18.6.4

K-means vs. Hierarchical Clustering

Hierarchical clustering seems like a fairly interesting idea, but you wonder what the differences are between K-means and hierarchical.

The K-means algorithm is the main algorithm we used in this module. It is easy, runs relatively quickly, and can scale to large datasets. This is not to say there aren't drawbacks to the K-means algorithm.

Behind the scenes, K-means is dependent on random initialization, so the outcome depends on a random seed. With K-means, you need to have an idea of how many clusters you're looking for ahead of time, which might not always be known. This can be an issue when the points of data are not so clearly grouped into clusters, as K-means works best for spherical-looking data with similar density points closely grouped together.

With hierarchical clustering and the use of dendrograms, it's easier to pick how many clusters we want without making any assumptions since a *K* value does not need to be known ahead of time.

The dendrogram might not always create as clear of a choice as we would like, and it leaves the final decision up to the analyst. With the iris dataset, we knew the *K* value ahead of time, so using K-means in that situation would make more sense. Hierarchical clustering might not work as well on larger datasets because it is slower at run time, and there are a lot of decisions to be made about when to merge groups of clusters.

NOTE

Both clustering algorithms have their pros and cons. Read the **No Free Lunch (NFL) theorem (https://en.wikipedia.org/wiki/No_free_lunch_theorem)** , which states that there will always be times when one algorithm outperforms the other, and vice versa.

Match the following clustering algorithm with the appropriate description.

		K-means	Hierarchical	Both
A	Requires the clusters known ahead of time.	○	○	○
B	Still leaves it up to the user to analyze.	○	○	○
C	Uses a dendrogram.	○	○	○
D	Uses the elbow curve.	○	○	○
E	Groups based on a distance metric.	○	○	○
F	Is good for large data.	○	○	○
G	Is good for grouping data.	○	○	○

Check Answer

Finish ▶