CS145 Homework Assignment #2 (Fall 2015)

Due: Monday, Oct 26 at noon (the beginning of the class)
Only **hardcopy** is accepted.

1. Constraint based Pattern Mining (10 pts)

Assume that the price of each item is non-negative, and we are only interested in items with a minimum price less than $100. What properties does this constraint have to allow us to prune pattern/data space during the mining process?

Answer:
pattern space: monotone (3pts), succinct (2pts)
data space: data anti-monotone (3pts), data succinct (2pts)

2. Sequential Pattern Mining (20 pts)

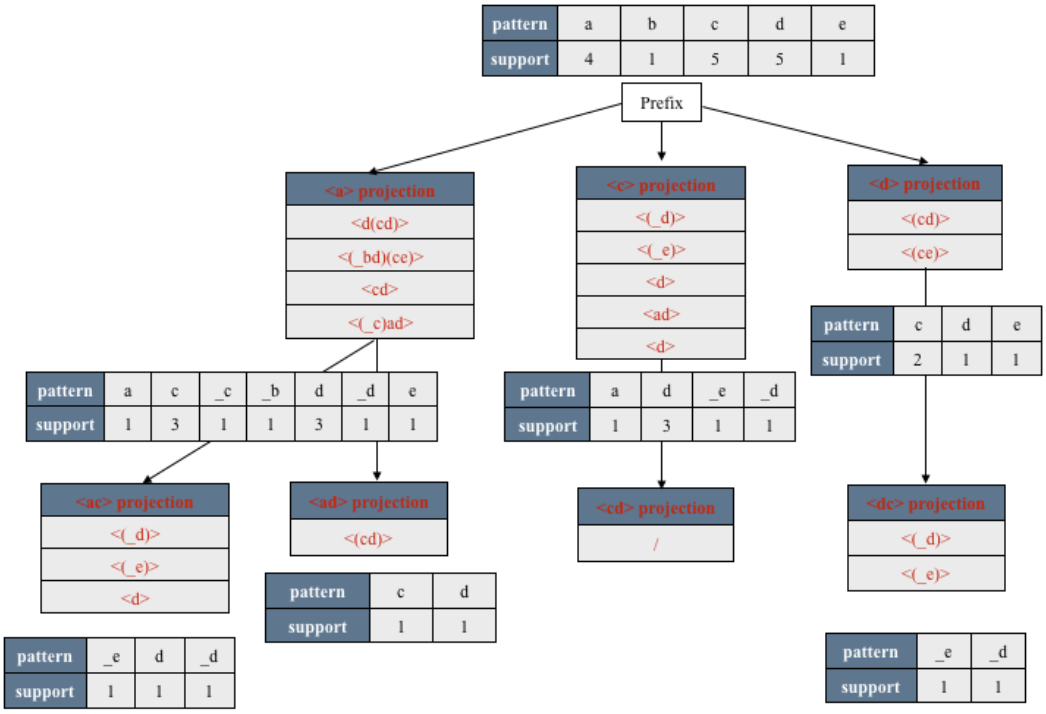
Consider the following database:

Customer ID	Data Sequence
1	<ad(cd)>
2	<(abd)(ce)>
3	<acd>
4	<(ac)ad>
5	<cd>

2.1 Compute the support of these three sequences: (6 pts) (1) <ac> (2) <(ac)> (3) <cd>
2.2 Assume that the minimum support is 2, show the steps of using PrefixScan to find out all the frequent sequential patterns. (14 pts)

Answer:
2.1
<ac> 3; (2pts)
<(ac)> 1; (2pts)
<cd> 3; (2pts)
2.2 Frequent Sequential Patterns are: (7pts)
<a>, <c>, <d>, <ac>, <ad>, <cd>, <dc>.

(list steps) (7pts)



pattern	a	b	c	d	e
support	4	1	5	5	1

<a> projection
<d(cd)>
<(_bd)(ce)>
<cd>
<(_c)ad>

pattern	a	c	_c	_b	d	_d	e
support	1	3	1	1	3	1	1

<ac> projection
<(_d)>
<(_e)>
<d>

pattern	_e	d	_d
support	1	1	1

<ad> projection
<(cd)>

pattern	c	d
support	1	1

<c> projection
<(_d)>
<(_e)>
<d>
<ad>
<d>

pattern	a	d	_e	_d
support	1	3	1	1

<cd> projection
/

<d> projection
<(cd)>
<(ce)>

pattern	c	d	e
support	2	1	1

<dc> projection
<(_d)>
<(_e)>

pattern	_e	_d
support	1	1

3. Hierarchical Clustering (20 pts)

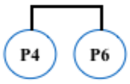
Consider the following points in a 2 dimensional space: P1(2,10), P2(2,5), P3(8,4), P4(5,8), P5(6,4), P6(4,9). Assuming that we use Euclidean distance, show the steps of applying an agglomerative hierarchical clustering algorithm using minimum distance between clusters. You may use dendrogram to illustrate the merge.

Answer:

The initial distance matrix (5pts)

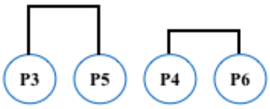
The order of the merge (15pts; 3 each)

	P1	P2	P3	P4	P5	P6
P1	0	√25	√72	√13	√52	√5
P2		0	√37	√18	√17	√20
P3			0	√25	√4	√41
P4				0	√17	√2
P5					0	√29
P6						0



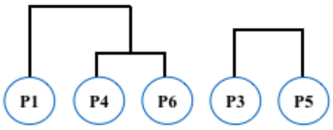
Merge clusters P4 and P6

	P1	P2	P3	P4P6	P5
P1	0	√25	√72	√5	√52
P2		0	√37	√18	√17
P3			0	√25	√4
P4P6				0	√17
P5					0



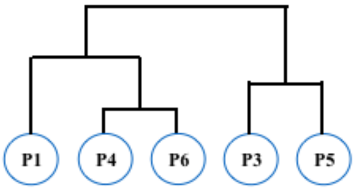
Merge clusters P3 and P5

	P1	P2	P3P5	P4P6
P1	0	√25	√52	√5
P2		0	√17	√18
P3P5			0	√17
P4P6				0



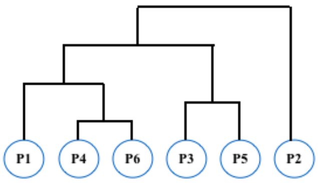
Merge clusters P1 and P4P6

	P1P4P6	P2	P3P5
P1P4P6	0	√18	√17
P2		0	√17
P3P5			0



Merge clusters P1P4P6 and P3P5 *or Merge clusters P2 and P3P5 (Both are right!)*

	P1P4P6P3P5	P2
P1P4P6P3P5	0	√17
P2		0



Merge clusters P1P3P4P5P6 and P2

Use the following points in a 2 dimensional space to answer question 4 and 5.
P1(2,10), P2(2,5), P3(8,4), P4(5,8), P5(7,5),P6(6,4), P7(1,2), P8(4,9).

4. K-means Clustering (20 pts)

The distance function is Euclidean distance. Assuming that K= 3 and initially we assign P1, P4, and P7 as the center of each cluster, show the steps of applying the k-means algorithm to cluster points into three clusters.

Answer:

	P1	P2	P3	P4	P5	P6	P7	P8
Cluster1 (2,10)	0	√25	√72	√13	√50	√52	√65	√5
Cluster2 (5, 8)	√13	√18	√25	0	√13	√17	√52	√2
Cluster3 (1,2)	√65	√10	√53	√52	√45	√29	0	√58

Cluster 1: P1, center (2,10);
Cluster 2: P3,P4,P5,P6,P8, center (6,6); Iteration 1 (5pts)
Cluster 3: P2,P7, center (1.5,3.5).

	P1	P2	P3	P4	P5	P6	P7	P8
Cluster1 (2,10)	0	√25	√72	√13	√50	√52	√65	√5
Cluster2 (6, 6)	√32	√17	√8	√5	√2	√4	√41	√13
Cluster3 (1.5,3.5)	√42.5	√2.5	√42.5	√32.5	√32.5	√20.5	√2.5	√36.5

Cluster 1: P1,P8 center (3,9.5);
Cluster 2: P3,P4,P5,P6, center (6.5,5.25); Iteration 2 (5pts)
Cluster 3: P2,P7, center (1.5,3.5).

	P1	P2	P3	P4	P5	P6	P7	P8
Cluster1 (3,9.5)	√1.25	√21.25	√55.25	√6.25	√36.25	√39.25	√60.25	√1.25
Cluster2 (6.5, 5.25)	√42.81	√20.31	√3.81	√9.81	√0.31	√1.81	√40.81	√20.31
Cluster3 (1.5,3.5)	√42.5	√2.5	√42.5	√32.5	√32.5	√20.5	√2.5	√36.5

Cluster 1: P1,P4,P8 center (3.67,9);
Cluster 2: P3,P5,P6, center (7,4.33); Iteration 3 (5pts)
Cluster 3: P2,P7, center (1.5,3.5).

	P1	P2	P3	P4	P5	P6	P7	P8
Cluster1 (3.67,9)	√3.79	√18.79	√43.75	√2.77	√27.09	√30.43	√56.13	√0.11
Cluster2 (7,4.33)	√57.14	√25.45	√1.11	√17.47	√ 0.45	√1.11	√41.43	√30.81
Cluster3 (1.5,3.5)	√42.5	√2.5	√42.5	√32.5	√32.5	√20.5	√2.5	√36.5

Clusters are the same as the previous iteration, mean values stop change.

Final clusters:

Cluster 1: P1,P4,P8	center (3.67,9);	
Cluster 2: P3,P5,P6,	center (7,4.33);	Iteration 4 (5pts)
Cluster 3: P2,P7,	center (1.5,3.5).	

5. DBSCAN (10 pts)

Assume that (1) we use Euclidean distance, (2) we use DBSCAN with *Eps* = 2 and *MinPts* = 2 (not include the core point). How many core points are there? How many core points are there if we increase *Eps* to 3? (please list the core points of each question)

Answer:

(1) 3; (3pts) P3, P5, P6 (2pts)

(2) 4; (3pts) P3, P5, P6, P8 (2pts)

6. Clustering Algorithms (20 pts)

Describe each of the following clustering algorithms in terms of the following criteria: (1) input parameters that must be specified; (2) limitations.

(a) k-means (b) PAM

(c) BIRCH (d) DBSCAN

Answer:

(a) k-means

1. Number of clusters K (3pts)

2. Sensitive to outliers (1pt), and often terminate at a local optimum (1pt)

(b) PAM

1. Number of clusters K (3pts)

2. Small data sets (not scalable) (2pts)

(d) BIRCH

1. Branching factor: the maximum number of children (1.5pts) Threshold: max diameter of sub-clusters stored at the leaf nodes (1.5pts)

2. Can handle only numeric data (1pt) and sensitive to the order of the data records (1pt)

(d) DBSCAN

1. Radius (1.5pts) Minimum number of points in the neighborhood (1.5pts)

2. All cluster has the same density (in other words, one global density requirement cannot identify clusters with different densities) (1pt) Sensitive to parameters (1pt)