Homework Assignment 11 [30 points]

STAT430 Unsupervised Learning - Fall 2021

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

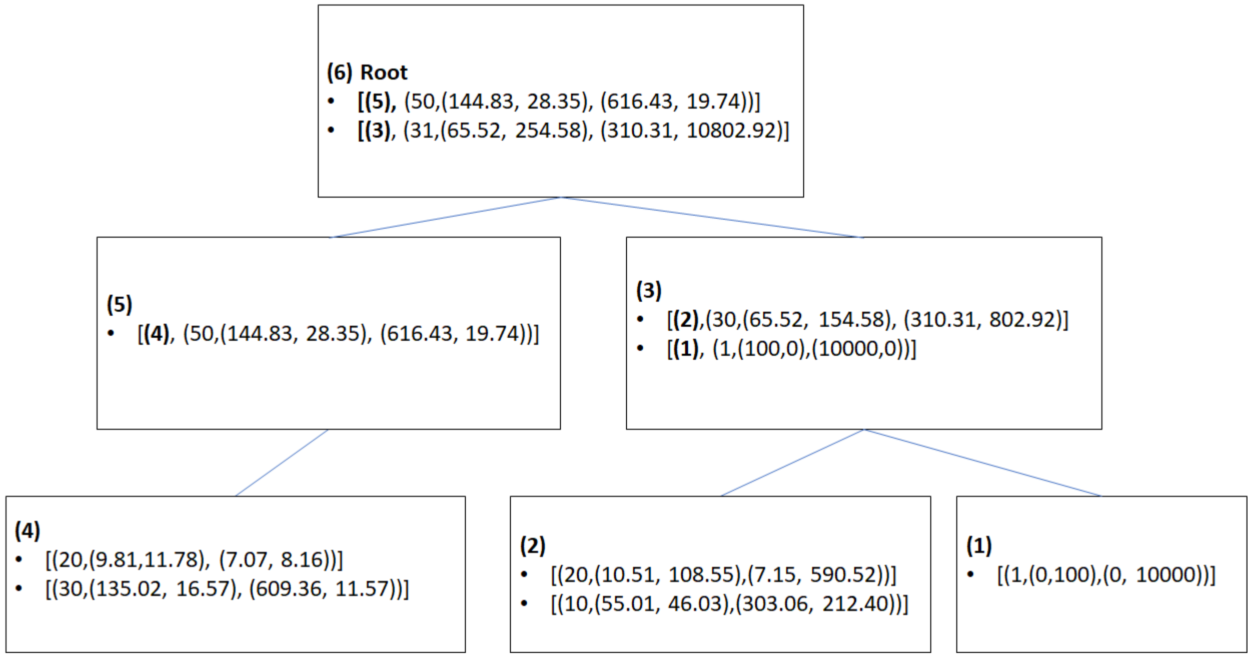
Part 2 and 3 in the Jupyter Notebook

Part 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:
  - : $R_k = \frac{\sum_{i \in C_k} \boldsymbol{dist}(\boldsymbol{x}_{\boldsymbol{i}*}, \boldsymbol{c}_{\boldsymbol{k}})}{|C_k|} = \frac{\sum_{i \in C_k} ||(\boldsymbol{x}_{\boldsymbol{i}*} - \boldsymbol{c}_{\boldsymbol{k}})||^2}{|C_k|}$
  - *Or in other words, we are defining the distance between an object $\boldsymbol{x}_{\boldsymbol{i}*}$ and a centroid $\mathbf{c}_{\boldsymbol{k}}$ as the squared Euclidean distance.*
- The **radius threshold** T = 2.5

Current CF Tree



Interpreting the Current CF Tree

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

1b[1.5 pt]: Calculate the centroids of each of the 5 subclusters in the leaf nodes.

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

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[1.5 pt]: Cluster the 5 centroids of the 5 leaf subclusters from the tree using hierarchical agglomerative clustering, using single linkage. Display the *approximate* dendrogram below.

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

1f[1.5 pt]: Find the new centroids (ie. averages) of the three clusters of centroids that you extracted in 1e.

1g[1.5 pt]: 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
…	…	…

Adding a New Observation to the CF Tree

1h [3 pt]: 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)).

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

1i [3 pt]: 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.

<u>Hint</u>: You may need to use what you calculated in ***1h.***