VGP337 - Neural Network & Machine Learning

Instructor: Peter Chan

Unsupervised Learning

Unsupervised Learning

- Unsupervised machine learning algorithms find patterns from a dataset without reference to known, or labeled, examples
- Unlike supervised learning, it cannot be directly applied to regression or classification problems because you have no idea what the values for the output data might be
- Instead, unsupervised learning can be used to discover the underlying structure of the data

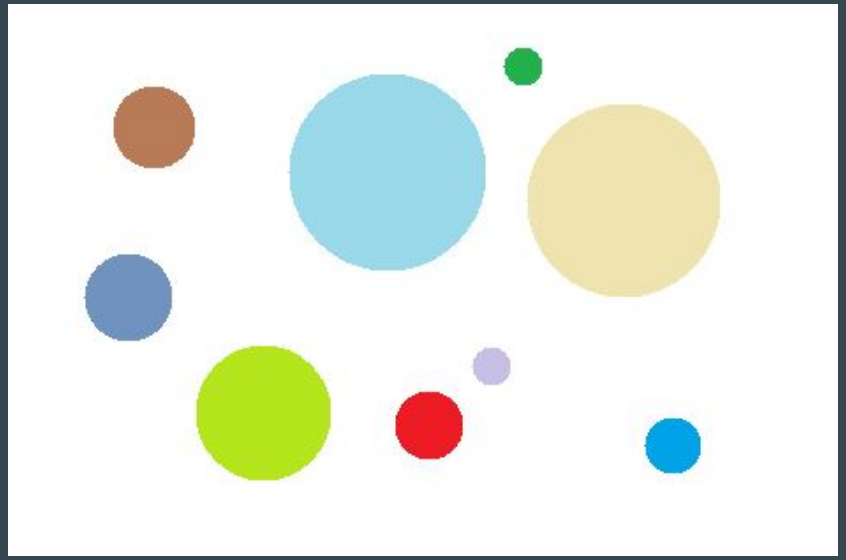
What is Clustering?

- Clustering is a set of techniques used to partition data into groups, or clusters
- Clusters are loosely defined as groups of data objects that are more similar to other objects in their cluster than the rest of the dataset
- In practice, clustering helps us better understand our data. This may include:
  - Recognizing key characteristics
  - Summarizing the dataset in a more compact representation
  - Identifying the presence of outliers

Clustering Approaches

- There are many different approaches to data clustering
- Each approach differ in their understanding of what constitutes a cluster
- Typical cluster models include:
  - Connectivity-based clustering (hierarchical clustering)
  - Centroid-based clustering
  - Distribution-based clustering
  - Density-based clustering
  - Grid-based clustering
- You can find more information about them here: https://en.wikipedia.org/wiki/Cluster_analysis#Algorithms

k-Means Clustering



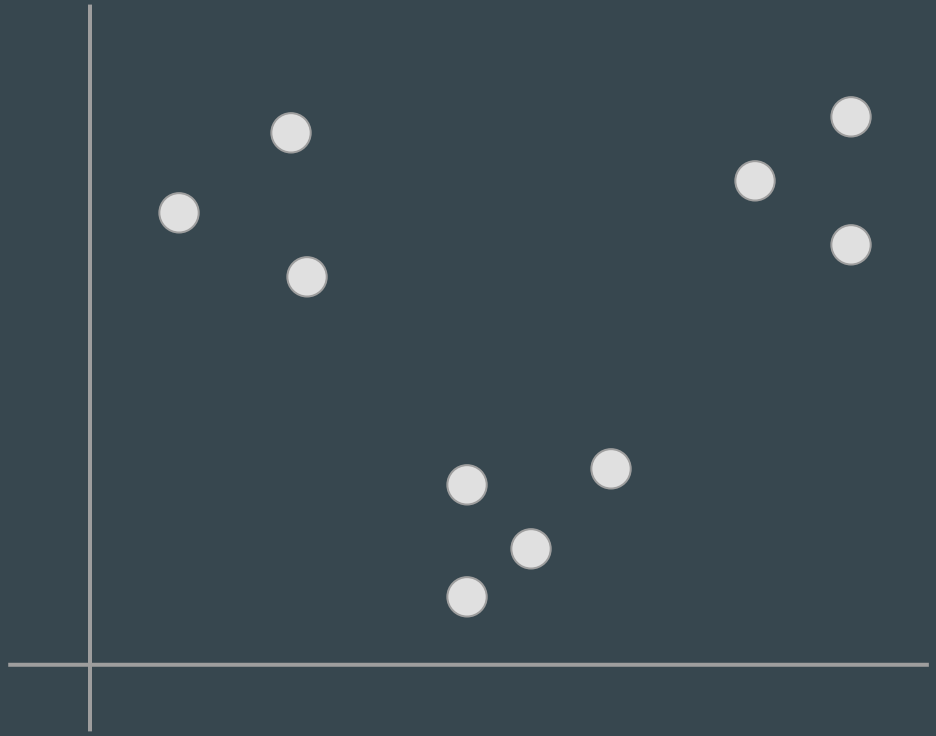
k-Means Clustering

- k-Means Clustering is an unsupervised learning algorithm that aims to partition a unlabeled dataset into k clusters
- It is a centroid-based technique in that the clusters are represented by a central vector, which may not be part of the data set
- The idea is simple: find k cluster centers and assign each example to the nearest center such that the total squared distances from the cluster are minimized

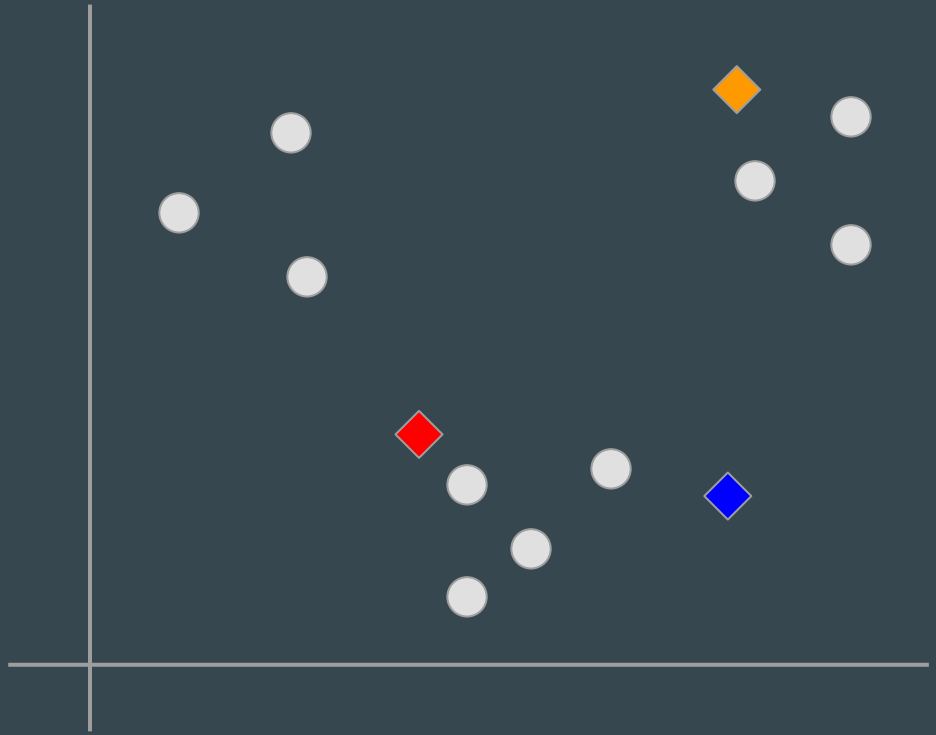
Algorithm

```
let k = number of clusters to assign
randomly initialize k centroids
while centroids changed && # of iteration < max iteration:
    for point in dataset:
        assign point to the closest centroid
    update each centroid to the new mean
    iteration++
```

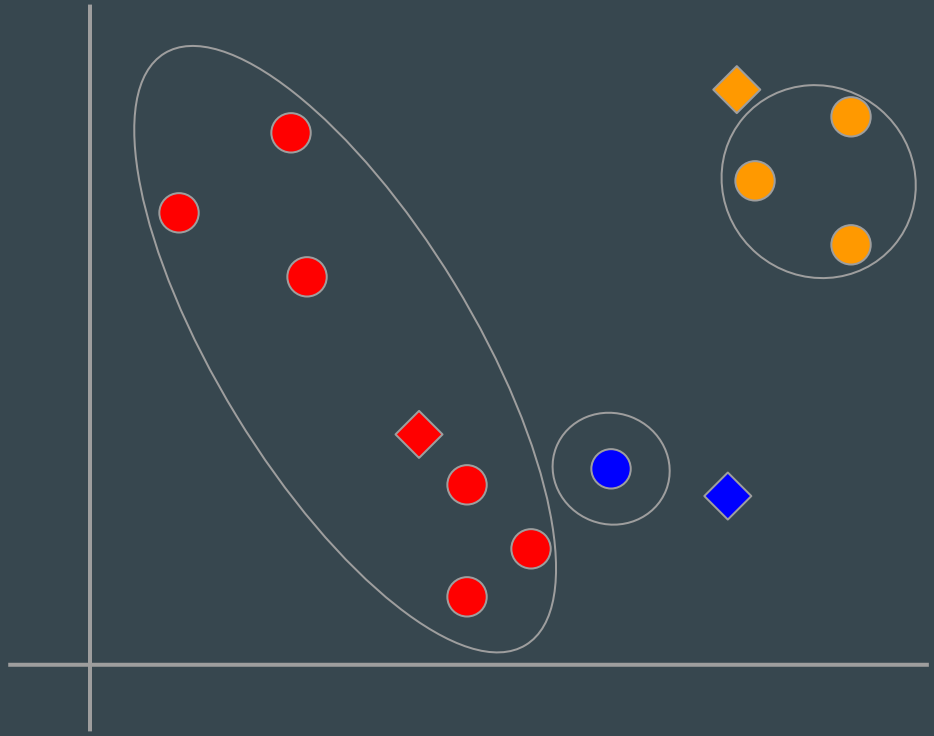
k-Means - Dataset



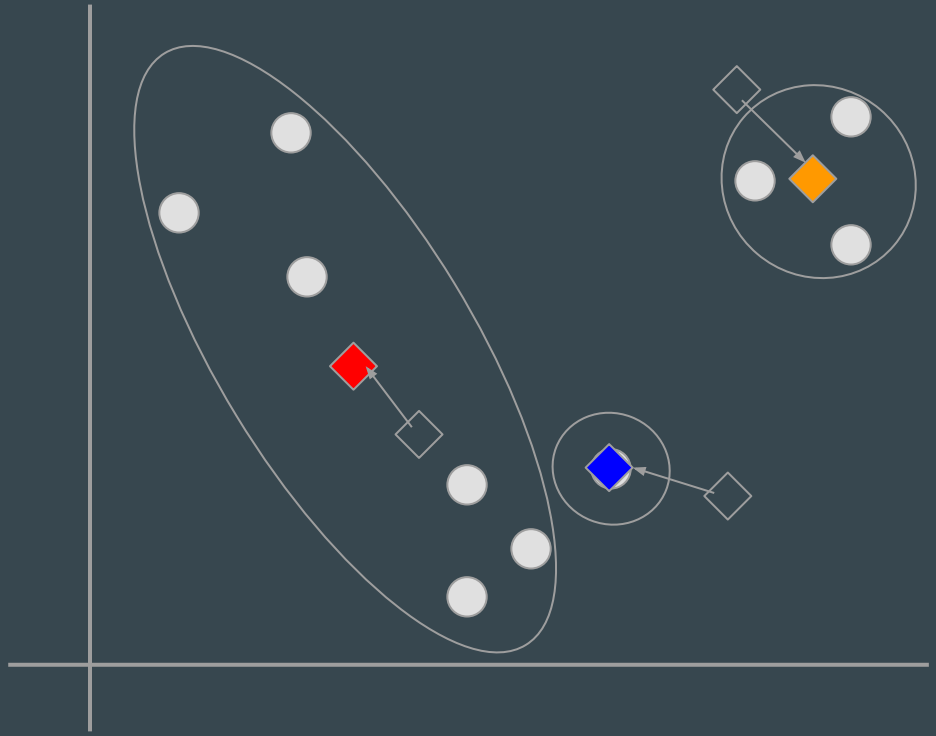
k-Means - Random select



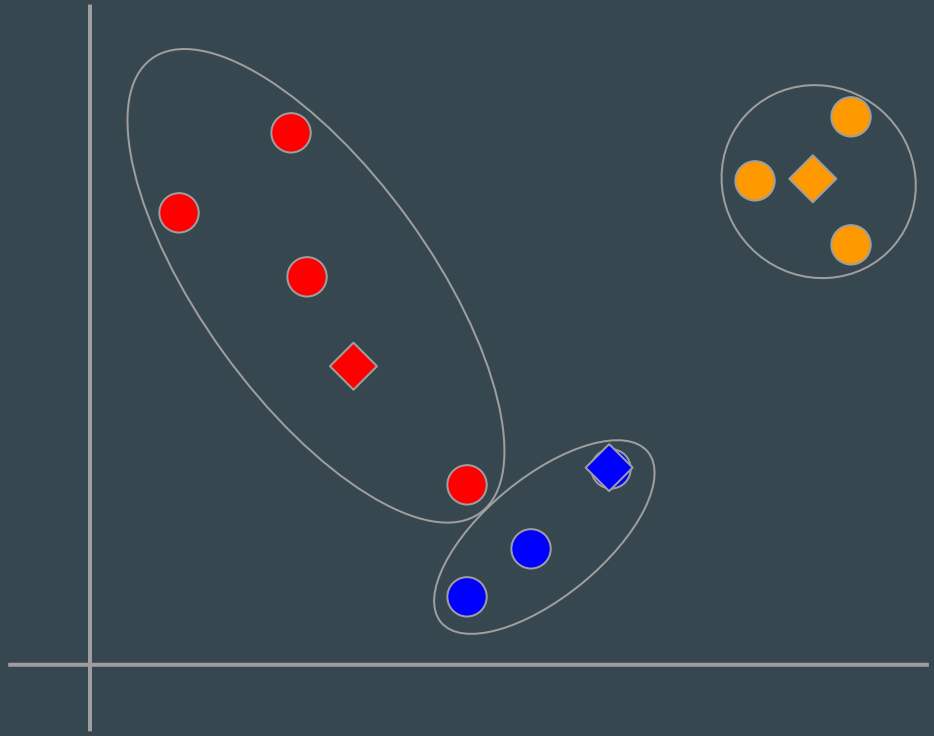
k-Means - Assign membership



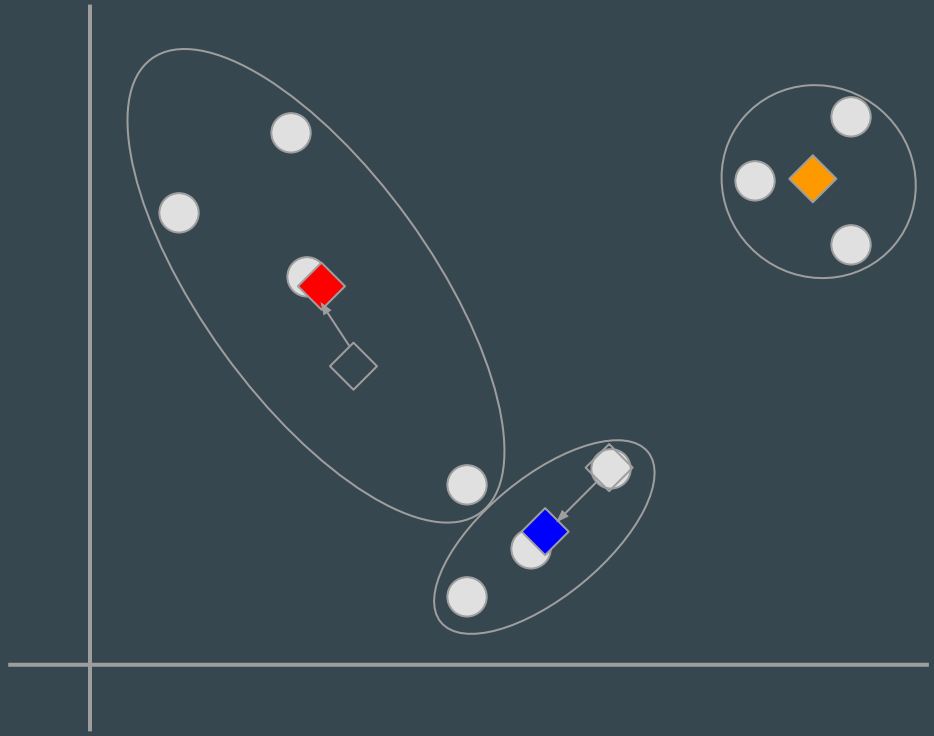
k-Means - Update centers



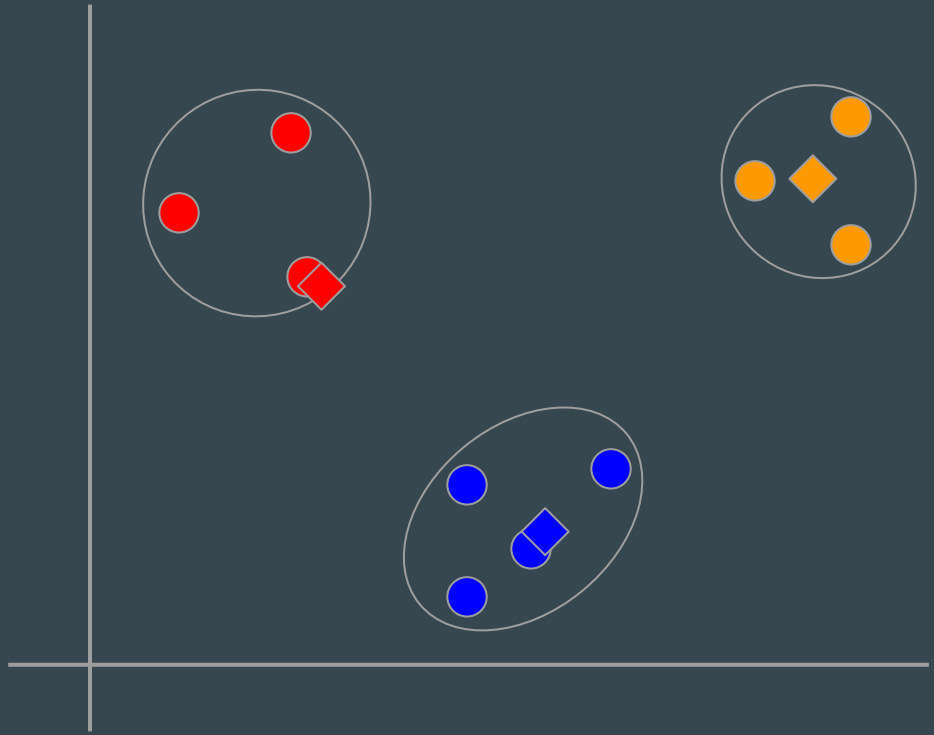
k-Means - Assign membership #2



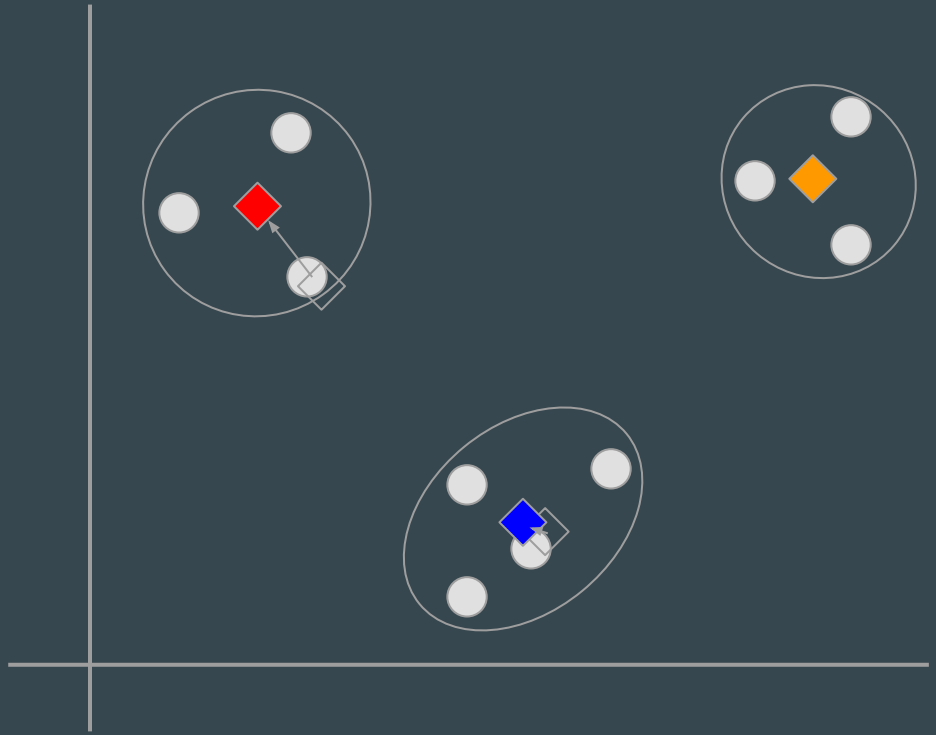
k-Means - Update centers #2



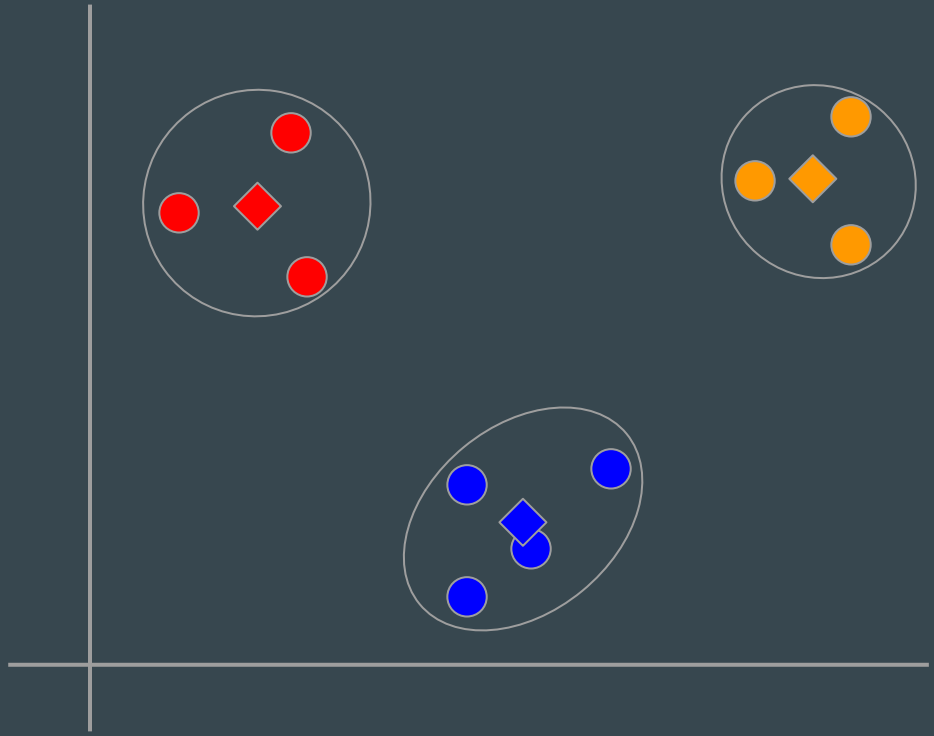
k-Means - Assign membership #3



k-Means - Update centers #3



k-Means - We are done!

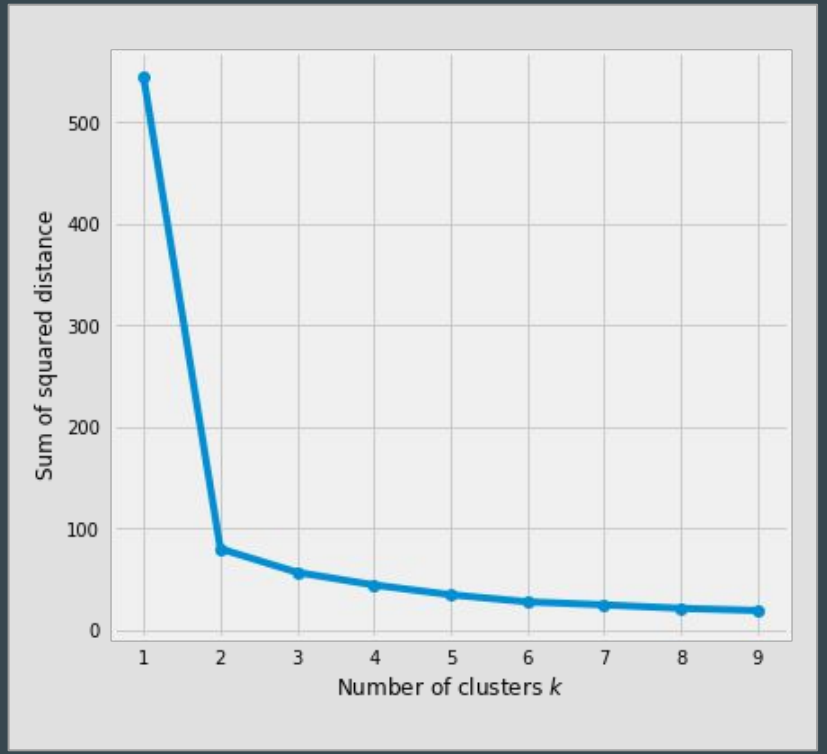


How to find k?

- Unlike supervised learning, clustering analysis does not have a ground truth to evaluate the accuracy of a model
- Furthermore, for k-Means, there is no right answer in terms of the number of clusters we should have for a particular dataset
- Nevertheless, there are some metrics that can help us decide what is a good *k*, namely:
  - The Elbow Method
  - Silhouette Analysis

The Elbow Method

- The Elbow Method works by computing the sum of squared error (SSE) between each data point and their assigned cluster centroid for each values of *k*
- This is easiest to see when plotted in a graph
- The *k* where the error starts to flatten out (the elbow) will be selected



Silhouette Analysis

- Silhouette analysis works by computing the degree of separation between clusters
- For each sample:
  - Compute the average distance from all data points in the same cluster (a^i)
  - Compute the average distance from all data points in the closest cluster (b^i)
  - Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

- This coefficient can be anything between -1 to 1
  - If the value is 0, then the sample is very close to the neighboring clusters
  - If the value is -1, then the sample is assigned to the wrong cluster
  - If the value is 1, then the sample is far away from the neighboring clusters

References

- Understanding K-means Clustering in Machine Learning
- K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks
- Selecting the number of clusters with silhouette analysis on KMeans clustering