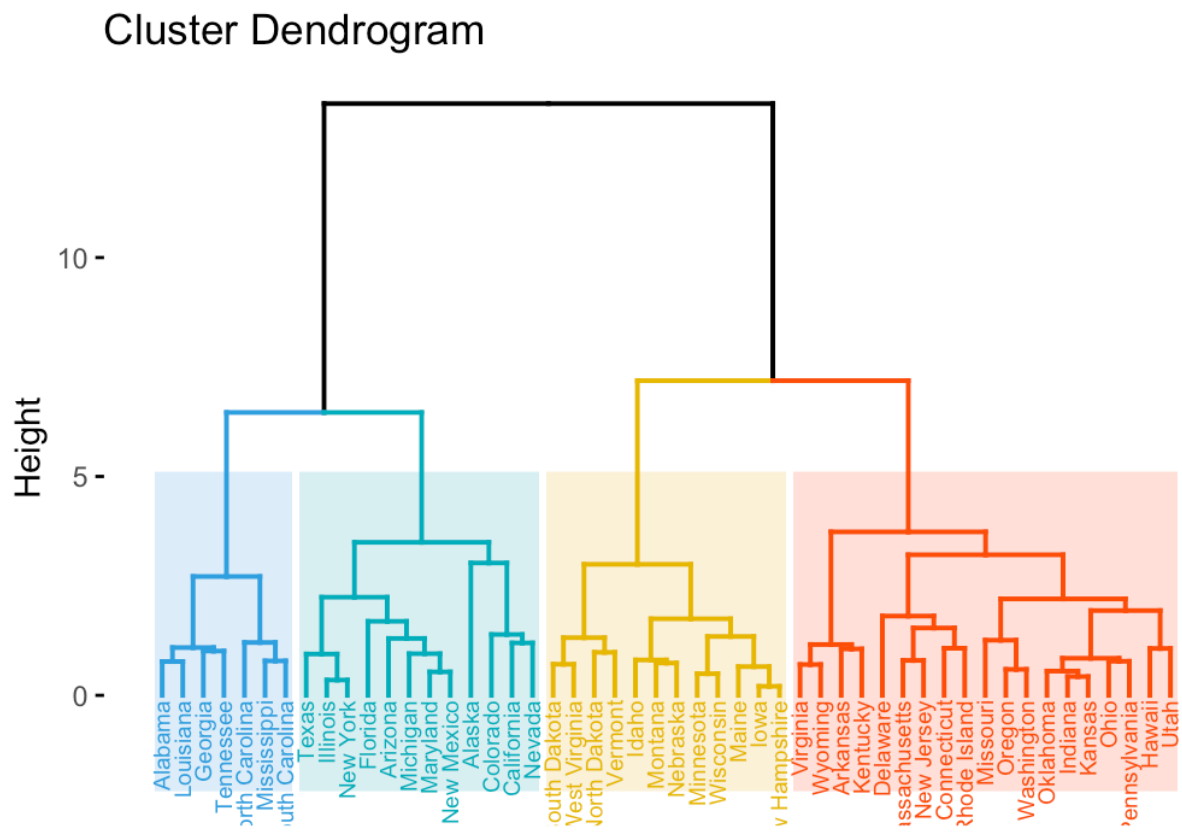
Reflection Week 6

Summary

This week's lectures introduces hierarchial clustering, which applies to cluster problems that cannot be achieved using nearest distance neighbor algorithm used in the previous lectures.

Concepts

- Hierarchial clustering motivation:
    - Avoid choosing number of clusters beforehand (unsupervised learning), in case we don't know how many to begin with.
    - We can use dendrograms to help visualize granularities of clusters.
    - Allows flexibility in the use of distance metrics.
    - More complex cluster shapes.
- Algorithms used in hierarchial clustering:
    - Divisive Clustering:
        - Start with all data in one cluster, then recursively split data using k-means into smaller subclusters.
        - Need to decide some of the hyperparameters: how many clusters per split, when to split and when to stop splitting (max cluster size, cluster 'radius' (which can be determined using distance metrics to determine threshold of the furthest point), and number of final clusters).
    - Agglomerative Clustering:
        - Start with each data point in its own cluster, then recursively merge closest clusters until all points are in one big chunk cluster, based on a predefined distance metric between each cluster.
        - Can be visualized using a dendrogram.
        - Need to decide the distance metric $d(x_i, x_j)$, linkage function (single linkage, complete linkage, and centroid linkage are some of the options), and when/how to "cut" the dendrogram.
        - Very expensive in terms of timing and resources (calculating time for distance between pairs)
- Linkage function calculates the distancebetween each merged pair and other samples:
    - Complete linkage: similarity of farthest pair
    - Single linkage: similarity of the closest pair
    - Group average: similarity between groups
    - Centroid similarity (Ward's linkage): merge clusters with the most similar central point
- Dendrogram



A dendrogram has x and y axes:

- x-axis shows the data point, which is arranged in a very particular order as one can see above for readability
- y-axis shows the distance between pairs of clusters

"Cut" the dendrogram at a distance "D" (in the image above, value of D is 5)

Concerns

- In which case do we prefer each hierarchial clustering algorithm to another?
- How to choose the perfect linkage function? Or is it a matter of trial and error?