Unsupervised Learning, Part 4

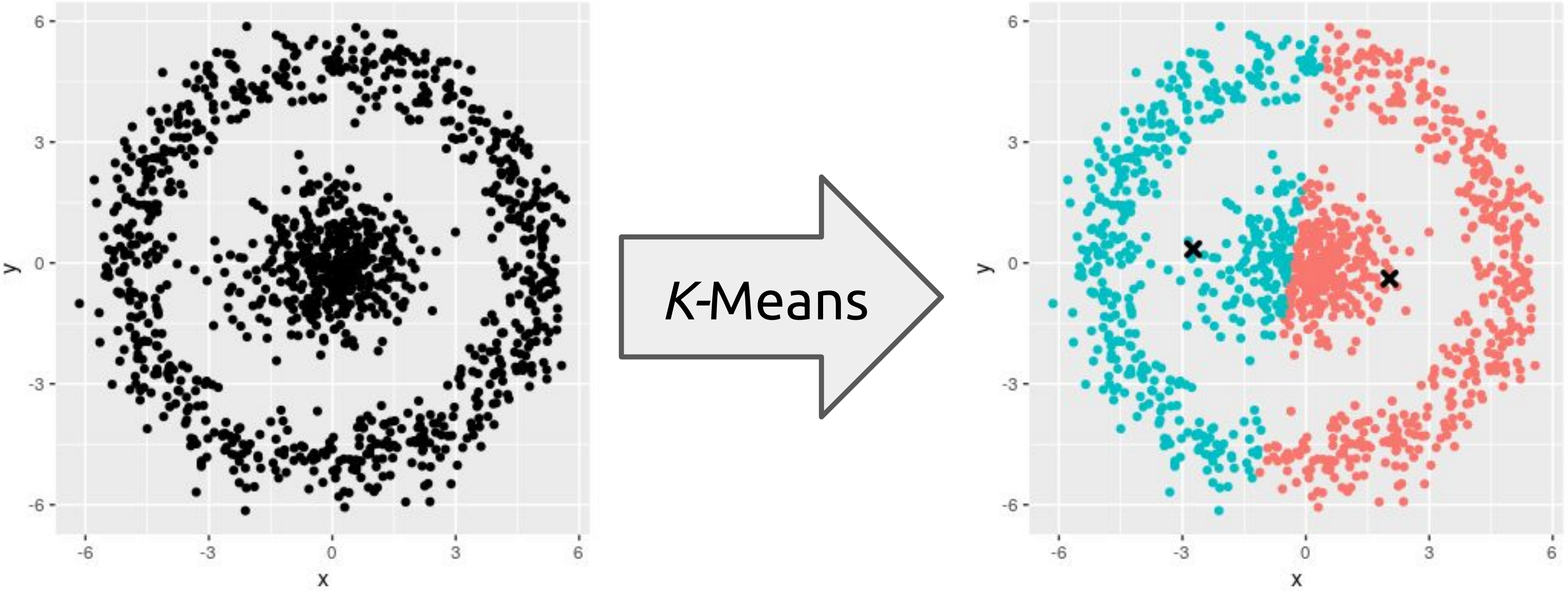
Density Based Clustering and (H)DBSCAN

K-Means Clustering

Drawbacks of K-Means Clustering:

- Must decide how many clusters to use
- Assumes that clusters are spherical
- Does not behave well in the presence of outliers - all points must be assigned to a cluster

K-Means Clustering



K-Means

DBSCAN

DBSCAN = **D**ensity-**b**ased **S**patial **C**lustering of **A**pplications with **N**oise

Density determined by the number of nearby points.

Assumes that clusters will have high density near the “core” and lower density near the edges.

Labels low-density points as outliers

Requires specifying a distance and number of neighbors cutoff for determining density.

HDBSCAN

HDBSCAN = **H**ierarchical Density-based Spatial Clustering of Applications with Noise

Builds clusters from the "ground up".

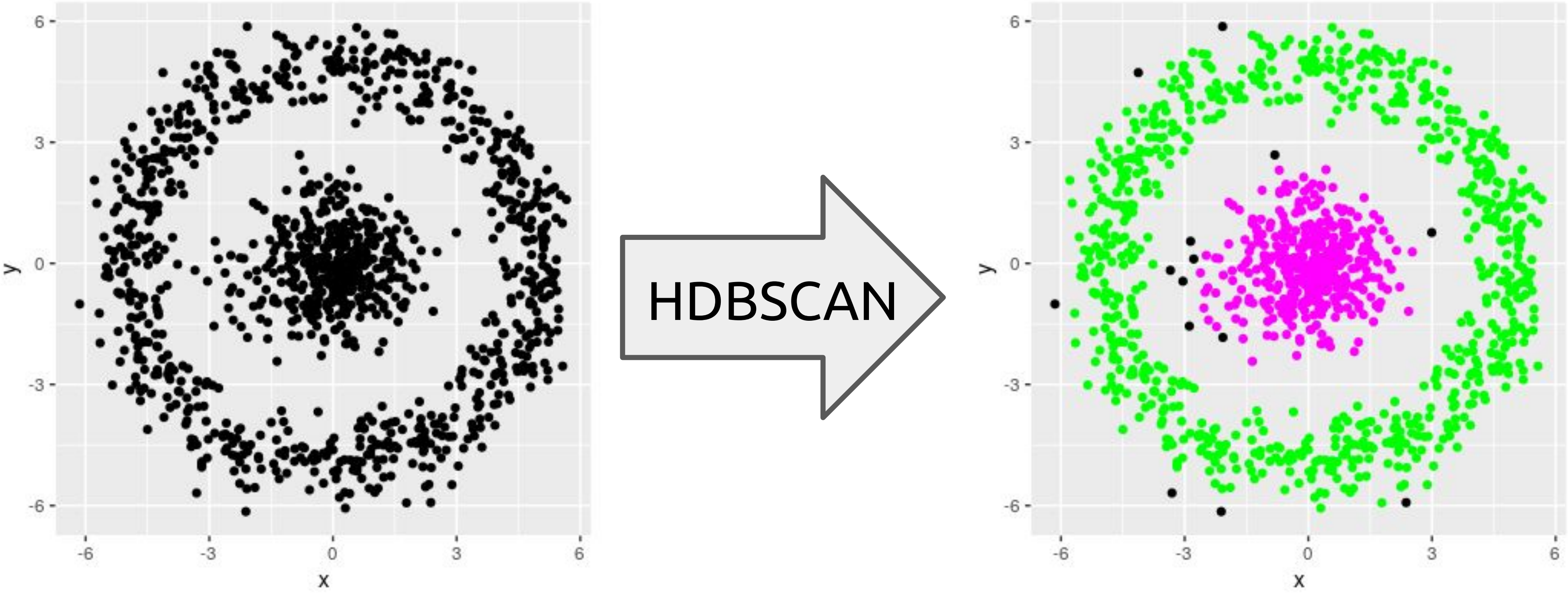
Looks through a range of possible hyperparameter values and chooses a stable clustering.

Don't have to specify the distance, density cutoff, or desired number of clusters.

See https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html for more details of the algorithm.

HDBSCAN

Black points in the result are “outliers”



Example Notebook

HDBSCAN.Rmd