CS412 Introduction to Data Mining and Principles
Homework 5

Note: Please show the major calculation steps in your solution.

1 Question 1 (30 points)

Suppose we want to predict whether a restaurant is popular based on its price, delivery, and cuisine, and we collected the training data as shown in table 1. Answer the following questions.

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

ID	Cuisine	Price	Delivery	Popularity
1	Thai	$	Yes	P
2	Korean	$$$	No	P
3	Thai	$$	Yes	NP
4	American	$	Yes	P
5	American	$	No	P
6	Korean	$$	No	NP
7	Thai	$	Yes	P
8	Korean	$$	Yes	P
9	American	$$$	No	P
10	American	$	Yes	NP

1a. Based on the training data, we want to construct a Naive Bayes classifier. (No smoothing is required.) Please estimate the following terms:
1a(i). [4] Pr(Popularity = 'P')
1a(ii). [4] Pr(Popularity = 'NP')
1a(iii). [4] Pr(Price = '$', Delivery = 'Yes' , Cuisine = 'Korean' | Popularity = 'P')
1a(iv). [4] Pr(Price = '$', Delivery = 'Yes' , Cuisine = 'Korean' | Popularity = 'NP')
1b.[6] Suppose a restaurant has the values: Price = '$', Delivery = 'Yes' , Cuisine = 'Korean'. Based on the calculation in part (1a.), is this restaurant classified as popular?

1c. [4] Design an ensemble method for Naive Bayes to further improve the accuracy and briefly describe the steps.
1d. [4] Describe the metrics that can effectively evaluate the classification of data with rare positive examples.

2 Question 2 (40 points)

We have eight training points, which are plotted in figure 1.

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

Table 2: Dataset

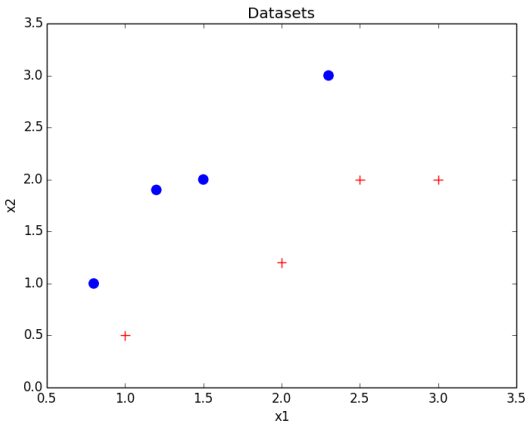

Figure 1: 2-D scatterplot of the Dataset

Also, we have four test points with their true labels. Please answer following questions.

Table 3: Test Dataset

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

2a.[10] Perform k-nearest neighbor classification with K = 1. What's the testing error? (Please use Euclidean distance. Show your reasoning.)

2b. [10] Do the same thing as question 2a with K = 3.

2c. [10] A linear classifier $f(x) = a * x_1 + b * x_2 + c$ works as follows: if $f(x) >= 0$, predict x as +1; otherwise, predict x as -1. Design a reasonable linear classifier(i.e. choose proper a, b, c). What's the training error? What about the testing error? Show your reasoning(Your design doesn't need to be optimal).

2d. [10] Compare KNN and linear classification method(e.g. SVM). You may draw conclusions based on your experience with question 2a-2c.

3 Question 3 (30 points)

Suppose we want to cluster the following 13 points:

index	x1	x2
1	1	3
2	1	2
3	2	1
4	2	2
5	2	3
6	3	2
7	5	3
8	4	3
9	4	5
10	5	4
11	5	5
12	6	4
13	6	5

Table 4: Dataset

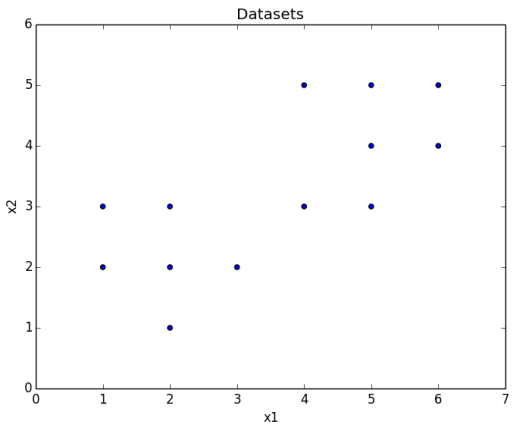

Figure 2: 2-D scatterplot of the Dataset

3a. [10] Cluster above points using K-means algorithm with k = 2. Please use (0, 4) and (6, 5) as the initial center points for the two clusters. Show your reasoning.
3b. [10] Now we want to use DBSCAN, a density-based algorithm, with MinPts = 2, and Eps = 1.5. Outline your clustering process.
3c. [10] Please perform AGNES, a hierarchical clustering algorithm on above points. Please use single link method and adopt Euclidean distance as the dissimilarity measure.