🤖

Lecture 12: Hierarchical Clustering

◴ Created	@August 5, 2024 12:09 PM

- We use clustering to group together unlabeled datasets
    - clusters are defined with their centroid and spread
    - the k-means algorithm initializes to random points and then continuously updates the centroids until it hopefully converges to some local optima
        - cons:
            - it doesn’t work well for clusters with different spreads or sizes
- We can use the Mixture of Gaussians approach to mitigate this using the Expectation Maximization algorithm
    - this assigns probabalistic values to how likely a point is in a cluster

Hierarchical Clustering

- Uses the natural relationship between real-world entities (for example species diagram) to create clusters
- allows us to not pick how many clusters we want
- use dendrograms to visualize different granularities of clusters
- pros
    - any distance metric can be used
    - can model more complex cluster shapes by establishing relationships
- the goal of a hierarchical clustering model is to create dendrograms

- there are 2 types of models
    - Divisive (top down)
        - starts with one large cluster and then recursively divides those clusters until we have what we want
        - e.g. recursive k-means
    - Agglomerative (bottom up)
        - starts with a large number of clusters and then combines them until they are together in one big cluster
        - e.g. single linkage
        -
- Assessing performance for clustering algorithms
    - Don’t know
    - more distance between the outer boundaries of each cluster is better
- heterogeneity objective
    - the model is trying to minimize the distance between each of the points and the centroid
    - this is like the error metric
    - we are trying to minimize this value

Detecting outliers for k-means and hierarchical clustering

- if there are clusters with <2-3 datapoints that are far away from all of the other clusters