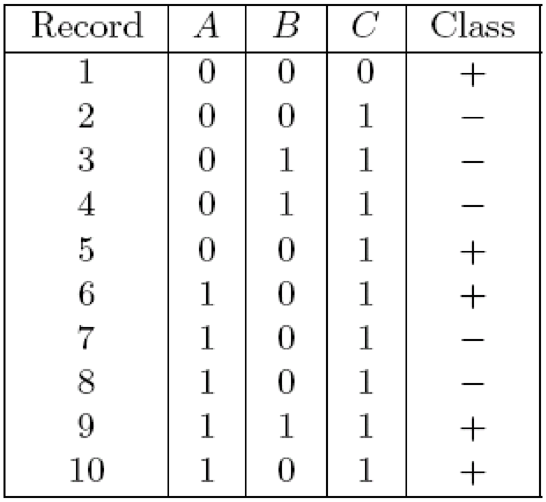
CS 5525: Data Analytics I

Homework 4

Due Date: November 8th , 2016 (4:00PM) **Total: 100 Points**

Problem 1. Bayesian Classification **(15 Points; 2 + 4 + 3 + 4 + 2)**

Consider the data set shown in the following Table:



Record	A	B	C	Class
1	0	0	0	$+$
2	0	0	1	$-$
3	0	1	1	$-$
4	0	1	1	$-$
5	0	0	1	$+$
6	1	0	1	$+$
7	1	0	1	$-$
8	1	0	1	$-$
9	1	1	1	$+$
10	1	0	1	$+$

(a) Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|−), P(B|−), and P(C|−).

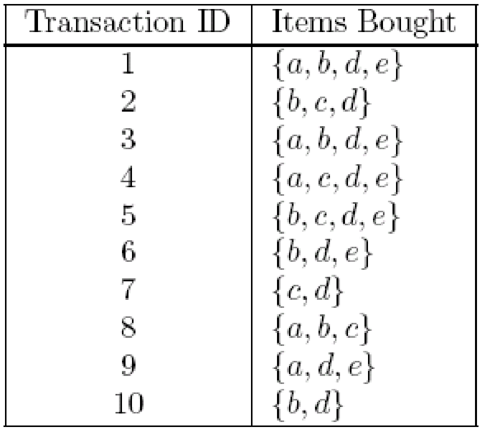
(b) Use these estimates of conditional probabilities to predict the class label for a test sample (A = 0, B = 1, C = 0) using the naive Bayes approach.

(c) Estimate the conditional probabilities using the m-estimate approach, with p = 1/2 and m = 4.

(d) Repeat part (b) using the conditional probabilities given in part (c).

(e) Compare the two methods for estimating probabilities. Which method is better and why?

Problem 2. Association Analysis **(15 Points; 6 + 9)**



Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

(a) Draw a contingency table for each of the following rules using the transactions shown in Table above:

Rules: $\{b\} \longrightarrow \{c\}$, $\{a\} \longrightarrow \{d\}$, $\{b\} \longrightarrow \{d\}$, $\{e\} \longrightarrow \{c\}$, $\{c\} \longrightarrow \{a\}$.

(b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to support and confidence

Problem 3. Association Analysis

(15 Points)

Consider the contingency tables shown below for 3 pairs of items: (bread, milk), (pepsi, coke), and (caviar, wine).

	milk	$\overline{\text{milk}}$
bread	80	120
$\overline{\text{bread}}$	20	30

	coke	$\overline{\text{coke}}$
pepsi	20	80
$\overline{\text{pepsi}}$	100	50

	wine	$\overline{\text{wine}}$
caviar	15	5
$\overline{\text{caviar}}$	20	210

Rank the following six rules (in increasing magnitude): bread --> milk, milk --> bread, coke --> pepsi, pepsi --> coke, wine --> caviar, and caviar --> wine according to the following measures: support, confidence, and Lift.

Problem 4. Support and Confidence **(10 Points)**

Consider the data set shown in the Table below. The first attribute is continuous, while the remaining two attributes are asymmetric binary. Compute the support and confidence for the following rules:

A	B	C
1	1	1
2	1	1
3	1	0
4	1	0
5	1	1
6	0	1
7	0	0
8	1	1
9	0	0
10	0	0
11	0	0
12	0	1

Rule 1:

$$\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}$$

Rule 2:

$$\{(\, A \text{ is odd} \,), B = 1\} \rightarrow \{C = 1\}$$

Rule 3:

$$\{(\, A \text{ is even} \,), C = 1\} \rightarrow \{B = 1\}$$

Problem 5. K-means Clustering **(15 Points; 10 + 5)**

PART I Consider the following six points (with (*x*, *y*) representing location in a 2D space) and let us try to group them into three clusters.

$$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), C_1(1,2), C_2(4,9)$$

The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster. Using the *k-means* algorithm show the:

(i) Cluster assignment of each data point after the first iteration

(ii) Centroids after the first iteration

PART II Consider the following one-dimensional dataset {1,2,3,5,9}. Perform k-means algorithm with 2 clusters and initial centroids are 0 and 9. Compute the following: (i) Final centroids (ii) Cohesion (iii) Separation.

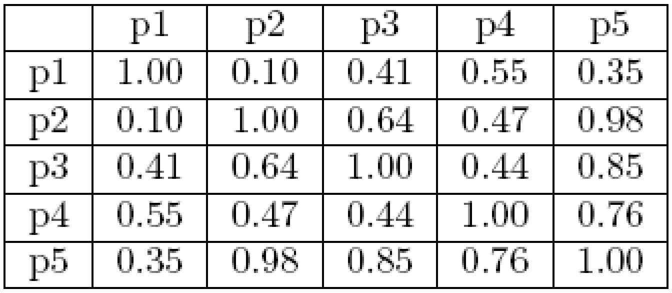
Problem 6. Hierarchical Clustering **(20 Points; 5 + 7 + 8)**

(a) Perform Hierarchical clustering (single Linkage) on the following one-dimensional dataset {0.1, 1, 1.7, 3.4, 3.9, 4.7} (i) If we want to obtain two clusters, show the cluster membership for each datapoint. (ii) Draw the Dendrogram.

(b) Consider the following four data points and use the cosine similarity measure and perform the hierarchical clustering using the single linkage clustering algorithm. Give the proximity matrix and draw the corresponding Dendrogram obtained after clustering.

A: (0 2 0 0); B. (2 0 1 2); C: (2 1 0 2); D: (2 2 1 0)

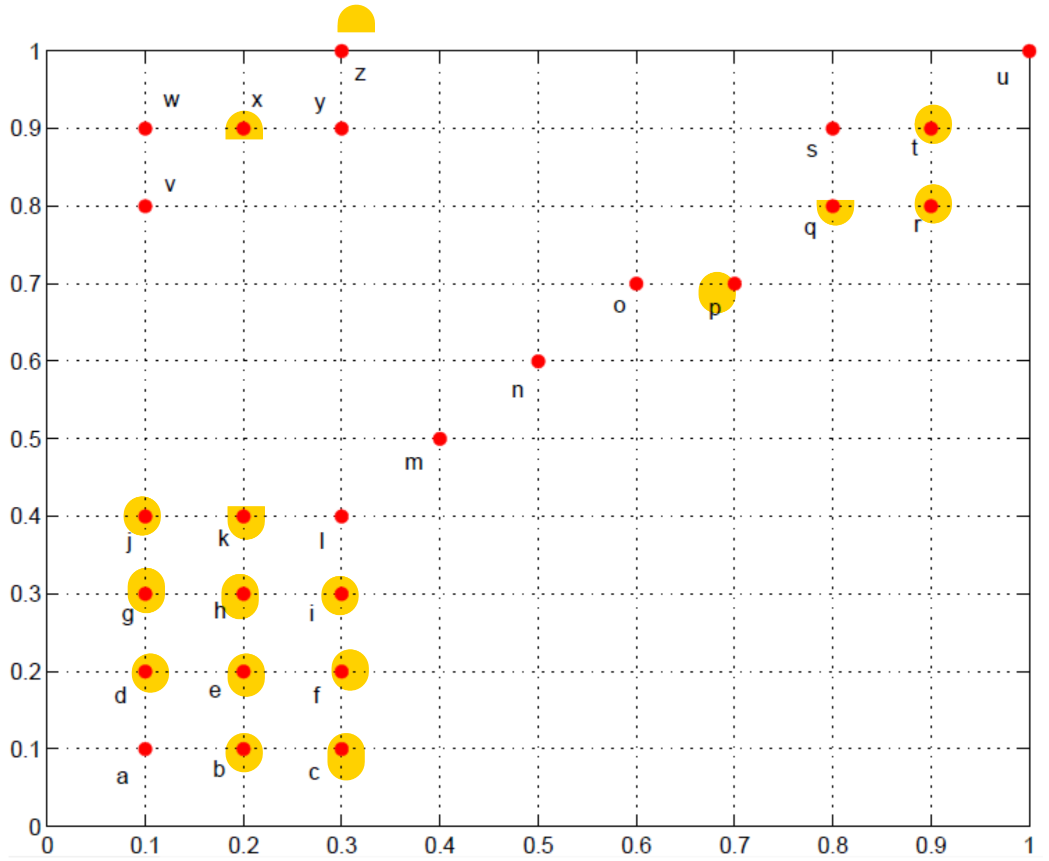
(c) Use the similarity matrix in the following Table to perform single and complete linkage hierarchical clustering. Show your results by drawing a dendrogram that will clearly show the order in which the points are merged. Also, give the updated similarity matrix after each merge.



	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Problem 7. DBSCAN Clustering **(10 Points)**

Consider the data set shown in Figure 3. Suppose we apply DBSCAN algorithm with Eps=0.15 (in Euclidean distance) and MinPts = 3.



List all the core points in the diagram (you can use the labels of the data points in the diagram). Note: a point is considered a core point if there are more than MinPts number of points (including the point itself) within a neighborhood of radius Eps. List all the border points in the diagram. List all the noise points in the diagram.