Assignment_11

Friday, November 5, 2021 7:59 PM

Assignmen...

Homework Assignment 11 [30 points]

STAT430 Unsupervised Learning - Fall 2021

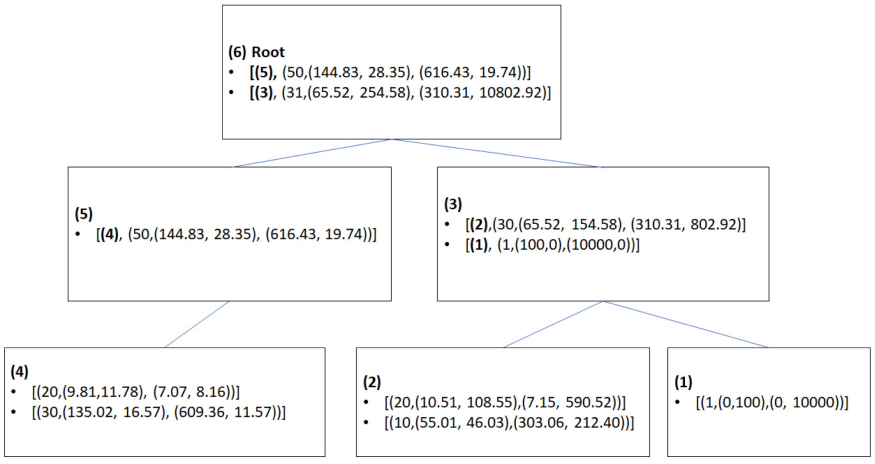
Due: Friday, November 12 on Compass at 11:59pm CST.

#1. BIRCH Clustering ("By Hand")

First, suppose we have already read in several observations into the CF tree below in Phase 1 of the BIRCH algorithm. The BIRCH algorithm that we are using has the following specifications and parameters.

- B = 2
- L = 2
- We are using **radius** to measure cluster size and specifically we will be defining the radius of a cluster to be:
  - : $R_k = \frac{\sum_{i \in C_k} \boldsymbol{dist}(\boldsymbol{x_{i*}}, \boldsymbol{c_k})}{|C_k|} = \frac{\sum_{i \in C_k} ||(x_{i*} - \boldsymbol{c_k})||^2}{|C_k|}$
  - *Or in other words, we are defining the distance between an object $\boldsymbol{x_{i*}}$ and a centroid $\boldsymbol{c_k}$ as the squared Euclidean distance.*
- The **radius threshold** T = 2.5

Current CF Tree



Interpreting the Current CF Tree

1a: How many observations have been read into the CF tree so far?

81 observations. There are 5 subclusters represented in the leaf nodes, represented by the 5 clustering features which show us that these subclusters have 20, 30, 20, 10, and 1 observations, respectively.

1b: Calculate the centroids of each of the 5 subclusters in the leaf nodes.

- Centroid 1: (9.81, 11.78)/20 = (0.49, 0.59)
- Centroid 2: (135.02, 16.57)/30 = (4.5, 0.55)
- Centroid 3: (10.51, 108.55)/20 = (0.53, 5.43)
- Centroid 4: (55.01, 46.03)/10 = (5.5, 4.6)
- Centroid 5: (0,100)/1 = (0,100)

1c: Do you think that any outlier observations have been read into this CF tree? If so, what is this outlier observation?

Yes. There is a subcluster in one of the leaf node entries that is comprised of just 1 observation (represented by the clustering feature (1, (0,100), (0, 10000)). Furthermore, the centroid of this singleton cluster is (0,100), so the single observation in this cluster is also (0,100).

We know that this observation is an outlier for the following reasons.

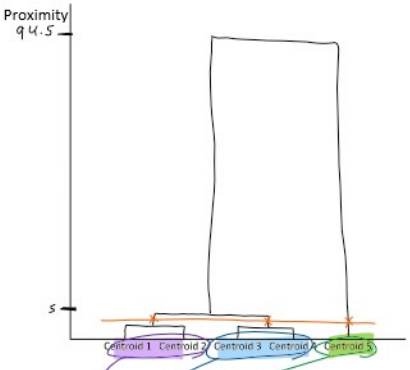
1. This centroid/observation is comparatively much further away from the other four leaf node centroids and
2. The maximum radius allowed for any of these leaf node subclusters is 2.5. So each of the observations in subclusters 1, 2, 3, and 4 will be comparatively close to their respectively centroids and thus, also very far away from the observation (0,100).

Cluster Refinement and Extracting Cluster Labels (Phase 3 and Phase 4 of BIRCH)

What we would like to do next is the following.

a. Refine the clustering structure of the CF tree by using a global clustering algorithm (ie. Phase 3 of BIRCH).
b. Create a global clustering with k=3 clusters. (ie. Phase 3 of BIRCH)
c. And finally, re-read in the first four objects in the dataset and assign them a cluster label (1, 2 or 3).

1d: Cluster the 5 centroids of the 5 leaf subclusters from the tree using hierarchical agglomerative clustering, using single linkage. Display the dendrogram below.



- Centroid 1: (9.81, 11.78)/20 = (0.49, 0.59)
- Centroid 2: (135.02, 16.57)/30 = (4.5, 0.55)
- Centroid 3: (10.51, 108.55)/20 = (0.53, 5.43)
- Centroid 4: (55.01, 46.03)/10 = (5.5, 4.6)
- Centroid 5: (0,100)/1 = (0,100)

1e: Extract the clustering of centroids from the dendrogram above that has k=3 clusters.

- Centroid Cluster 1: {(0.49, 0.59), (4.5, 0.55)}
- Centroid Cluster 2: {(0.53, 5.43), (5.5, 4.6)}
- Centroid Cluster 3: {(0,100)}

1f: Find the new centroids (ie. averages) of the three clusters of centroids that you extracted in 1e.

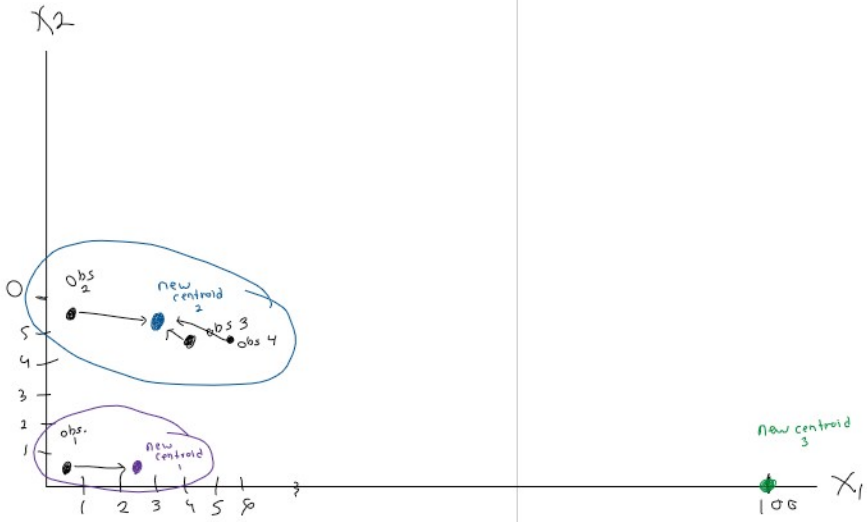
- Centroid Cluster 1 Centroid: ((0.49, 0.59) + (4.5, 0.55))/2 = **(2.5, 0.57)**
- Centroid Cluster 2 Centroid: ((0.53, 5.43) + (5.5, 4.6)))/2 = **(3.015, 5.015)**
- Centroid Cluster 3 Centroid: **(0,100)**

1g: Suppose that the observations below are the first 4 observations in the dataset that were read into the CF tree. Use the procedure discussed in Phase 4 of the BIRCH algorithm to assign cluster labels to each of these 4 observations.

	Dataset	
	x1	x2
Observation 1	0.85	0.34
Observation 2	0.72	5.27
Observation 3	5.09	4.44
Observation 4	5.31	4.90
...	...	...

→ cls. 1

→ cls. 2



Adding a New Observation to the CF Tree

1h: Calculate the cluster radius (using the definition of cluster radius that we defined at the beginning of #1) of a subcluster that has a clustering feature of (21, (10.31, 11.78), (7.32, 8.16)).

Hint: In this problem, you will have to calculate the radius of a (potential) subcluster by just using the clustering feature. You have all the information that you need to solve this problem. This is just an algebraic manipulation problem.

→ Let $C_k = (C_{k1}, C_{k2})$ represent the cluster centroid and $X_i = (X_{i1}, X_{i2})$ be a point.

→ Let $LS_k = (LS_{k1}, LS_{k2})$

$C_k = (C_{k1}, C_{k2})$

$= \left(\frac{LS_{k1}}{N_k}, \frac{LS_{k2}}{N_k}\right)$

$= (10.31, 11.78)$

the cluster centroid and $X_i = (X_{i1}, X_{i2})$ be a point.

→ Let $LS_k = (LS_{k1}, LS_{k2})$

$SS_k = (SS_{k1}, SS_{k2})$

$= \frac{(LS_{k1}, LS_{k2})}{N_k}$

$= \frac{(10.31, 11.78)}{21}$

$= (.49, .56)$

$$R_k = \left[\sum_{X_i \in C_k} (X_{i1} - c_{k1})^2 + (X_{i2} - c_{k2})^2 \right] / N_k$$

$$= \left(\sum_{X_i \in C_k} (X_{i1}^2 - 2X_{i1}c_{k1} + c_{k1}^2) + (X_{i2}^2 - 2X_{i2}c_{k2} + c_{k2}^2) \right] / N_k$$

$$= \left[\left(\sum_{X_i \in C_k} X_{i1}^2 \right) - 2c_{k1} \left(\sum_{X_i \in C_k} X_{i1} \right) + |C_k| c_{k1}^2 + \left(\sum_{X_i \in C_k} X_{i2}^2 \right) - 2c_{k2} \left(\sum_{X_i \in C_k} X_{i2} \right) + |C_k| c_{k2}^2 \right] / N_k$$

$$= \left[SS_{k1} - 2c_{k1}(LS_{k1}) + N_k c_{k1}^2 + SS_{k2} - 2c_{k2}(LS_{k2}) + N_k c_{k2}^2 \right] / N_k$$

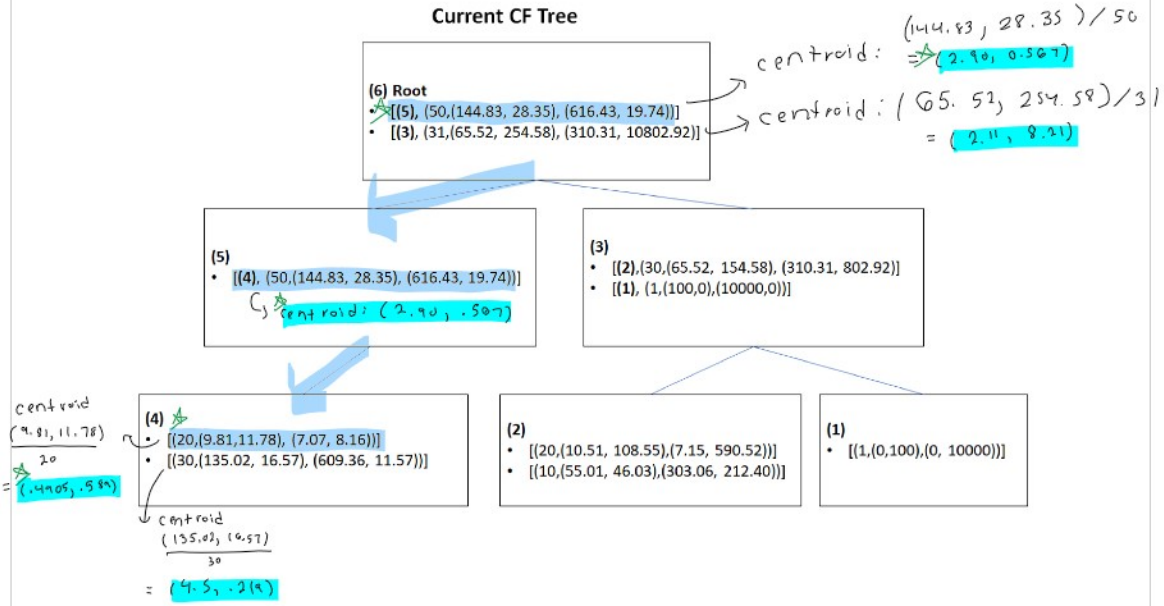
$$= \left[7.32 - 2(.49)(10.31) + 21(.49)^2 + 8.16 - 2(.56)(11.78) + 21(.56)^2 \right] / N_k$$

$$= \boxed{.1814}$$

1h: Finally, suppose that we would like to add a new observation, $(0,0.5)$, to the CF Tree. Add this observation to the CF tree and give the updated CF tree below.

Hint: You may need to use what you calculated in ***1g.***

1. First, we create a singleton cluster with just the point (0,0.5).
2. The clustering feature of this singleton cluster is (1, (0,0.5), (0, 0.25)) and the centroid is (0,0.5).
3. In the root node, the centroid (0,0.5) is closest to the centroid that corresponds to node 5.
4. In node 5, the centroid (0, 0.5) is closest to the centroid that corresponds to node 4.
5. In node 4, the centroid (0, 0.5) is closest to centroid of the first leaf entry in the node.
6. If we were to merge the subcluster comprised of the singleton cluster (0, 0.5) and the subcluster described in the first entry of leaf 4, then we could get the clustering feature of this merged cluster by adding the clustering features of these subclusters together
   a. Merged cluster clustering feature = **(21, (10.81, 11.78), (7.32, 8.16))** = (1, (0, 0.5), (0, 0.25)) + (20, (9.81, 11.78), (7.07, 8.16)).
   b. This is the clustering feature from 1g, which we discovered had a radius of 0.184.
   c. This merged subcluster would have a radius which was below the threshold of 2.5. Therefore we could keep this merged subcluster and update the CF tree accordingly, see below.




CF Tree
(6) Root
(51, (145.83, 28.35), (616.68, 19.73))
[(3), (31,(65.52, 254.58), (310.31, 10802.92)]
(5)
(51, (145.83, 28.35), (616.68, 19.73))
(3)
[(2),(30,(65.52, 154.58), (310.31, 802.92)]
[(1), (1,(100,0),(10000,0))]
(4)
(21, (10.81, 11.78), (7.32, 8.16))
[(30,(135.02, 16.57), (609.36, 11.57))]
(2)
[(20,(10.51, 108.55),(7.15, 590.52))]
[(10,(55.01, 46.03),(303.06, 212.40))]
(1)
[(1,(0,100),(0, 10000))]