Data Mining Research & Practices – Midterm HW

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**	**0.95**	**0.60**	**0.30**	**0.50**
P2	**0.95**	**1**	**0.90**	**0.70**	**0.20**
P3	**0.60**	**0.90**	**1**	**0.65**	**0.75**
P4	**0.30**	**0.70**	**0.65**	**1**	**0.80**
P5	**0.50**	**0.20**	**0.75**	**0.80**	**1**

(b) (2%) Explain the idea of "complete link" for hierarchical clustering.

2. 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	Android	High	Yes
C	Android	Low	No
D	iOS	High	Yes
E	iOS	Middle	No
F	Other	High	No
G	iOS	Middle	No
H	Other	Middle	No
I	Android	Low	No
J	iOS	Low	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) (4%) 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 = "Other" and Income = "Low"*** according to the constructed decision tree.
(d) (4%) Predict the class label of an unknown data sample with the values ***Platform =" Android" and Income = "Low"***, using the **Naive Bayesian classification**.

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

3.

(8%) 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 patients who don't have heart disease, there are 20% positive by C1. Suppose that the ***Accuracy*** of the classifier C1 is 80%. What are the ***Precision*** and ***Recall*** of the classifier C1?

(b) For patients who don't have heart disease, there are 60% negative by C2. Suppose that the ***Precision*** of the classifier C2 is 10%. What are the ***Recall*** and ***Accuracy*** of the classifier C2?

4.

(a) (2%) Explain the idea of stratified K-fold cross validation.

(b) (2%) Explain the Random Sampling with Holdout method.

(c) (3%) Explain the idea of Boosting approach and how to adjust the weights of training samples.

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 C_f denote the set of data points in a leaf node f; and let Y_i be the value (target variable) of a data point i in C_f. Use above symbols (*leaves*(T), f, C_f. i, Y_i) to answer the following questions. Use examples or diagrams to aid your explanations.

   (a) (4%) Derive the equation for measuring the impurity of the regression tree in **CART**.

   (b) (2%) Briefly explain the differences in building a classification tree and a regression tree.

   (c) (2%) 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.

   (d) (2%) Explain how to derive the training sets of Random forest with K trees.

   (e) (2%) Explain how each node is determined in constructing Random Forest.

   (f) (4%) 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.

6. Briefly explain the idea of collaborative filtering with implicit rating (1: purchased, 0: non-purchased). Use examples or diagrams to aid your explanations.

   (a) (3%) How to derive the similarity measures between users.

   (b) (3%) How to derive the recommendation score on the target item for the target user based on his/her neighbors?

7.

(a) (4%) Assume that the probability distribution of D1 is 1/9999 for each class C_i, i = 1 to 9999, and the probability distribution of D2 is 1/16, 1/8, 3/16, 1/4, 1/8, 1/16, 3/16 for each class, respectively. Which one (D1 or D2) has higher value of ***Gini Index***? Which one has higher impurity? Explain why.

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

(c) (2%) Explain the basic concept of ***Gini Index***.

8. There are two Classes, C1 and C2. The total number of documents in the training set is 50, and the number of documents belonging to C1 is 30. The following table shows the probability of P(Xi | Cj). Given a document D1 that contains some terms shown in the table.

(a) (6%) Please use the Naive Bayesian document classification method to determine which Class does D1 belong to.

Table

Term(Xi)	P(Xi|C1)	P(Xi|C2)
X1	3/16	1/16
X2	1/16	1/8
X3	1/4	1/4
X4	5/32	3/16
X5	1/32	1/8
X6	1/16	1/16
X7	1/8	1/16
X8	1/8	1/8

Document D1

Document term	Frequency
X1	0
X2	2
X3	2
X4	0
X5	1
X6	2
X7	1
X8	0

(b) (3%) Briefly explain the main idea of Probabilistic Model (Multinomial model) for document classification.

(c) (4%) Briefly explain the Vector Space Model (VSM). Explain tf_{ij}, $\log\frac{N}{df_j}$ and the formula $w_{ij} = tf_{ij} \times \log\frac{N}{df_j}$ (for term t_j of the document D_i)

9. Given the following equation for Matrix factorization (MF).

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

(a) (4%) Explain the idea of Gradient Descent. You should draw a diagram and use an example to aid your explanations.

(b) (4%) Briefly explain how to update $\boldsymbol{qi}$ by using the gradient descent approach for solving the MF.

(c) (3%) Explain why the regularization term is added in the equation.

(d) (4%) Show the equation of MF for handling implicit rating (1: purchased, 0: non-purchased). Explain the usage of the confidence parameter for dealing with the uncertainty of implicit feedback.

10.

True Class	N	N	P	N	P	N	P	N	P	P
Prob. Score	0.25	0.4	0.5	0.7	0.76	0.80	0.82	0.85	0.90	0.95

(a) (4%) Draw the ROC curve (TPR vs. FPR) for the above classification result.

TPR = number of true positives / the total number of actual positives

FPR = number of false positives / the total number of actual negatives

(b) (4%) Give an example to show the classifier with best classification result. Draw the ROC curve.