BIRCH

Unsupervised Machine Learning

Isaac Lee
Jared Mindel
Jeannine Hall
Jed Dryer

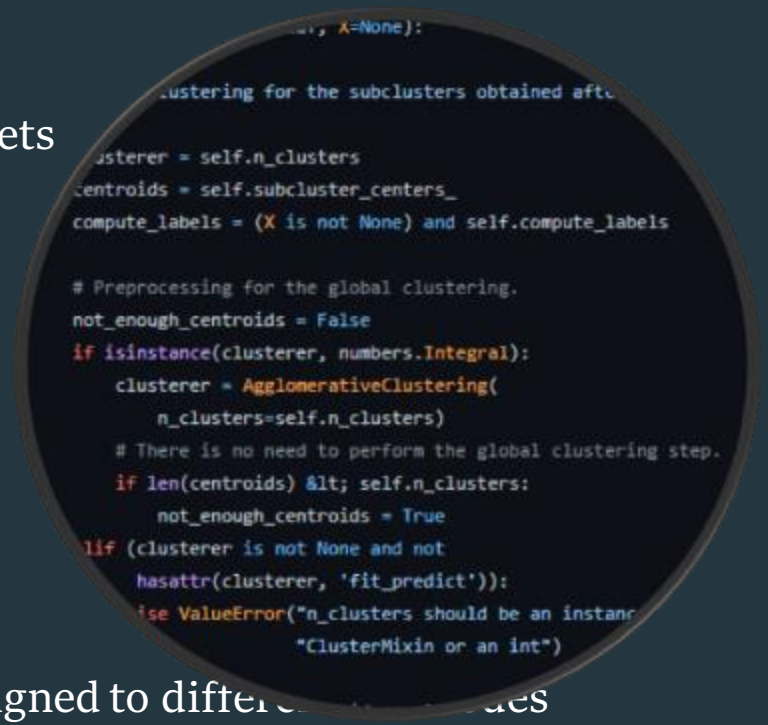
Friday, April 29th, 2022

What is BIRCH?

- **B**alanced **I**terative **R**educing and **C**lustering Using **H**ierarchies
- Unsupervised data mining algorithm
- Hierarchal Clustering

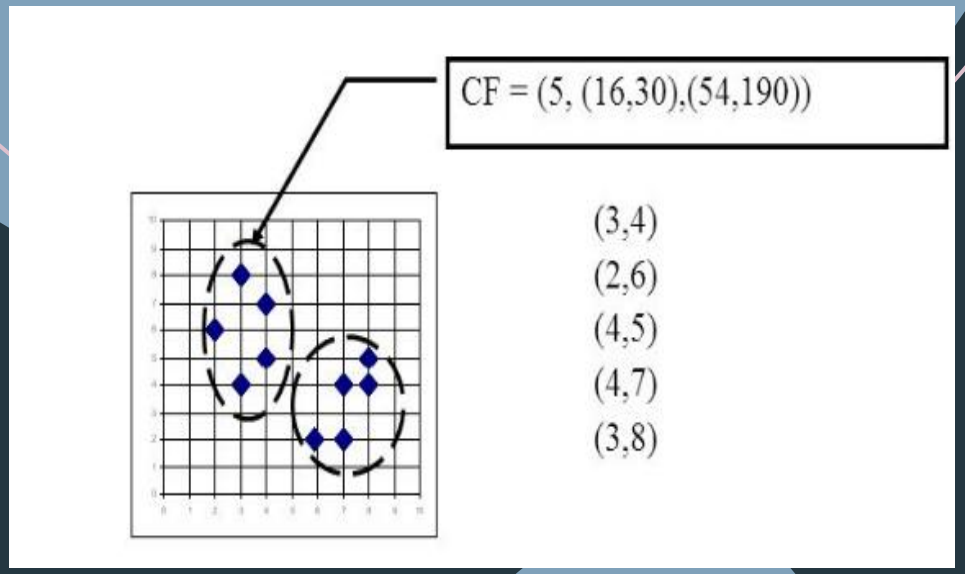
How does the algorithm work?

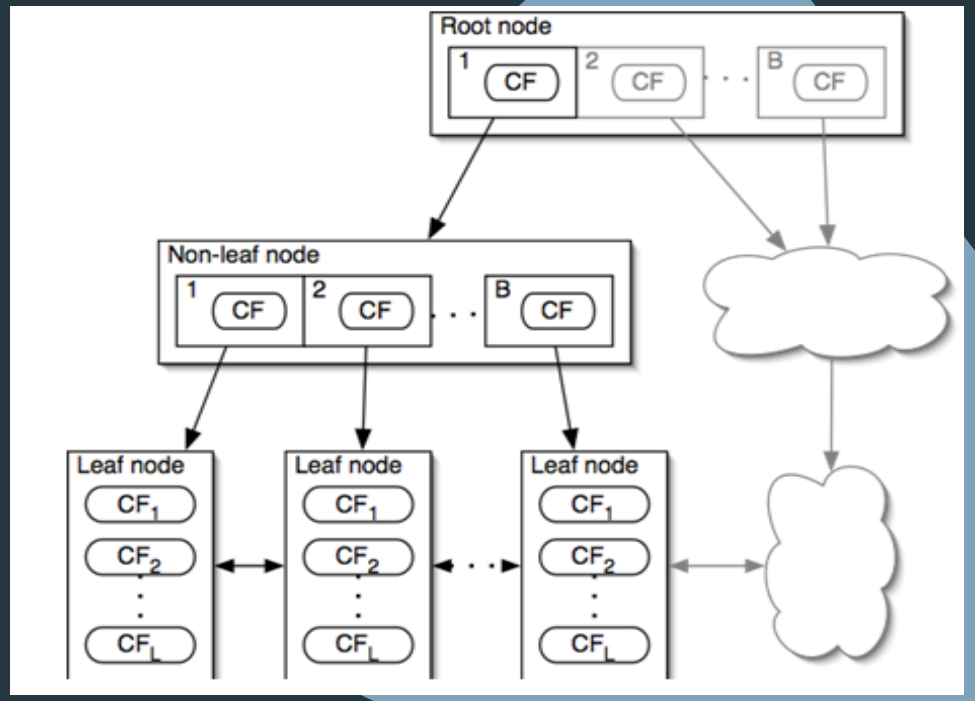
- Scanning data into memory
  - A one-time scan that makes the algorithm efficient for very large datasets
  - Data is fit into Cluster Feature (CF) trees
- Condense data (resize data) (optional)
  - Achieved by adjusting the branching factor and the threshold
- Global clustering
  - Applies an existing clustering algorithm on the leaves of the CF tree
- Refining clusters (optional)
  - Corrects the problem of CF trees where the same valued points are assigned to differ… …ues



How does the algorithm work? (cont.)

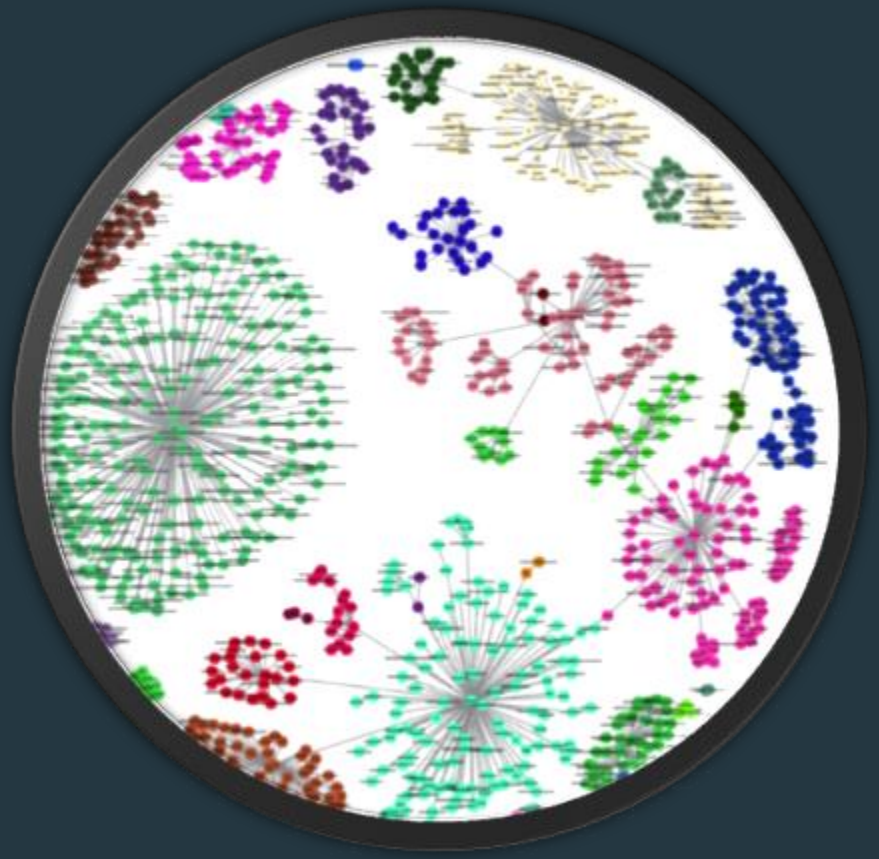
- Cluster Feature (CF)
  - A summary of statistics that represent a set of data points in a given cluster.
  - Count
  - Linear Sum
  - Squared Sum
- Cluster Feature Tree (CF Tree)
  - Height balanced tree that stores cluster feature for hierarchical clustering





What data processing steps are required?

- Data Import
- Data Cleaning
  - Restrict the data to numerical only
    - May require dropping categorical or dummying
  - Check for null values
    - Drop features with majority nulls
    - Consider your goals with the other nulls
  - Standardizing the data is recommended



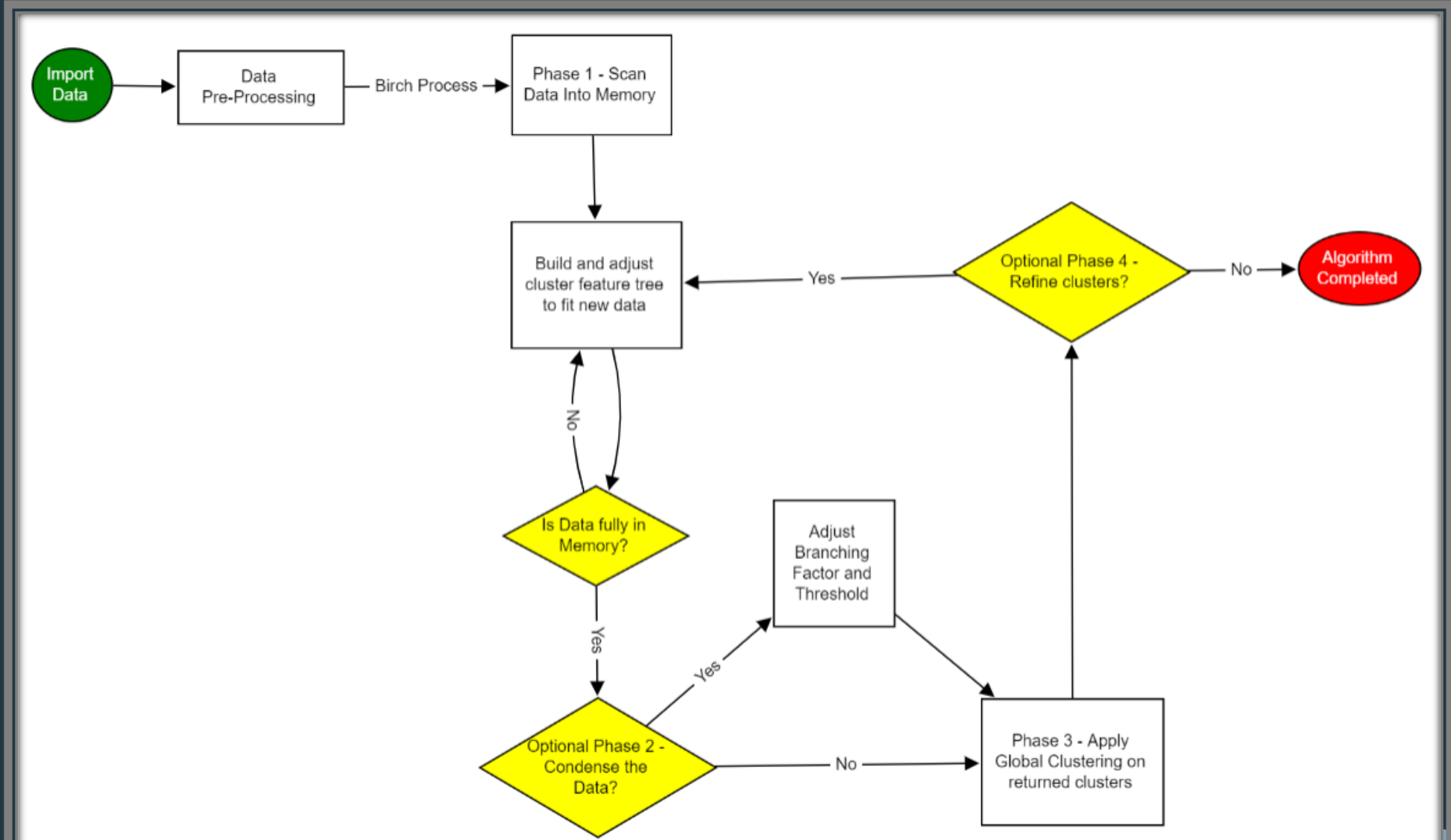
What are the hyper-parameters?

- Threshold
  - Maximum number of subclusters per leaf node
  - Default = .5
- Branching Factor
  - Maximum quantity of CF subclusters per node
  - Default = 50
- Number of Clusters
  - Default = 3

Advantages and Disadvantages

- Advantages
  - Clusters the set of summaries which is more compact than the original dataset
  - Makes full use of available memory, minimizing I/O costs
  - Do not have to select number of clusters (k) at outset
  - Only requires a single scan of the dataset
    - Does not require the whole dataset in advance
- Disadvantages
  - May be dependent on the input ordering of data records
  - Cannot process categorical data
  - Having three hyper-parameters can make it tricky to tune properly

Any Questions?



Appendix

AI for Aspiring Researchers. 2020, July 13th. Clustering with K-Means and Birch Algorithm. YouTube. https://youtu.be/YWcDgX_pN-8?t=501

- Nice code-along video that covers both K-means and Birch algorithms.

Bashirian, M. (n.d.). *Birch Clustering Clearly Explained*. BIRCH Clustering Clearly Explained. Retrieved April 28, 2022, from https://morioh.com/p/c23e0d680669

- Explains how BIRCH works, the clustering feature, and the cluster feature tree

Birch: An efficient hierarchic clustering for large data. Medium. Retrieved April 28, 2022, from https://rafirahim.medium.com/birch-an-efficient-hierarchic-clustering-for-large-data-84f5b9e5c91d

- Good discussion of the algorithm itself and how it works; compares with the K-Means Clustering

Birch in Data Mining - Javatpoint. www.javatpoint.com. (n.d.). Retrieved April 28, 2022, from https://www.javatpoint.com/birch-in-data-mining

- Says how it works and compares it to K-Means. Says advantages. Superfluous relative to the other ones I have, but it can still be used.

Brownlee, J. (2020, August 20). *10 clustering algorithms with python*. Machine Learning Mastery. Retrieved April 28, 2022, from https://machinelearningmastery.com/clustering-algorithms-with-python/

- There is well-commented example code, but there are other algorithms covered too so there isn't really a whole lot specifically on BIRCH but gives a lot of background info on clustering in general

Clustering example with birch method in Python. Clustering Example with BIRCH method in Python. (2019, September 26). Retrieved April 28, 2022, from https://www.datatechnotes.com/2019/09/clustering-example-with-birch-method-in.html

- Example code that uses BIRCH

Clustering. scikit learn. (n.d.). Retrieved April 28, 2022, from https://scikit-learn.org/stable/modules/clustering.html#birch

- Discusses clustering and Birch broadly

Extensive survey on hierarchical clustering methods in … (n.d.). Retrieved April 28, 2022, from https://www.irjet.net/archives/V3/i11/IRJET-V3I11115.pdf

- There is well-commented example code and information on clustering in general

Gupta, A. (2021, June 3). *Balanced iterative reducing and clustering using hierarchies-birch*. Medium. Retrieved April 28, 2022, from https://medium.com/geekculture/balanced-iterative-reducing-and-clustering-using-hierarchies-birch-1428bb06bb38

- Explains history, drawbacks, how it works, parameters

Kharwal, A. (2021, June 26). *Birch clustering in machine learning*. Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript. Retrieved April 28, 2022, from https://thecleverprogrammer.com/2021/03/15/birch-clustering-in-machine-learning/

- This very short blog post provides a very clear and concise example of what BIRCH is, and provides a code along example that clearly illustrates how to use the algorithm with Python

Maklin, C. (2019, July 14). *Birch clustering algorithm example in Python*. Towards Data Science. Retrieved April 28, 2022, from https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9

- This is a helpful code-along and algorithm explanation that is blocked by a paywall. The article is included in the resource file as a pdf

Appendix (Cont.)

Rani, Y., & Rohil, H. (n.d.). *A Study of Hierarchical Clustering Algorithm*. Research India Publications. Retrieved April 28, 2022, from https://www.ripublication.com/irph/ijict_spl/20_ijictv3n10spl.pdf

- This is a study of Clustering Algorithms and has some short segments that very clearly summarize the BIRCH algorithm (section 3.2). The section is two paragraphs long and provides a very high-level summary

sklearn.cluster.Birch. scikit learn. (n.d.). Retrieved April 28, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html

- The documentation for the Birch algorithm in Python; very helpful

Verma, Y. (2021, November 11). *Guide to birch clustering algorithm(with python codes)*. Analytics India Magazine. Retrieved April 28, 2022, from https://analyticsindiamag.com/guide-to-birch-clustering-algorithmwith-python-codes/

Yousaf, S. (n.d.). *What is the sklearn.cluster.birch() function in python?* Educative. Retrieved April 28, 2022, from https://www.educative.io/edpresso/what-is-the-sklearnclusterbirch-function-in-python

- Parameters are explained and it has good code samples

ZHANG, T., RAMAKRISHNAN, R., & LIVNY, M. (1997). (tech.). (U. Fayyad, Ed.) *BIRCH: A New Data Clustering Algorithm and Its Applications* (Vol. 1, Ser. Data Mining and Knowledge Discovery, pp. 141–182). Netherlands: Kluwer Academic Publishers. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.325.7171&rep=rep1&type=pdf

- This resource provides an in-depth look at a wide range of aspects related to BIRCH and how the algorithm functions at its base level. Some of the topics covered include:
  - Contributions and Limitations (2.3)
  - Background (3)
  - Anomalies (4.4)
  - Memory Management (5.2)
  - Parameters and Settings (6.3)