CS412: An Introduction to Data Warehousing and Data Mining **Fall 2014**

Assignment 5

Due Date: Dec/3/2014

General Instruction

- Feel free to talk to other members of the class in doing the homework. We are more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.

Assignment Submission

- The homework is due at **11:59 PM on the due date**. **We do NOT accept late homework!**
- We will be using Compass for collecting the homework assignments. Please submit your answers via Compass (http://compass2g.illinois.edu). Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment.
- The homework **MUST** be submitted in **pdf** format. Answers to the written part and mini-MP should be included in one .pdf file.
- If scripts are used to solve problems, you are required to submit the source code, and use the file names to identify the corresponding questions. For instance, 'Question1.netid.py' refer to the python source code for Question 1, replace 'netid' with your netid. Please **DO NOT** zip the files.
- For each question, you will **NOT** get full credit if you only give out a final result. Necessary calculation steps and reasoning are required.

Dataset

- The data for Question 3 and Mini-MP can be download from compass 2g or the course website.

Question 1 (30 points)

Suppose we want to predict whether a restaurant is popular based on its price, parking, and cuisine, and we collected training data as in table 1. Popularity is the label, and (Price, Parking, Cuisine) are the features. Answer following questions.

Table 1: Training dataset (P - popular, NP - not popular)

ID	Price	Parking	Cuisine	Popularity
1	Medium	Available	Mexican	P
2	High	Available	Italian	NP
3	Low	Available	American	P
4	Medium	No	Mexican	NP
5	Low	Available	American	P
6	Medium	No	Italian	P
7	High	Available	Italian	NP
8	High	Available	Mexican	P
9	High	No	American	NP
10	Low	No	Italian	P
11	Low	Available	Mexican	NP
12	Low	Available	Italian	P

• **Purpose:** First, understand information gain and gini index as attribute selection measure for decision tree; second, understand Naive Bayes.

• **Requirement:** Please show your calculation in addition to the final results.

1. [10'] What's the information gain for the "Price" attribute? Please show your calculation.

2. [5'] Now suppose we want to use Gini Index as attribute selection measure. What's the Gini index for the attribute "Parking"? What's the reduction in impurity in terms of Gini Index? Please show your calculation.

3.[10'] Based on the training data, we want to construct a Naive Bayes classifier. Please estimate the following terms (No smoothing is required, and please show your calculation):

- Pr(Popularity = 'P') and Pr(Popularity = 'N')
- Pr(Price = 'Low', Parking = 'Available' , Cuisine = 'Mexican' | Popularity = 'P')
- Pr(Price = 'Low', Parking = 'Available' , Cuisine = 'Mexican' | Popularity = 'N')

4.[5'] Suppose a restaurant has the values: Price = 'Low', Parking = 'Available' , Cuisine = 'Mexican'. Based on the calculation in 3, is this restaurant classified as popular? Please show your reasons.

Question 2 (20 points)

We have eight training points, which are listed and plotted in figure 1; also four test points with their true labels are shown in table 2. Please answer following questions.

- **Purpose:** First, understand K-nearest neighbor classifier; second, understand Perceptron algorithm.
- **Requirement:** Please show your calculation in addition to the final results.

x1	x2	y
1	0.5	+1
2	1.2	+1
2.5	2	+1
3	2	+1
1.5	2	-1
2.3	3	-1
1.2	1.9	-1
0.8	1	-1

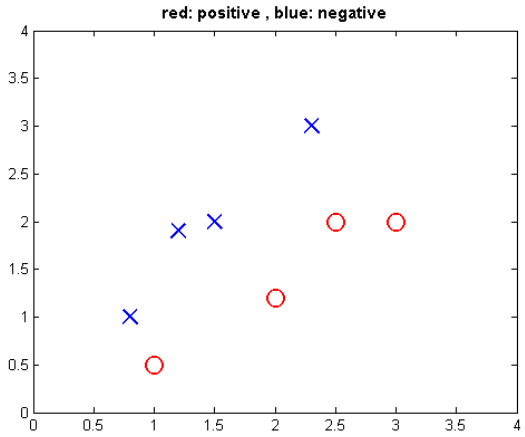

Figure 1: training data

Table 2: Test data with true labels

x1	x2	y
2.7	2.7	+1
2.5	1	+1
1.5	2.5	-1
1.2	1	-1

1.[10'] Perform k-nearest neighbor classification with K = 3. What's the testing error? (Please use Euclidean distance and ties are broken at random. Show your reasoning.)

2.[10'] Is the **training data** linearly separable? Suppose the initial weight of the classifier is $(w_0, w_1, w_2) = (1, 0.5, -1)$ and the learning rate $\eta = 0.1$. Suppose we pick the point $(0.8, 1)$ with label -1. Is this point correctly classified? If not, how to adjust the weight according to the perceptron algorithm?

Question 3 (30 points)

We have nine data points, which are listed and plotted in figure 2.

- **Purpose:** Understand K-Means, one of its issues, and a potential solution to alleviate the issue.
- **Requirement:** All claims/argument need reasonable and brief explanations (2+ sentences).

x	y
1	1
1.5	2
2.5	5
6	5
4	5
4.5	5
3.5	4.5
5	1
6	1

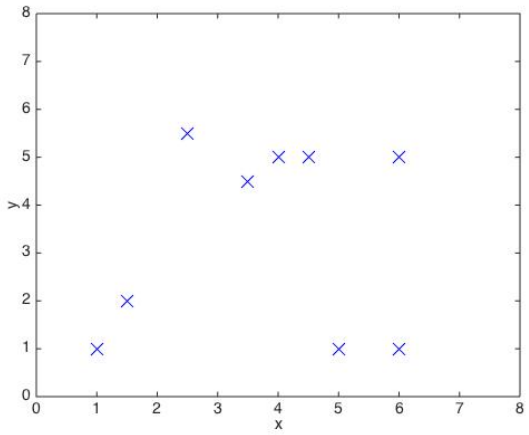

Figure 2: Data

1.[10'] Let k=3. We want to perform k-means using Euclidean distance with initial cluster centers (4.0, 5.0), (3.5, 4.5), and (6.0, 5.0). For each iteration, you need to annotate the graph in figure 2 to show which points belong to which clusters. For example, you might draw a red circle at each point belonging to cluster 1, a green circle at each point belonging to cluster 2, and a blue circle at each point belonging to cluster 3. Also, you need to plot the mean of each cluster with the same color you use for data points in this cluster, but in different point shape to differentiate them from data points. We provide you with the original data graph as a JPG file, you might use annotation or image processing tools such as Mac Preview or Microsoft Paint to annotate the file. Please also show us the coordinates of cluster centers in each iteration. Note:

- We don't allow scanned pictures in your solution. Thus, please use annotation tools instead.
- Each iteration, excluding the final one, needs a graph. We don't need the graph for the final one because it should be the same to its previous one according to the stopping criterion of K-Means.
- You don't have to show numerical distances in each iteration.

2. [8'] Do the same as the above question but with initial cluster centers (2.5, 5.5), (3.5, 4.5) and (6.0, 1.0) instead.

3. [12'] You should realize that the clustering outputs in the above questions are different. Answer the following questions:

1. Why is the difference possible?
2. Which one is better, and why?
3. Could you describe a reasonable quality measurement, and show that we can determine the better one using the measurement?
4. With the example above, you might realize that with wrong choice of initial cluster centers, K-Means will give us very bad clustering output. Could you propose a method to alleviate the issue?

Mini-MP(20 points)

In this Mini-MP, you'll use Weka to do email spam filtering task. We assume you should now be familiar with Weka since you've already seen two demos, and tried frequent pattern mining with Weka.

As shown in the demo, a classification task usually involves four steps: collecting data with labels, preprocessing data, learning a classifier, and testing performance on unseen instances. In this assignment, you're provided with preprocessed email data in ARFF format with the name spam.arff. Please finish the following steps.

- **Purpose:** Learn how to use weka to solve a real-world problem.
- **Requirement:** Please follow the instructions of each question.

1.[5'] The provided file has numeric attributes. Please turn each attribute value into binary value, and show a screen shot, which demonstrates your results.

2.[5'] Hereafter, you're required to train a classifier on 66% of the whole data, and report the performance(accuracy, confusion matrix) on the remain test data. Please apply J48 (decision tree) algorithm in Weka with default setting. Please include a screen shot of the visualized tree(top levels), and write down one rule that you find interesting.

3.[5'] Please play with at least two parameters of J48, discuss briefly about the meaning of each parameter, and report the performance on the test data.

4.[5'] Please apply Naive Bayes algorithm in Weka to the train data, and report the performance on the test data.