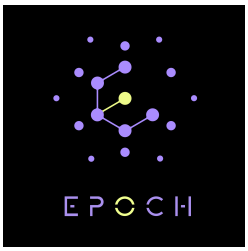

The ML Basket!
Epoch 3: Climb the Tree - Hierarchically

25th August 2024

Background

Hierarchical clustering is a powerful method for finding nested groupings in data by creating a hierarchy of clusters. Unlike flat clustering methods, hierarchical clustering allows you to visualize the data in a tree-like structure called a dendrogram, which can reveal deeper insights into the relationships between data points.

Objective

The goal of this epoch is to implement hierarchical clustering from scratch and apply it to a creative problem that simulates organizing an evolving ecosystem of species based on their traits. This will help you understand the underlying concepts of hierarchical clustering and its practical applications.

Problem Statement

Imagine you are a data scientist working for a futuristic zoo that's constantly evolving its ecosystem of species. Your job is to analyze and cluster species based on their physical and behavioral traits. However, this zoo is special—it evolves over time, meaning new species can emerge and existing species can mutate.

You are provided with a synthetic dataset that contains information about 1,000 species, each described by six traits: Size, Speed, Color, Diet, Habitat, and Aggression. The dataset simulates the evolution of species over 20 time steps, with each species potentially mutating and new species being introduced over time.

Your task is to implement hierarchical clustering and use it to track and visualize how the ecosystem evolves. You'll need to create a system that clusters species and adapts dynamically as new data is introduced, simulating the discovery of new species and mutations.

To make it more interesting, imagine this is a zoo in a digital world, where each species is represented by a unique combination of traits (e.g., color, size, speed, habitat). Your clustering algorithm will help the zoo keep the ecosystem balanced by identifying similar species and grouping them, as well as tracking the evolution of clusters over time.

Tasks

1. **Understanding Hierarchical Clustering:**
   - Research the concepts of agglomerative and divisive hierarchical clustering.
   - Write a brief summary of the differences between these two approaches.
   - Explore different linkage criteria (e.g., single, complete, average, ward) and their impact on the clustering results.
2. **Implementation:**
   - Implement agglomerative hierarchical clustering from scratch in Python, avoiding the use of libraries that directly perform clustering.
   - Implement a dendrogram visualization to display the hierarchy of clusters.
   - Test your implementation using the provided synthetic dataset that represents species with different traits.
3. **Simulating Ecosystem Evolution:**
   - Apply your hierarchical clustering algorithm to the evolving dataset.
   - Track how clusters change over time and visualize the evolution using dendrograms.
4. **Creative Application:**
   - Develop a "Zoo Evolution" tool that allows users to input traits for new species and observe how the ecosystem's clusters adapt.
   - Optionally, implement a feature where users can "mutate" an existing species and see how it affects the overall clustering.
   - Incorporate the ability to save and load different states of the ecosystem to observe and compare different evolutionary paths.

Deliverables

1. **Code:** A well-documented Jupyter notebook or Python script containing your hierarchical clustering implementation, ecosystem simulation, and visualization tools.
2. **Report:** A report summarizing your understanding of hierarchical clustering, the results of your ecosystem simulation, and any insights gained from observing cluster evolution.
3. **Zoo Evolution Tool:** If implemented, provide a link to your tool or instructions on how to run it locally.

Checkpoint Questions

1. Explain the difference between agglomerative and divisive hierarchical clustering.
2. How does the choice of linkage criteria affect the formation of clusters? Provide examples.
3. Discuss the advantages and limitations of using hierarchical clustering for dynamic datasets that evolve over time.
4. What are the computational challenges of hierarchical clustering for large datasets, and how can they be addressed?
5. Describe how dendrograms can be used to make decisions about the number of clusters in hierarchical clustering.
6. Propose a modification to the hierarchical clustering algorithm to better handle real-time data updates.
7. Discuss the ethical implications of using clustering algorithms in applications like wildlife management or conservation.
8. How would you modify your clustering approach to handle traits that are interdependent or correlated?
9. Explain how hierarchical clustering can be integrated with other machine learning techniques to improve ecosystem modeling.
10. Compare and contrast hierarchical clustering with other clustering methods like k-means and DBSCAN in the context of evolving datasets.