H_2O.ai

Redundant Data

- Features can be highly interrelated.
- Not all features are related to the target.

- Select representatives from clustered / grouped features
- Use dimensionality reduction techniques (PCA, GLRM, MCA, MDS)

Real Data

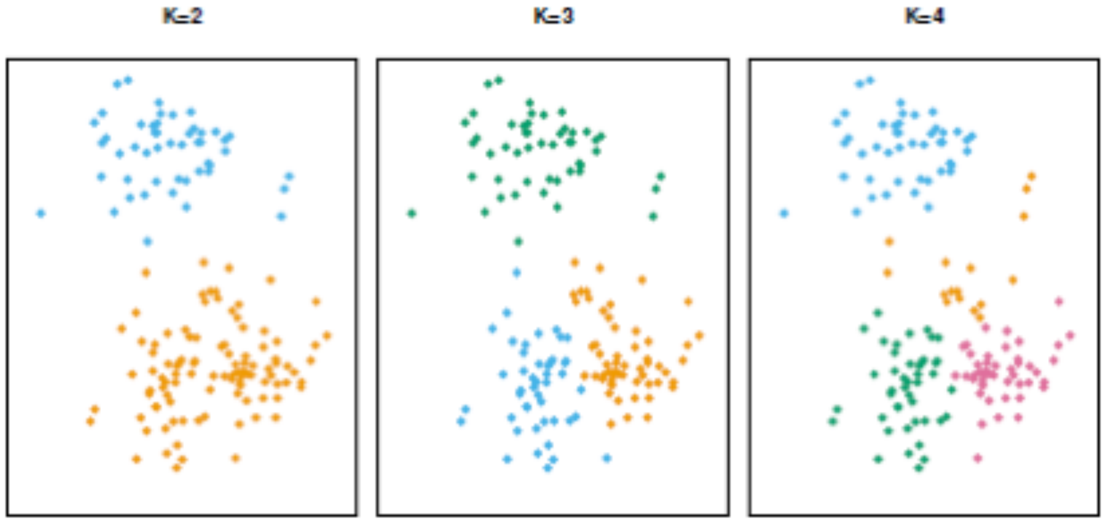
Solutions

Too much of the same is not a good thing

- Leads to numeric instability in machine learning algorithms (GLM)
- Overweights importance of redundant features (RF, GBM)

K-Means Clustering

- K-Means clustering groups observations based on numeric features
  - ○ Assumes clusters are roughly the same sized hyperspheres
  - ○ Minimize Euclidean distance between observations and cluster centers
- Number of methods for choosing the number of clusters, k
  - ○ Choose several and evaluate performance
  - ○ Use business rules



Redundant Data

Real Data

- Features can be highly interrelated.
- Not all features are related to the target.

Too much of the same is not a good thing

- Leads to numeric instability in machine learning algorithms (GLM)
- Overweights importance of redundant features (RF, GBM)

Solutions

- Select representatives from clustered / grouped features
- Use dimensionality reduction techniques (PCA, GLRM, MCA, MDS)