Jennifer MacDonald
604501712
02/22/17

Homework 2

CS 249: Big Data Analytic, Winter 2017
Instructor: Prof. Wei Wang

Due on: Wednesday, February 22, 2017

Instructions

- Please be short and concise in your answers. Do not present unnecessary details.
- Please submit your submission online in CCLE before the due date
- Please mention your full name, UCLA ID and your discussion section on the first page.

Table 1: Coodinates for Problem A

Point	Coodinate
x1	(1,1,1)
x2	(2,1,5)
x3	(6,7,2)
x4	(9,8,7)
x5	(3,5,1)
x6	(5,3,2)

Problem A. K-mean (9+1 pts)

1. Consider the points in table 1 mapped in 3-D space. Assume k=2 and the initial center points are x2 and x4. Using Euclidean distance, run the first iteration of the k-means algorithm. Answer the following: (9 pts)

   (a) Calculate the distances for x1,x2,...,x6 to the initial center points.

From	**To**	**Distance**

x1	**x2**	**$\sqrt{17}$**
x2	**x2**	**0**
x3	**x2**	**$\sqrt{61}$**
x4	**x2**	**$\sqrt{102}$**
x5	**x2**	**$\sqrt{33}$**
x6	**x2**	**$\sqrt{22}$**
x1	**x4**	**$\sqrt{149}$**
x2	**x4**	**$\sqrt{102}$**
x3	**x4**	**$\sqrt{35}$**
x4	**x4**	**0**
x5	**x4**	**9**
x6	**x4**	**$\sqrt{66}$**

(b) Which points are in the first cluster? The second cluster?

First cluster: x2, x1, x5, x6

Second cluster: x4, x3

(c) What are the new center for each cluster after the first iteration?
First cluster: (2.75, 2.5, 2.25)
Second cluster: (7.5, 7.5, 4.5)

Table 2: Coodinates for Problem B

Point	Coodinate
x1	(3,4)
x2	(0,2)
x3	(1,1)
x4	(1,5)

2. How do we know if the algorithm has converged? (1 pts)

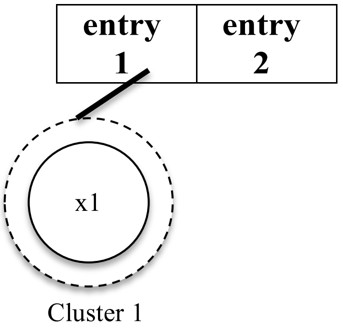
We know that the algorithm has converged when there is no change to the cluster membership.

Problem B. Birch (15 pts)

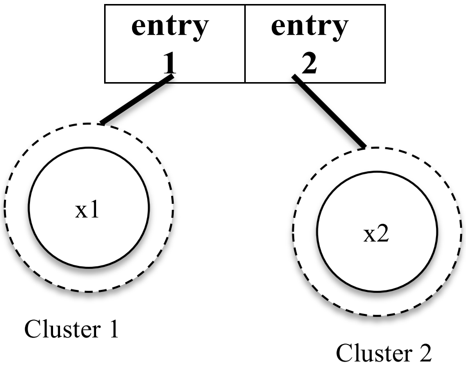
1. Consider the points in Table 2 mapped in 2-D space. Assume the branch factor = 2, # of entry in leaf = 2, and cluster threshold (diameter) = 2. Run the Birch algorithm. For each iteration, draw the corresponding CF-tree. For each leaf node, show the points that are in each sub-cluster. For simplicity, you may assume the memory space is unlimited.

   Note: If the node consist of an odd # of entries when splitting, split in such a way that the left partition consists of one more entry.

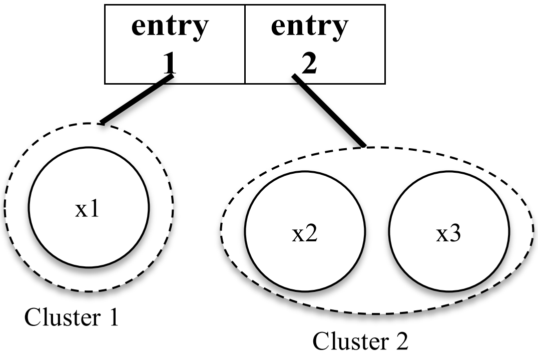
   Iteration 1:
   entry 1 CF: (1, (3, 4), (9, 16))



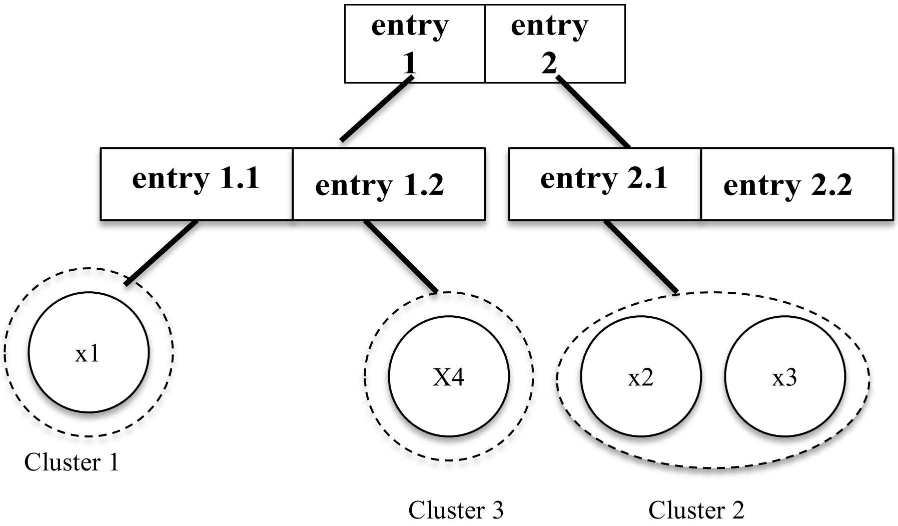
   Iteration 2:
   Tentative entry 1 CF: (2, (3, 6), (9, 20))
   $D = \surd((2(2)(9) - 2(3)^2) + (2(2)(20) - 2(6)^2))/(2(2 - 1))) = \surd 13$



Iteration 3:
x3 to x1: $\sqrt{(2^2 + 3^2)} = \sqrt{13}$
x3 to x2: $\sqrt{(1^2 + 1^2)} = \sqrt{2}$
Tentative entry 2 CF: (2, (1, 3), (1, 5))
$D = \sqrt{((2(2)(1) - 2(1)^2) + (2(2)(5) - 2(3)^2))/(2(2 - 1)))} = \sqrt{2}$



Iteration 4:
x4 to x1: $\sqrt{(2^2 + 1^2)} = \sqrt{5}$
x4 to x2: $\sqrt{(1^2 + 3^2)} = \sqrt{10}$
x4 to x3: $\sqrt{(0^2 + 4^2)} = 4$
Tentative entry 1 CF: (2, (4, 9), (10, 41))
$D = \sqrt{((2(2)(10) - 2(4)^2) + (2(2)(41) - 2(9)^2))/(2(2 - 1)))} = \sqrt{5}$


entry 1
entry 2
entry 1.1
entry 1.2
entry 2.1
entry 2.2
x1
X4
x2
x3
Cluster 1
Cluster 3
Cluster 2

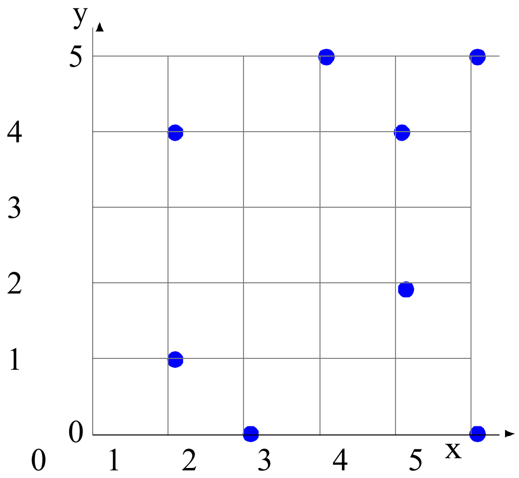


Figure 1. Coordinate Graph for Problem C and D

Table 3: Coodinates for Problem C and D

ID	Coodinate
1	(1,1)
2	(2,0)
3	(5,0)
4	(4,2)
5	(1,4)
6	(4,4)
7	(3,5)
8	(5,5)

Problem C. DBSCAN (15 pts)

1. Consider the points in Figure 1. Assume = 2 and the minPts = 2. Run the DBSCAN algorithm. Points are processed in ascending order of their IDs in Table 3. When determining if a point is a core, assume the point itself is excluded from the count. Identify all the formed clusters and label all the points as either core point, border point, or outlier point. Write your answers as a list of IDs according to Table 3.

ID	**Label**	**Neighbor(s)**	**Cluster**
1	**outlier**	**2**	
2	**outlier**	**1**	
3	**outlier**		
4	**border point**	**6**	**C1**
5	**outlier**		
6	**core point**	**4, 7, 8**	**C1**
7	**core point**	**6, 8**	**C1**
8	**core point**	**6, 7**	**C1**

Problem D. OPTICS (15+5 pts)

1. Describe the ordering scheme using OPTICS for Figure 1. Assume = 3 and minPts = 2. Run the OPTICS algorithm starting from point #8 (5,5). When determining if a point is a core, assume the point itself is excluded from the count. Assume the smaller ID has greater priority if there is a tie. Write down the order in which OPTICS examines the points. (15 pts)

ID	Neighbor(s)	Core Distance	Priority List
8	6, 7	2	6, 7
6	7, ~~8,~~ 4, 5	$\sqrt{2}$	7, 4, 5
7	~~6, 8,~~ 5	2	4, 5
4	~~6,~~ 3, 2	$\sqrt{5}$	3, 5, 2
3	~~4,~~ 2	3	5, 2
5	~~7,~~ 1, ~~6~~	3	2, 1
2	1, ~~4, 3~~	$\sqrt{8}$	1
1	~~2, 5~~	3	Empty

2. How is OPTICS different from DBSCAN? (5 pts)

Optics is different from DBSCAN in that OPTICS can handle clusters of different densities since they cannot all be detected using one global density parameter in DBSCAN. Additionally, OPTICS does not explicitly produce a data set clustering; instead, it outputs a cluster ordering.

Table 4: Dataset for Problem E

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-20	mild	overcast	normal	true	yes
07-21	mild	rain	high	true	yes
07-22	mild	overcast	normal	false	yes
07-23	mild	sunny	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	normal	false	yes

Problem E. Classification (15+10+5 pts)

Table 4 uses four attributes (temperature, weather outlook, humidity, and wind condition) to make a decision on whether to play golf or not. A client asks you to create classification models to predict future play.

1. Draw the entire decision tree. For each splitting decision, calculate the information gain for all attributes using entropy. (15 pts)

I(p,n) = I(4,2) = 0.918

Temperature	p_i	n_i	$I(p_i, n_i)$
mild	4	1	0.722
cool	0	1	0

E(Temperature) = 0.602 + 0 = 0.602

Outlook	p_i	n_i	$I(p_i, n_i)$
overcast	2	0	0
rain	2	1	0.918
sunny	0	1	0

E(Outlook)=0+0.459+0=0.459

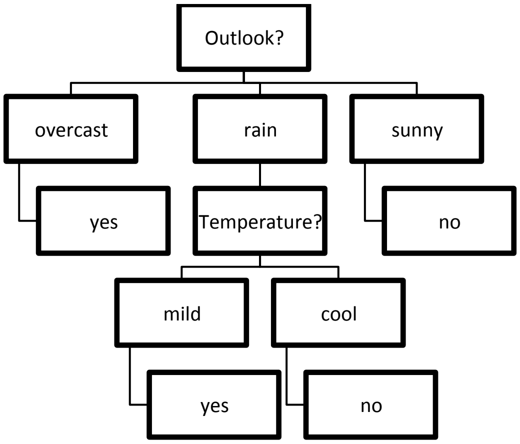
Humidity	p_i	n_i	$I(p_i, n_i)$
normal	3	1	0.811
high	1	1	1

E(Humidity) = 0.541 + 0.333 = 0.874

Windy	p_i	n_i	$I(p_i, n_i)$
true	2	2	1
false	2	0	0

E(Windy) = 0.667 + 0 = 0.667

Gain(Temperature) = 0.316
Gain(Outlook) = 0.459
Gain(Humidity) = 0.044
Gain(Windy) = 0.251



2. Using Naive Bayes Classification, predict the output value for the following condition. Temperature = mild, Outlook = rain, Humidity = high, Windy = false. In your answer, show the final probability values for each class. (10 pts)

P(Temperature = "mild" | Play Golf? = "yes") = 1
P(Temperature = "mild" | Play Golf? = "no") = ½
P(Outlook = "rain" | Play Golf? = "yes") = ½
P(Outlook = "rain" | Play Golf? = "no") = ½
P(Humidity = "high" | Play Golf? = "yes") = ¼
P(Humidity = "high" | Play Golf? = "no") = ½

P(Windy = “false” | Play Golf? = “yes”) = ½
P(Windy = “false” | Play Golf? = “no”) = 0
P(X | Play Golf? = “yes”) * P(Play Golf? = “yes”) = 0.063 * 0.667 = 0.042
P(X | Play Golf? = “no”) * P(Play Golf? = “no”) = 0 * 0.333= 0

The predicted output value is “yes”.

3. Does Naive Bayes take more or less data to make classification decision? Explain your answer. (5 pts)
   A decision takes more data than Naïve Bayes to make a classification decision. This is because using a decision tree requires entropy and information gain, and thus requires more data to build.

Problem F. Programming Assignment (70 pts)

In this assignment, you will be performing clustering methods on the wine dataset. Ignore the output column. You may program in either R, MATLAB, OCTAVE, or python. You are now free to import any libraries. Data Link.

1. One of the features of random forest method is its ability to select the most prominent features from given dataset. Use your language library to select the three features from the wine dataset. You can read more about random forest here and feature selection here.

2. Run the following clustering methods with number of clusters = 3. For any method that require initialization, pick the initialization points randomly. For each method 1. Plot the clusters in a 3-D graph. 2. Calculate the running time in milliseconds.

   (a) K-means
   (b) Birch
   (c) Agglomerative Clustering

3. Re-run the previous problem with number of clusters = {1,5,11}. Comment on any changes to the clustering pattern and/or running time.

4. Run the DB-scan method on the dataset. Plot the cluster in a 3-D graph.

5. Summarize your findings. Briefly explain the advantages and disadvantages of each clustering methods.