | **0.70** | **0.36** | **0.85** | **1** | **0.82** |
| P5 | **0.90** | **0.95** | **0.42** | **0.82** | **1** |

(b) (3%) Explain the idea of “complete link” for hierarchical clustering. (Why the two clusters can be merged when all the links between all elements in one cluster and all elements in the other are connected?)

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 | Other | Middle | No |
| C | Android | Middle | Yes |
| D | iOS | High | Yes |
| E | iOS | Middle | No |
| F | Other | Middle | No |
| G | Android | Low | No |
| H | iOS | High | Yes |
| I | Other | Low | No |
| J | Android | High | Yes |

(a) (3%) Which attribute will be selected as the first test attribute? Please use the **information gain** measure as the attribute selection measure.

(b) (3%) Show the final decision tree returned by the ID3 algorithm. You need to clearly indicate the class label of each leaf node.

(c) (3%) Predict the class label of an unlabeled data sample with ***Platform = “Android” and Income=“Middle”*** according to the constructed decision tree.

(d) (4%) Predict the class label of an unknown data sample with the values ***Platform =”*** ***iOS” and Income = “Middle”***, using the **Naive Bayesian classification**.

(e) (3%) What’s the assumption in “Naïve” Bayesian classifier which enable the posteriori probability P(X|Ci) can be calculated as the following format ?



1. (a) (3%) Assume that the probability distribution of D1 is 1/700 for each class *Ci*, *i* = 1 to 700, and the probability distribution of D2 is 1/7, 5/14, 1/28, 2/7, 5/28 for each class, respectively. Which one (D1 or D2) has higher value of entropy? Which one has higher impurity? Explain why.

(b) (3%) Explain the basic concept of ***Info(D)***(entropy of D).

1. (8%) There are two classifiers (test drugs) C1 and C2 for heart disease. Suppose that there are 30% people having heart disease in a city.

(a) For heart disease patients, there are 70% 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 Accuracy of the classifier C2 is 68%. What are the ***Precision*** *and* ***Recall*** of the classifier C2?

1. (3%) Explain the k-fold cross-validation.
2. (3%) Explain the idea of ensemble approach.
3. (4%) Explain the idea of bootstrap approach
4. (3%) Explain the Random Sampling with Holdout method..
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. (3%) Explain how the impurity is measured for the regression tree in **CART**.
7. (4%) Derive the equation for measuring the impurity of the regression tree in **CART**.
8. (3%) Explain how the prediction is made for numerical target variable in **CART**.
9. (a) (3%) Explain how the training set is generated for each tree in **Random Forest**.

(b) (3%) Explain how each node is determined in constructing **Random Forest.**

1. You need to draw diagrams to assist your explanation. (clearly indicate who is the target user and which item is the target item for recommendation)

(a) (5%) Explain the following equation.



(b) (5%) How to derive the similarity between items? Explain the equation you used.

1. (10%) Please apply the correlation and prediction techniques to compute **Ken**’s predicted rating score on **item 1**.

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| item # | Meg | Gin | Ken | Nan |
| 1 | 3 | 2 | ? |  |
| 2 |  | 3 | 4 |  |
| 3 |  | 1 | 2 |  |
| 4 | 2 | 2 | 3 | 1 |
| 5 | 4 |  | 1 | 3 |
| 6 | 3 |  | 5 | 2 |



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Log value mapping table (，，)** | | | | | | | |
| Log | Value | Log | Value | Log | Value | Log | Value |
|  | -1.5850 |  | -0.5850 |  | -2 |  | -1.22 |
|  | -0.5850 |  | -1.81 |  | -1 |  | -0.81 |

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