Unsupervised Machine Learning
CSCE 5215

K Means Clustering

K-means

- K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Motivation

In order to make sense of the world, sometimes we want to take a set of data samples and group them based on similarity.

This often simplifies learning, and can lead to insights about the nature of the data set being studied.

Examples:

- Direct Marketing: grouping people by their purchasing patterns
- Collaboration: identify groups based on citation patterns or paper content
- Genetics: noting which species are similar from an evolutionary point of view based on similarity between sequences

Advantage over dimensionality reduction

Dimensionality reduction (e.g. PCA)

- The components are linear combinations of features that transform to a response for that new feature, rather than a discrete grouping
- Requires a samples x features description

Clustering

- The clusters are discrete groups of points in state space
- Can be performed on the following data sets
  - samples x features matrix
  - similarity matrix of samples x samples

K-means

- Concept
  - Group points by similar distances
  - Groups are defined by the mean of the group for each feature
  - Membership is established by distance to a group means
- Algorithm
  - Initialize the group “centroids” by:
    - randomly choosing samples as the initial centroids
    - randomly assigning all samples to one of K classes
  - Re-assign membership of samples based on distance to the nearest centroid
  - Recalculate the centroids using the mean of the updated class memberships

No labeled data to train the model. Hence K-Means algorithm relies on the dynamics of the independent features to make inferences on unseen data.

(note distinction: K-means is clustering - an unsupervised learning technique. K-Nearest Neighbors is a supervised learning method)

Repeating the algorithm again…

- In the clustering problem, we are given a training set ${x^{(1)}, ... , x^{(m)}}$.
- We want to group the data into a few cohesive "clusters."
- Here, we are given feature vectors for each data point $x^{(i)} but no labels $y^{(i)}$

Our goal is to predict k centroids and a label $c^{(i)}$ for each datapoint.

K-means clustering algorithm

- Input: K, set of points $x_1 \ldots x_n$
- Place centroids $c_1 \ldots c_K$ at random locations
- Repeat until convergence:
  - for each point x_i:
    - find nearest centroid c_j $\quad \arg\min_j D(x_i, c_j)$
    - assign the point x_i to cluster j
  - for each cluster j = 1 ... K: $\quad c_j(a) = \frac{1}{n_{j x_i \to c_j}} \sum x_i(a) \quad$ for $a = 1..d$
    - new centroid c_j = mean of all points x_i assigned to cluster j in previous step
- Stop when none of the cluster assignments change

distance (e.g. Euclidian) between instance x_i and cluster center c_j

O (#iterations * #clusters * #instances * #dimensions)

The algorithm

1. The number of clusters along with the centroid value for each cluster is chosen randomly.

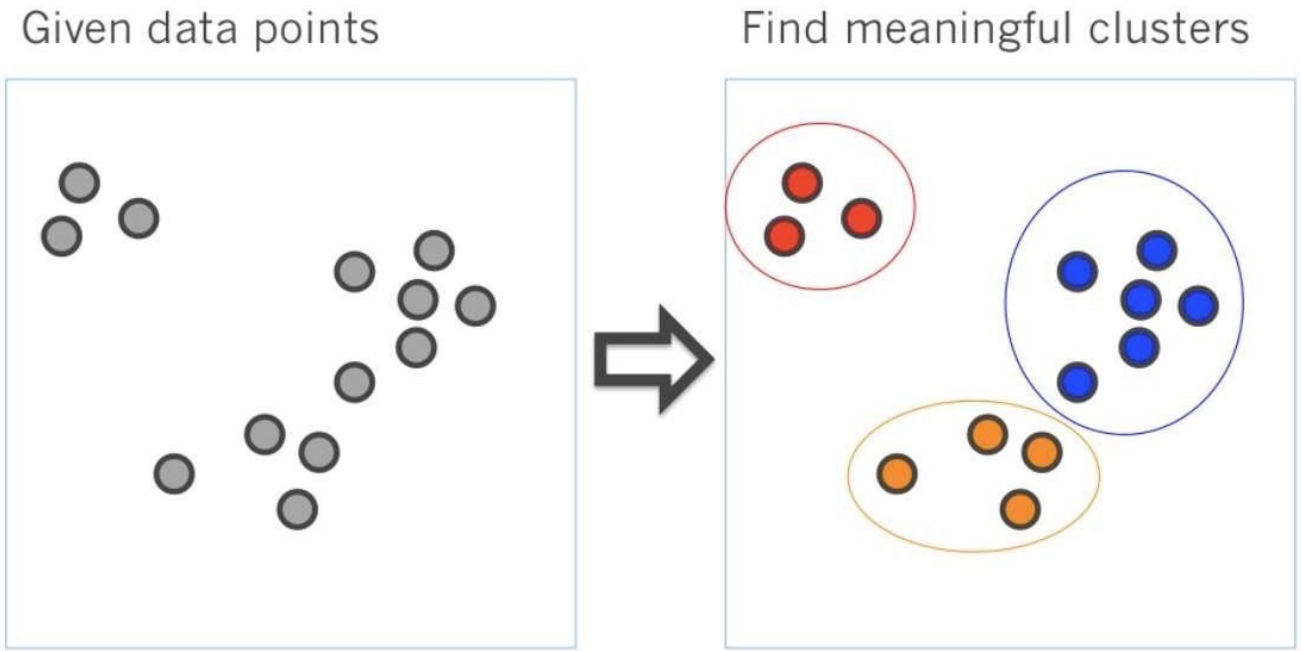
2. Euclidean distance between each data point and all the centroids is calculated.

3. The data points are assigned to the cluster whose centroid has the smallest distance to the data point.

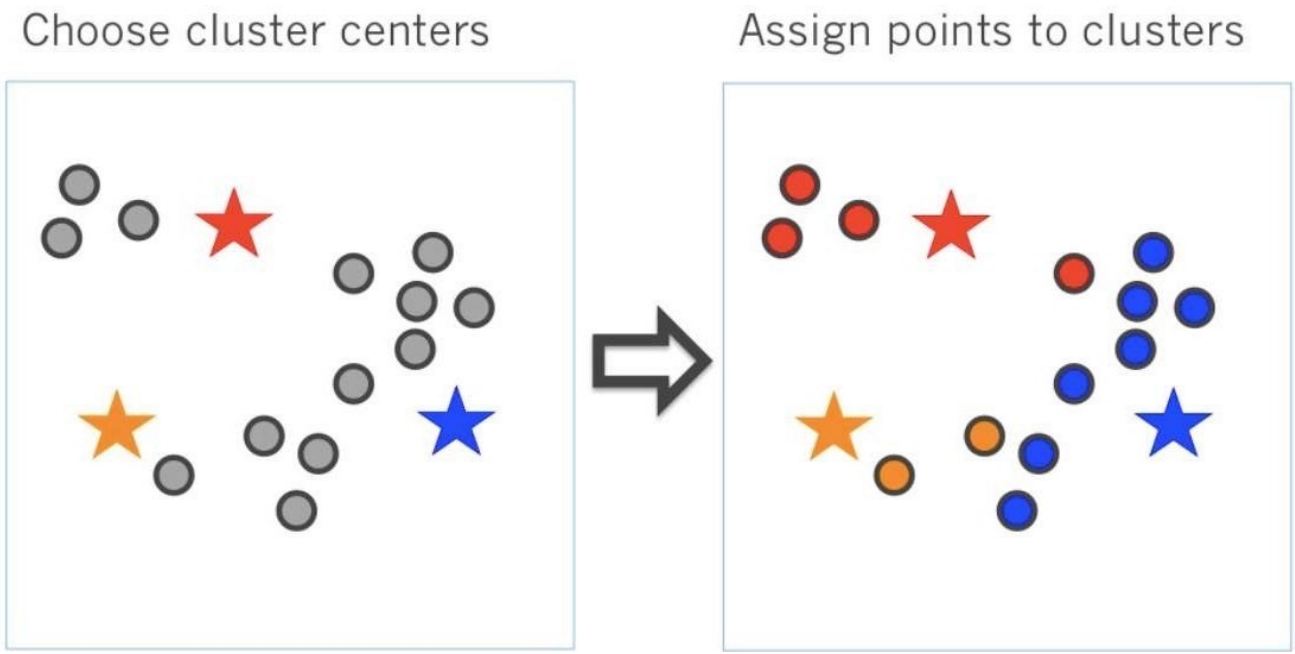
4. Centroid values for each cluster are updated by taking the mean of all the points in the cluster.

5. Steps 2, 3 and 4 are repeated until there is no difference between the previous centroid values and the updated centroid values for all the clusters.

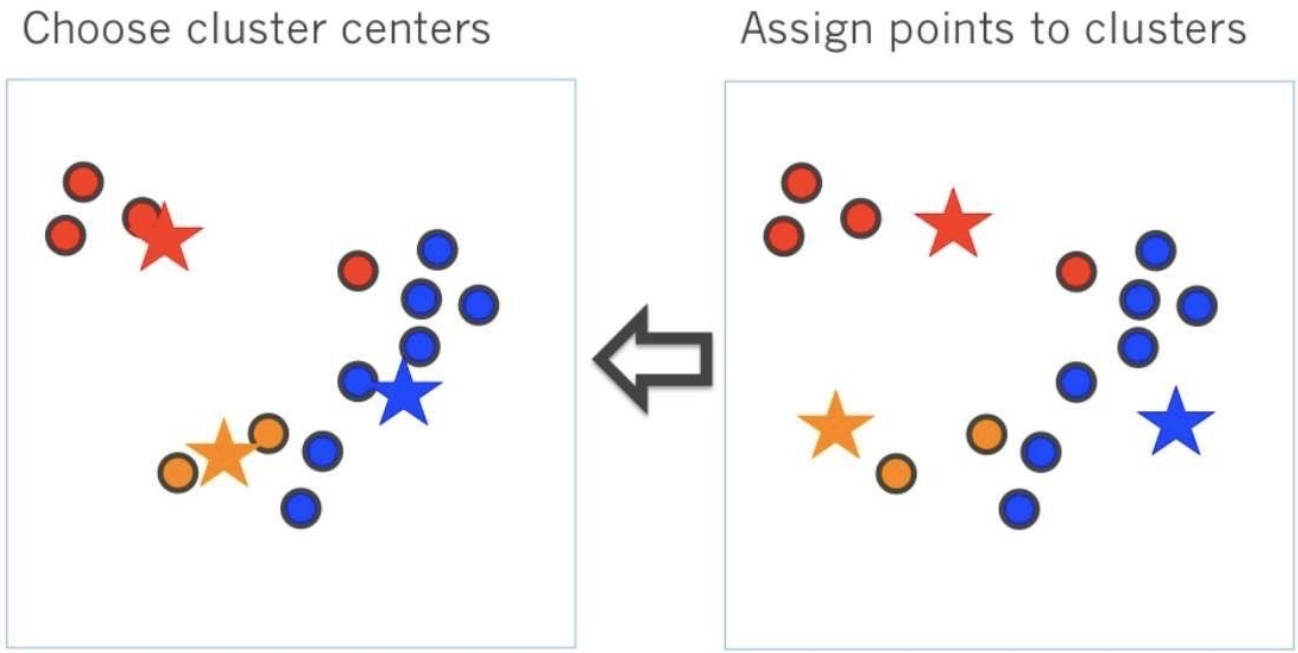
Clustering with K-Means



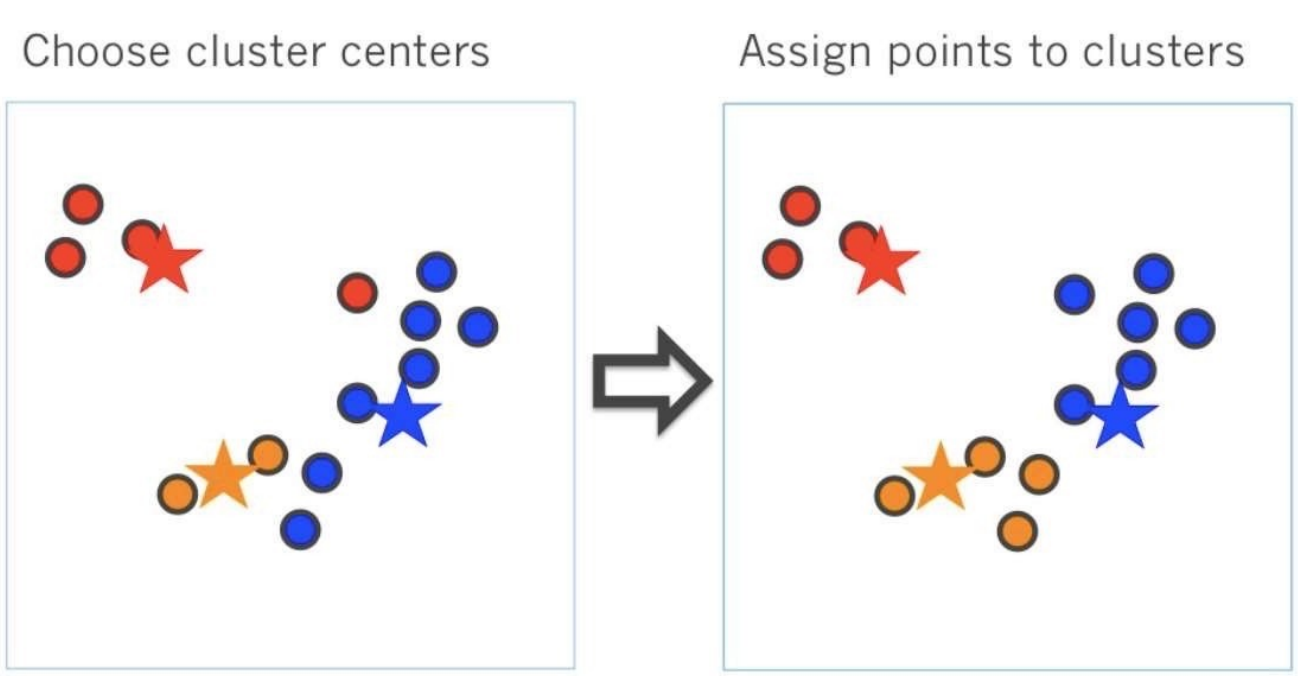
Clustering with K-Means



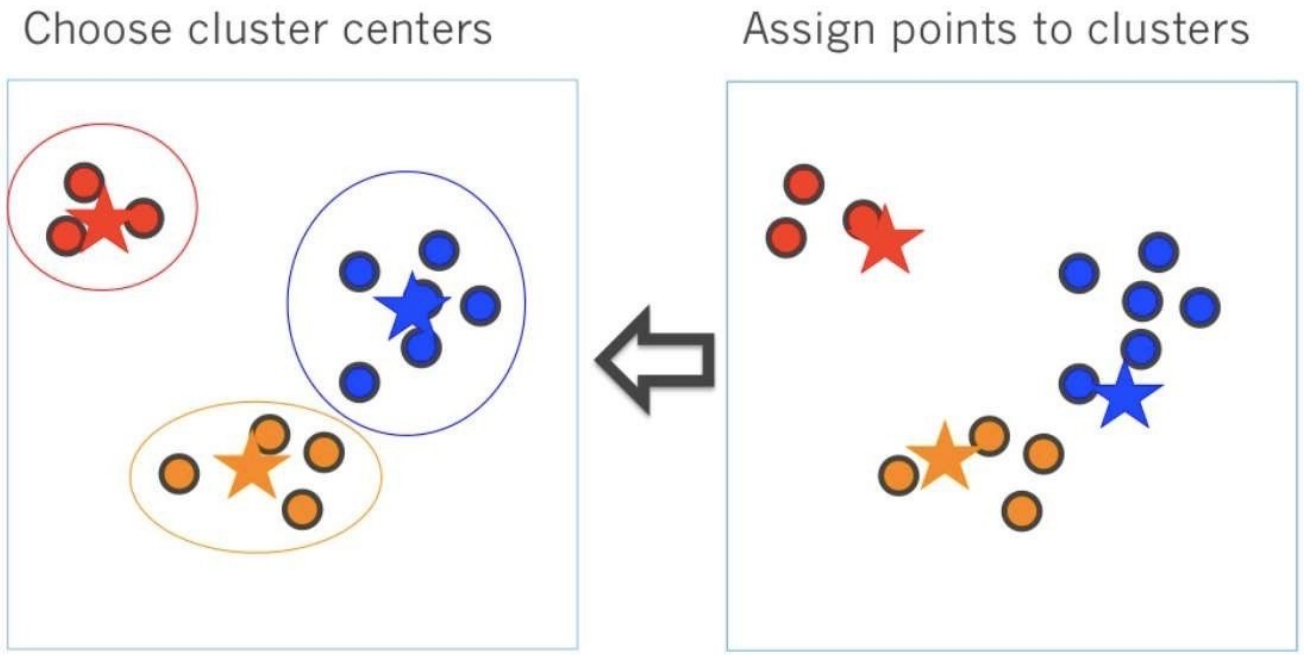
Clustering with K-Means



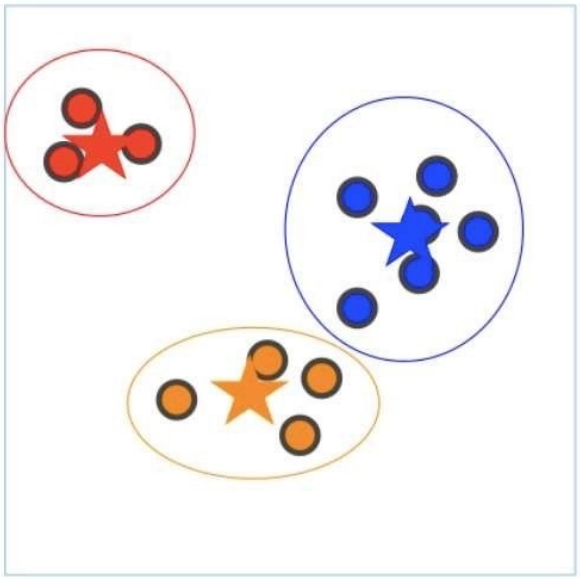
Clustering with K-Means



Clustering with K-Means



Clustering with K-Means



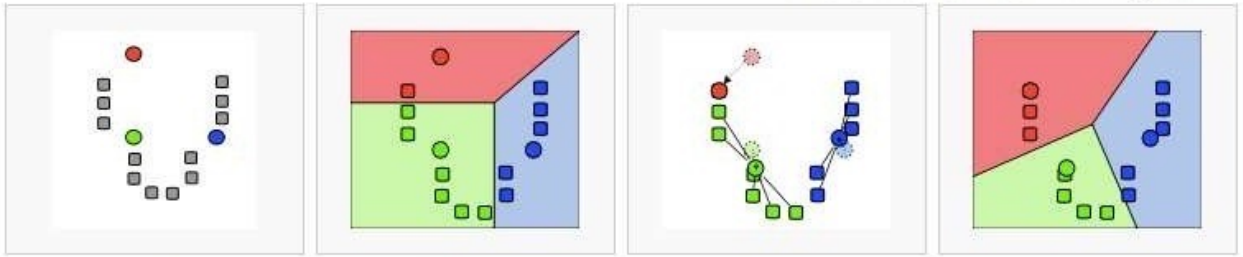
Data distributed by instance (point/row)

Smart initialization

Limited communication
(# clusters << # instances)

Summary

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

https://en.wikipedia.org/wiki/K-means_clustering

K-means comments

- The number of clusters found is fixed at K. Choosing the right K is important
  - there are methods to help select, but they are not bulletproof
- The resulting clusters can greatly depend on the initialization
  - local minima are often the result
- Only works for problems in which clusters are roughly spherical in shape (limitation)
  - otherwise, the centroid is not the best way to define membership