Introduction to Data Science

Daniel Gutierrez, Data Scientist

Los Angeles, Calif.

Course Outcomes

- Deploy unsupervised machine learning methods for knowledge discovery

Lesson Objectives

- Overview of unsupervised learning methods
- Manually step through process yielding distinct clusters showing groupings and similarities in the data
- Review the hierarchical clustering algorithm using R's `hclust()` function to compute clusters and use data viz to display
- Review the K-means clustering algorithm using R's `kmeans()` function to compute clusters and use data viz to display

Unsupervised Machine Learning

- Can we find things that are close together?
  - How do we define close?
  - How do we group things?
  - How do we visualize the grouping?
  - How do we interpret the grouping?

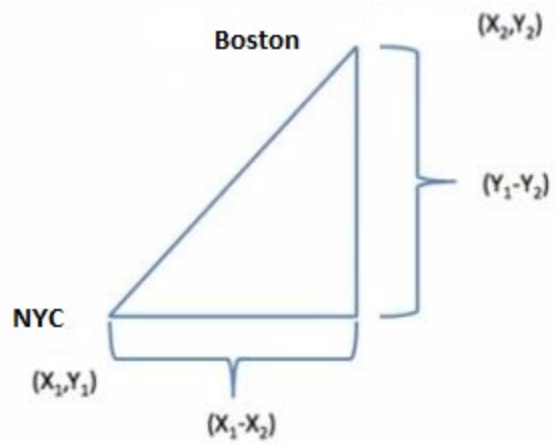
Unsupervised Machine Learning

- How do we define close?
  - Most important step is selecting appropriate distance measure - garbage in -> garbage out
  - Distance or similarity
    - Continuous – Euclidean distance
    - Continuous – correlation similarity
    - Binary – Manhattan distance
  - Pick a distance / similarity that makes sense for your problem

Unsupervised Machine Learning

Euclidean Distance

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$



$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \ldots + (Z_1 - Z_2)^2}$$

Unsupervised Machine Learning

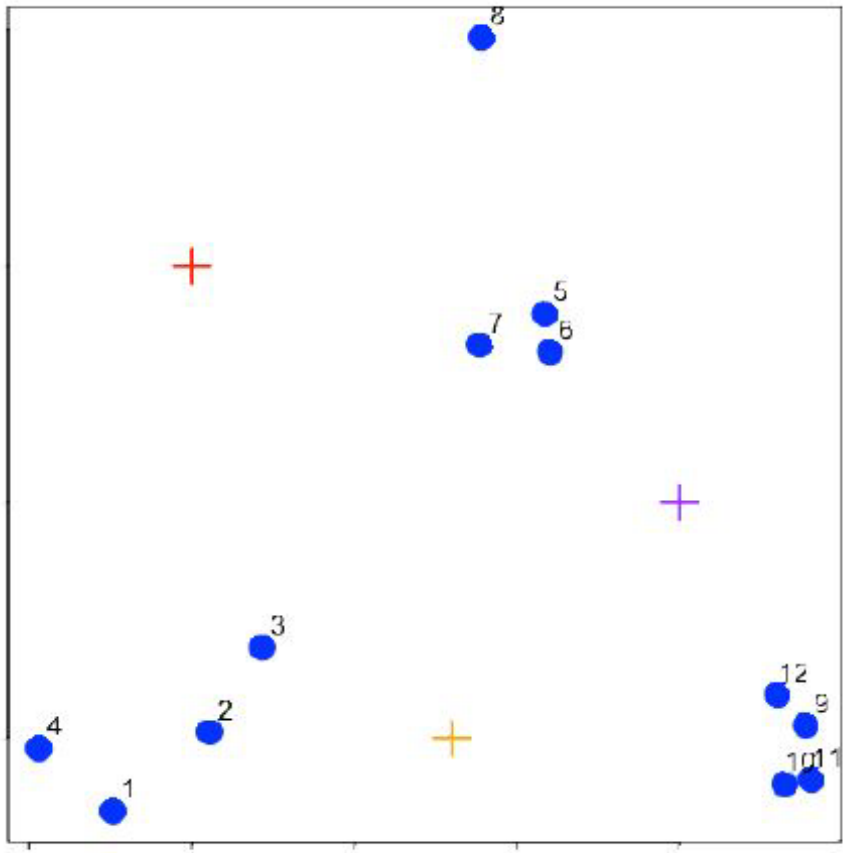
- Hierarchical clustering
  - Compute the distance between every pair of points/clusters:
    - Computing the distance between point A and point B is via the distance function
    - Computing the distance between point A and cluster B may first compute distance of all point pairs (one from cluster A and the other from cluster B) and then pick either min/max/avg of these pairs
  - Combine the two closest point/pairs into a cluster. Repeat step 1 until only one big cluster remains

Unsupervised Machine Learning

- K-means clustering
  - A partioning approach
    - Fix a number of clusters
    - Get “centroids” of each cluster
    - Assign things to closest centroid
    - Recalculate centroids
  - Requires
    - A defined distance metric
    - A number of clusters
    - An initial guess as to cluster centroids
  - Produces
    - Final estimate of cluster centroids
    - An assignment of each point to clusters

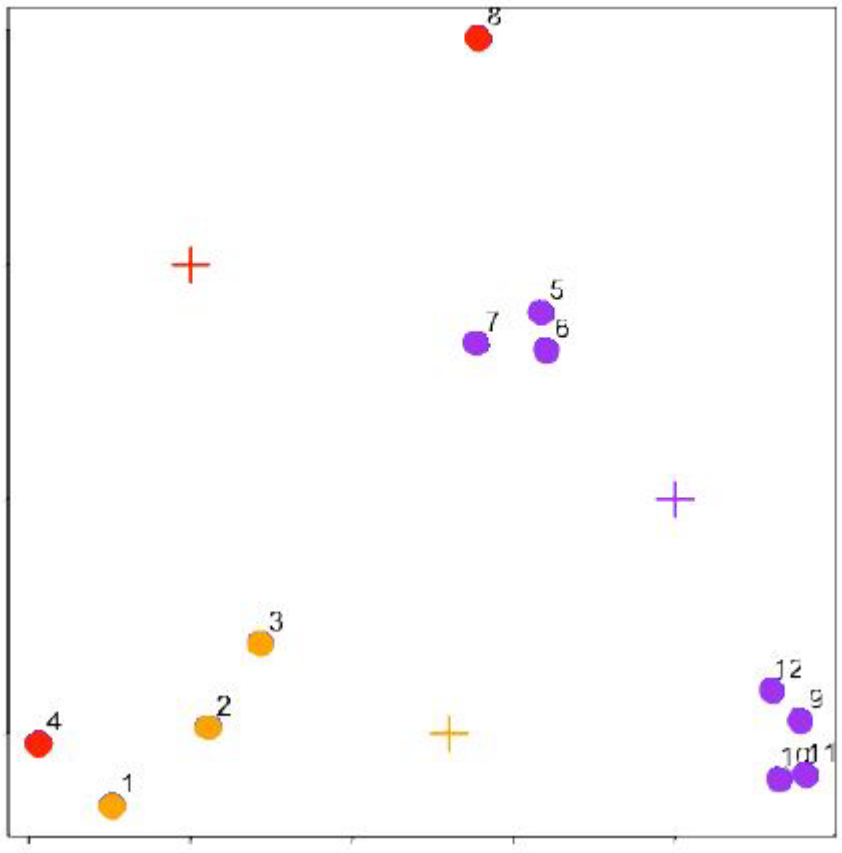
Unsupervised Machine Learning

- K-means clustering – starting centroids



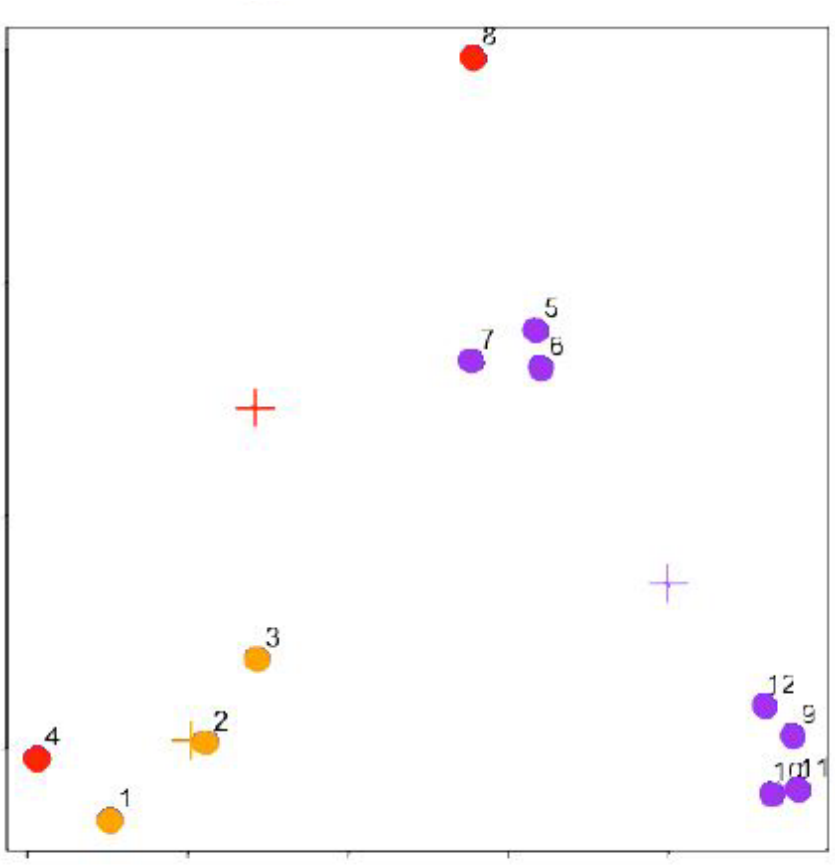
Unsupervised Machine Learning

- K-means clustering – assign to closest centroid



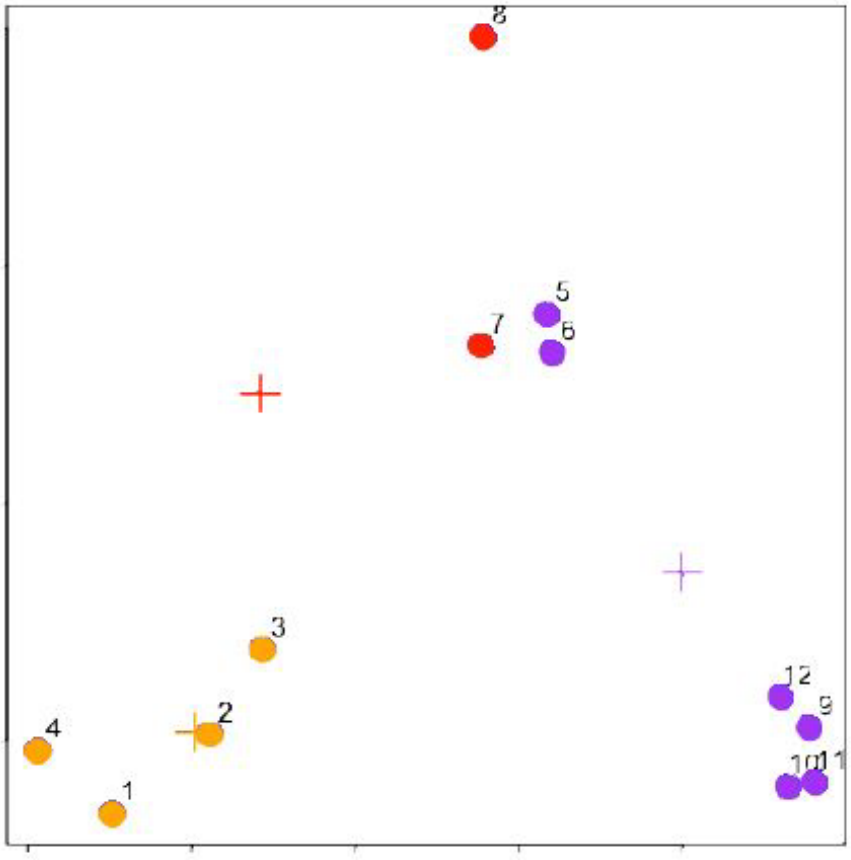
Unsupervised Machine Learning

- K-means clustering – recalculate centroids



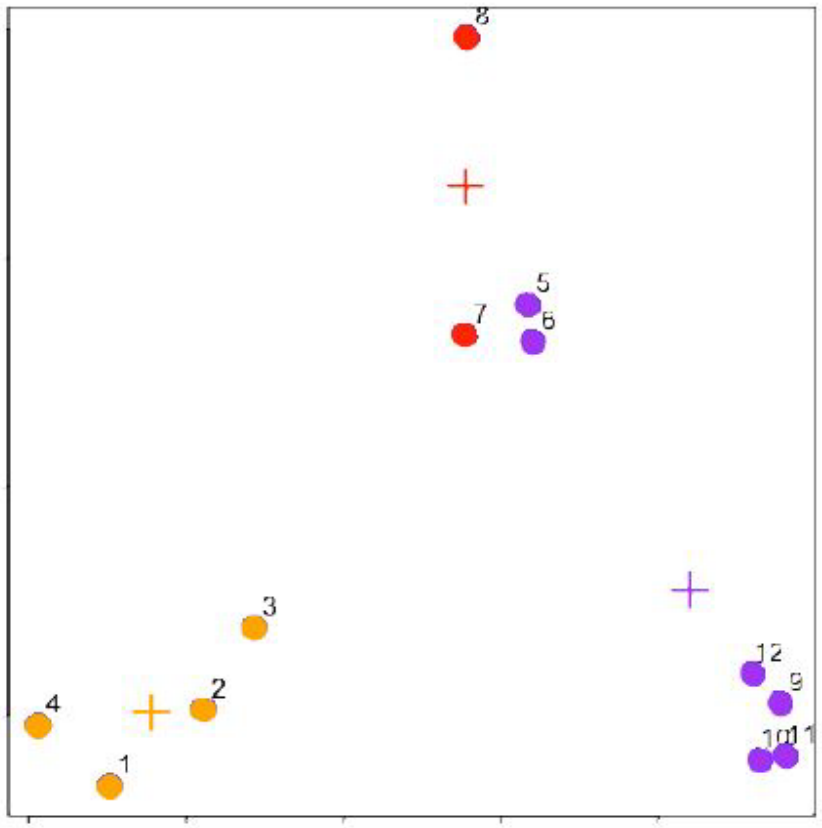
Unsupervised Machine Learning

- K-means clustering – reassign values



Unsupervised Machine Learning

- K-means clustering – update centroids



Code modules

WEEK 10-1 Code module – EDA on simulated data set

WEEK 10-2 Code module – hierarchical clustering

WEEK 10-3 Code module – data viz for hierarchical clusters

WEEK 10-4 Code module – K-means clustering

WEEK 10-5 Code module – data viz for K-means clusters

Summary

- In WEEK 10 of Introduction to Data Science, we continued the data science process by exploring two popular unsupervised machine learning algorithms.
- We used the `hclust()` algorithm for hierarchical clustering.
- We used the `kmeans()` algorithm for K-means clustering.