Clustering Chapter

When to use kmeans versus hierarchical

* Kmeans is relatively quick to converge, but does not preserve the underlying relationships between points and the clusters may be unstable due to the random initialization.
* Hierarchical clustering is quite flexible in its assumptions, captures rich interlinkages between each point and cluster, and can be reproduced without error. The challenge with hierarchical clustering is that it is computationally costly – depending on the linkage method, the number of calculations required are compounded with sample size. Thus, hierarchical clustering is generally suitable for smaller samples.
* In computer science, \*big O\* notation is used to describe the processing time or memory requirements of a function.

Hierarchical clustering

* Another clustering technique that captures how each observation is related to all others in a tree-like structure. Unlike k-means, hierarchical clustering provides context of the interrelationships amongst observations with a greater degree of flexibility, but with the tradeoff of large time costs.
* Two types: agglomerative and divisive.
  + Agglomerative
  + Divisive
* Describe how it works
  + Clustering pseudo code
  + Distances
* When to apply

|  |  |  |
| --- | --- | --- |
| Linkage distance | Formula | Choose clusters based on: |
| Single Linkage |  | distance between the two closest points in two clusters |
| Complete Linkage |  | Distance between the two farthest points in two clusters |
| Average Linkage |  | Average distance between all points |
| Centroid Distance |  | Distance between cluster centroids |
| Ward’s Distance |  | Clusters that minimize variance increase by using an ANOVA sum of squares overall all partitions |