Modul 3

Intro to K Means Clustering

Data Science Program

Purwadhika
Startup and Coding School

Reading Assignment

Chapter 10 of

Introduction to Statistical Learning

By Gareth James, et al.

Purwadhika
Startup and Coding School

K Means Clustering

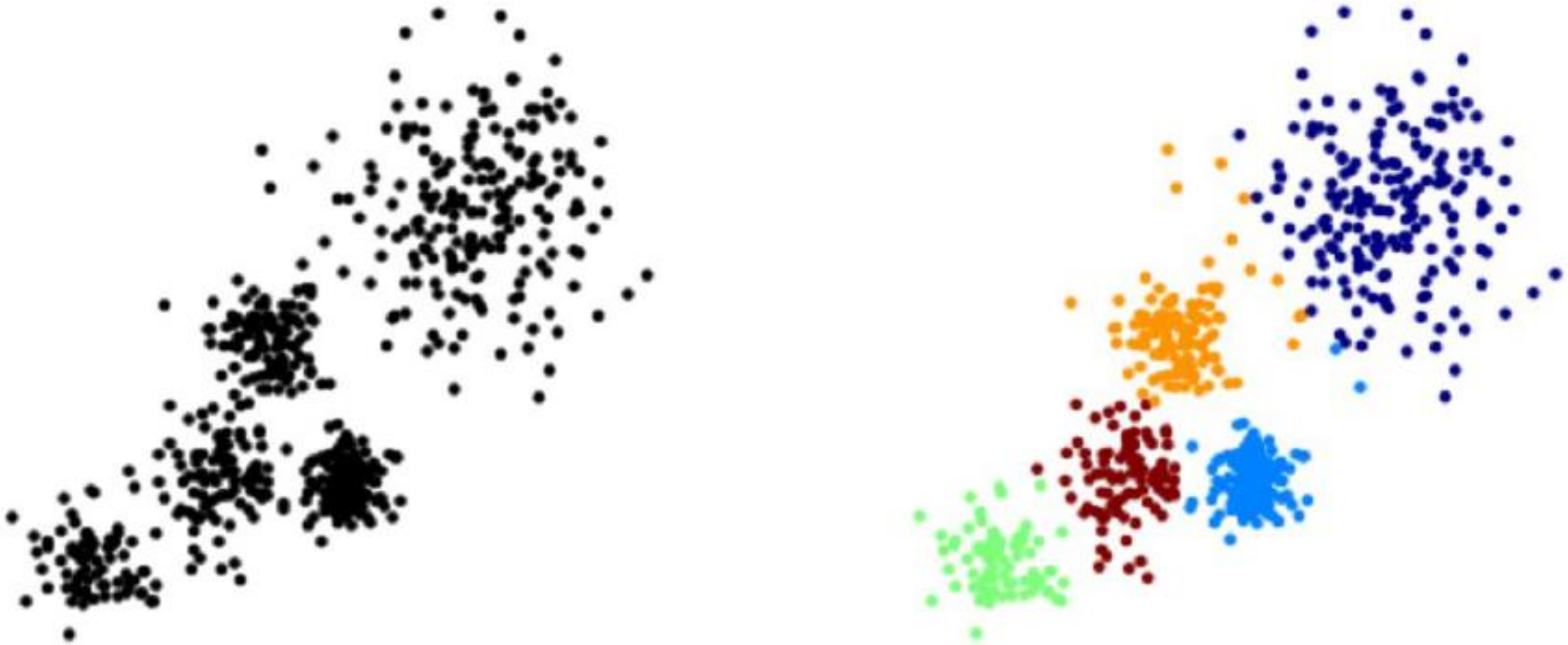
K Means Clustering is an unsupervised learning algorithm that will attempt to group similar clusters together in your data.

So what does a typical clustering problem look like?

- Cluster Similar Documents
- Cluster Customers based on Features
- Market Segmentation
- Identify similar physical groups

K Means Clustering

The overall goal is to divide data into distinct groups such that observations within each group are similar.

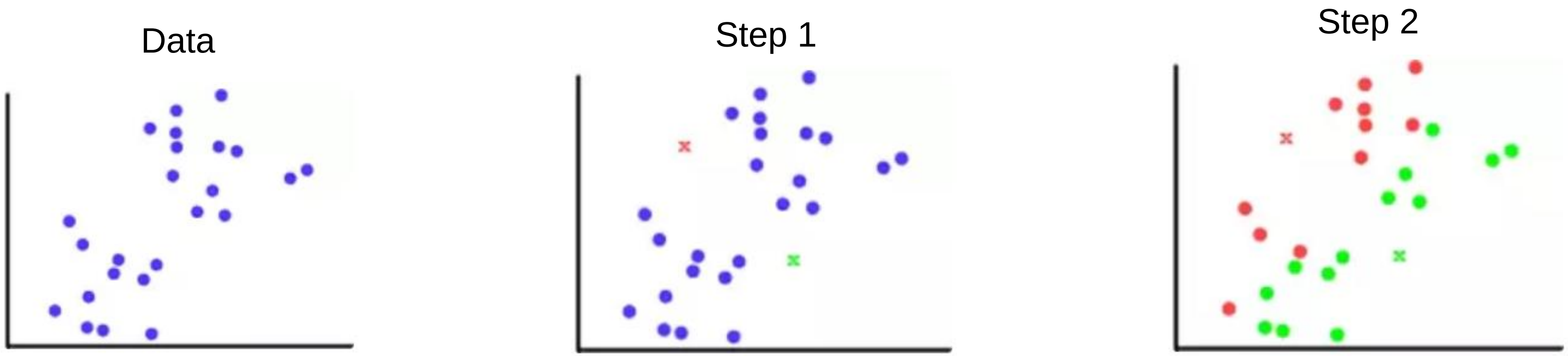


K Means Clustering

The K Means Algorithm

- Choose a number of Clusters “K”
- Randomly assign each point to a cluster
- Until clusters stop changing, repeat the following:
  - For each cluster, compute the cluster centroid by taking the mean vector of points in the cluster
  - Assign each data point to the cluster for which the centroid is the closest

K Means Clustering



Data

Step 1

Step 2

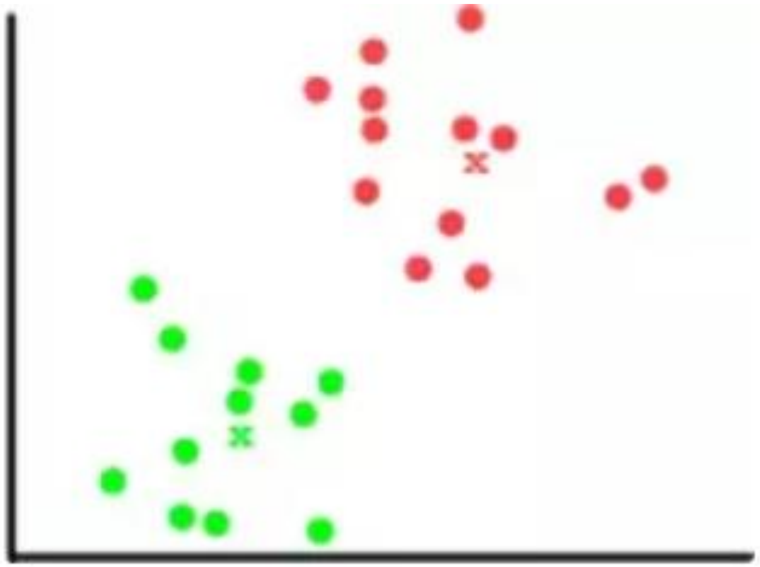
- The first thing we can do is to randomly initialize two points, called the cluster centroids.
- In k-means we do two things. First is a cluster assignment step and second is a move centroid step.
- In the first step, algorithm goes to each of the data points and divide the points into respective classes, depending on whether it is closer to the red cluster centroid or green cluster centroid.

Purwadhika
Startup and Coding School

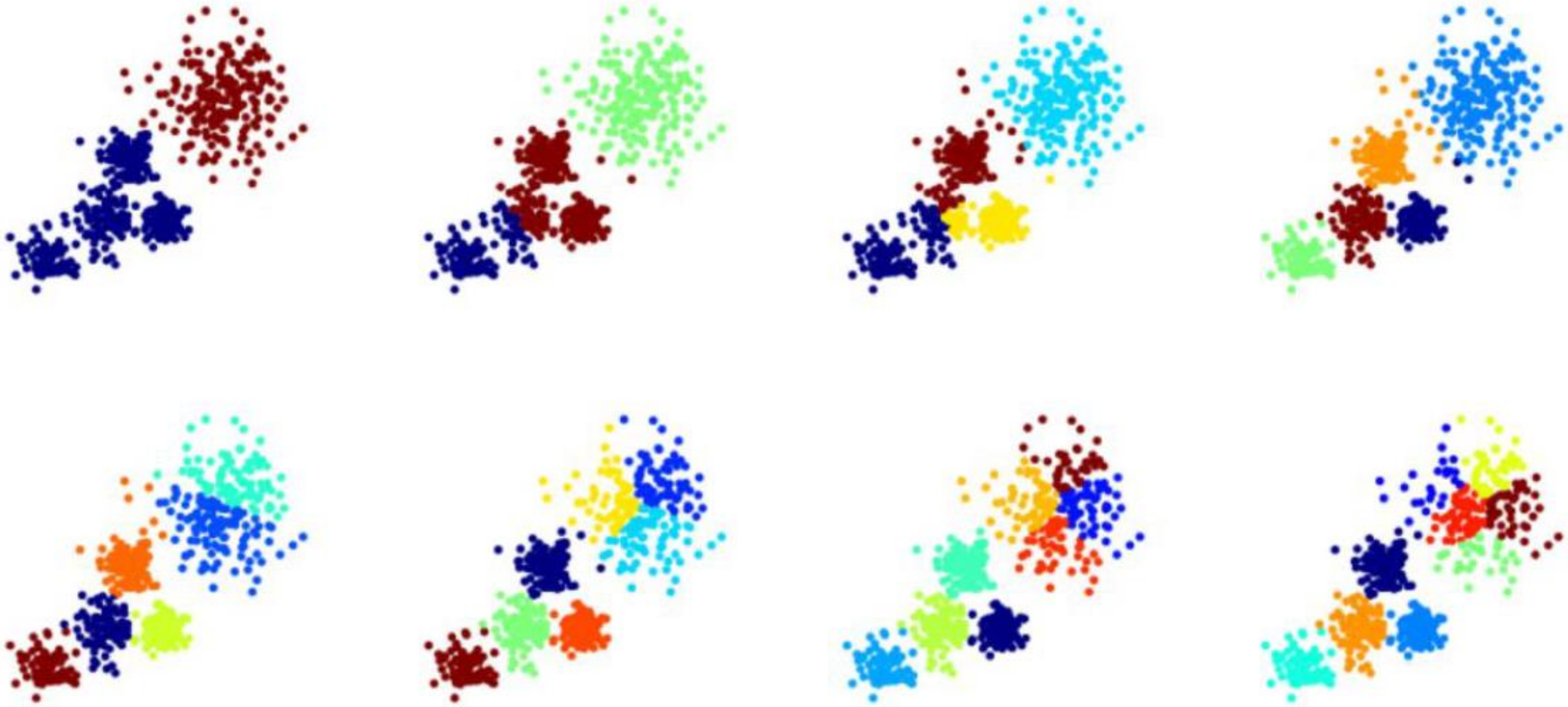
K Means Clustering

- In the second step, we move the centroid step. We compute the mean of all the red points and move the red cluster centroid there. We do the same thing for green cluster.
- This is an iterative step so we do the above step till the cluster centroid will not move any further and the colors of the point will not change any further.

Last Step



Choosing a K Value



Lets Practice with Python!