**Clustering Analysis**

Clustering is an unsupervised technique which divides the data points into several groups such that data points in the same groups are more like other data points in the same group than those in other groups.

**Types of Clustering:**

**Hard Clustering:** Each data point either belongs to a cluster completely or not.

**Soft Clustering:** Instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

Two most common methods are K-Means and Hierarchy clustering methods.

**K-Means Clustering**

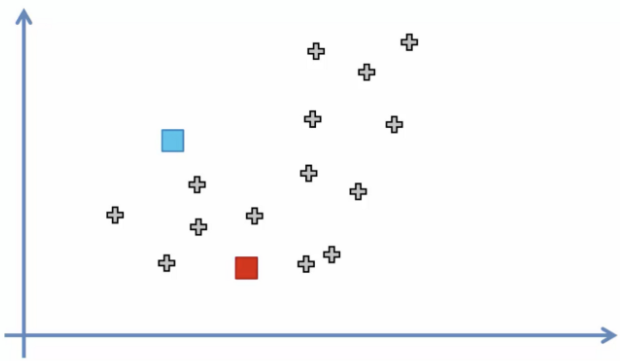
K means is an iterative clustering algorithm that aims to find local maxima in each iteration. In contrast to traditional supervised techniques, K-Means attempts to classify data without having first been trained with labeled data. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the most relevant group.

Application of K-Means:

* customer profiling
* market segmentation
* computer vision
* search engines
* astronomy

How it works:

Step 1: Select K (i.e. 2) random points as cluster centers called centroids

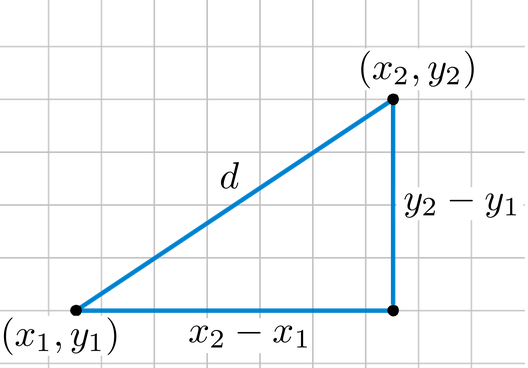


Step 2: Assign each data point to the closest cluster by calculating its distance with respect to each centroid.

**Distance metric**

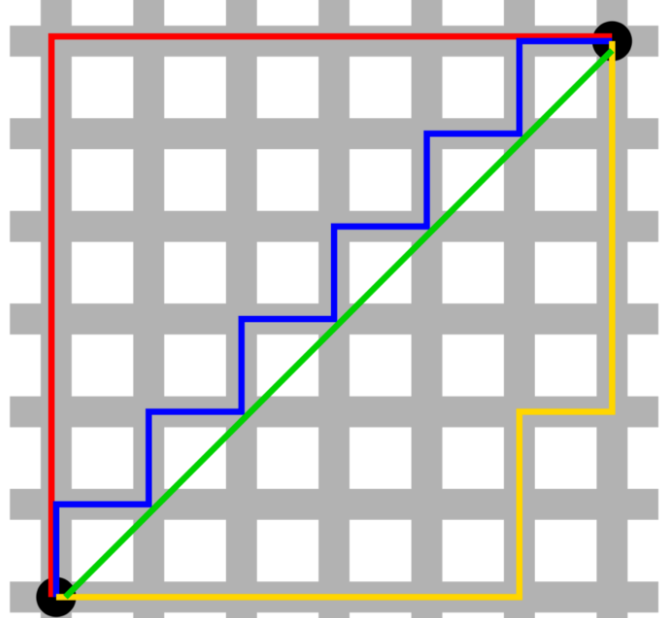
**Euclidean distance**:

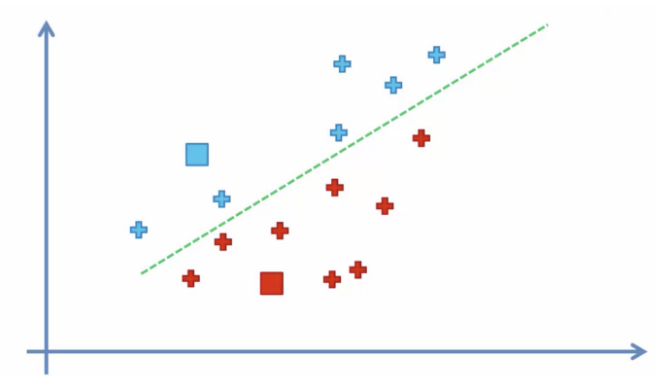
The shortest distance between two points. For example, if x=(a,b) and y=(c,d), the Euclidean distance between x and y is √(a−c)²+(b−d)².



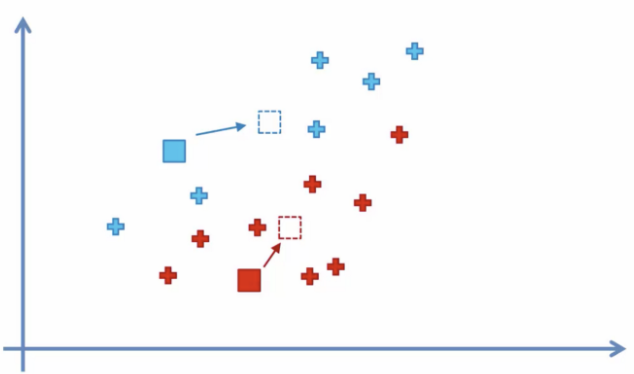
**Manhattan distance:**

Imagine you were in the downtown center of a big city and you wanted to get from point A to point B. You wouldn’t be able to cut across buildings, rather you’d have to make your way by walking along the various streets. For example, if x=(a,b) and y=(c,d), the Manhattan distance between x and y is |a−c|+|b−d|.

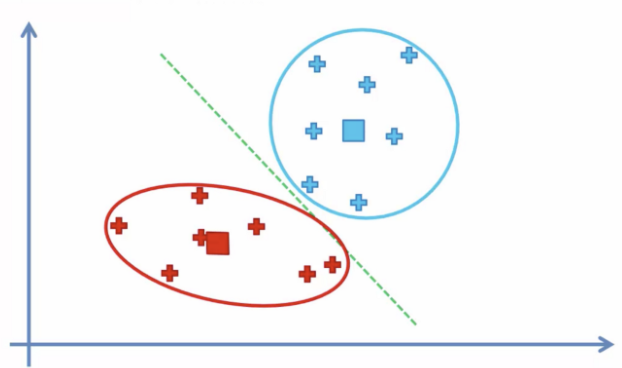




Step3: Determine the new cluster center by computing the average of the assigned points.

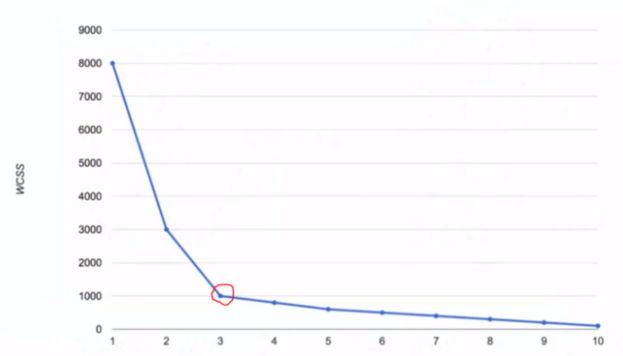


Step 4: Repeat steps 2 and 3 until none of the cluster assignments change.

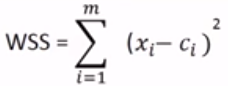


**How to choose no of clusters?**

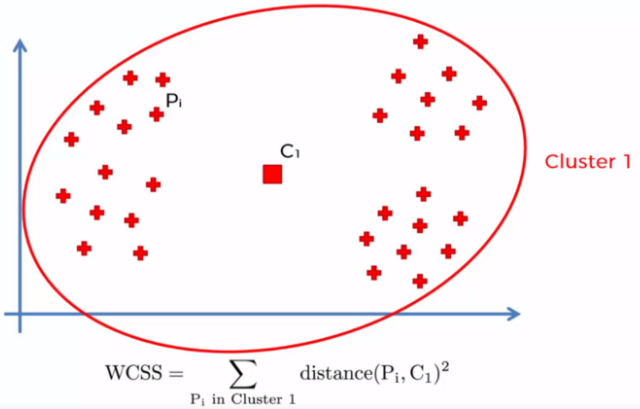
We graph the relationship between the number of clusters and Within Cluster Sum of Squares (WCSS) then we select the number of clusters where the change in WCSS begins to level off (elbow method).



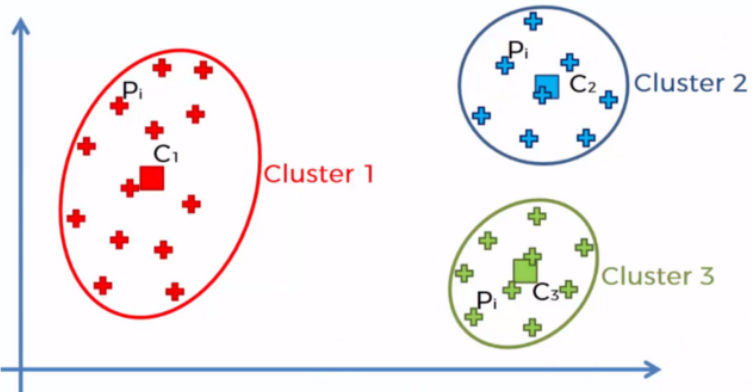
WCSS is defined as the sum of the squared distance between each member of the cluster and its centroid.



WCSS for one cluster



WCSS for three clusters





**Advantages:**

* Relatively simple to implement.
* Scales to large data sets.
* Guarantees convergence.
* Can warm-start the positions of centroids.
* Easily adapts to new examples.
* Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

**Disadvantages**:

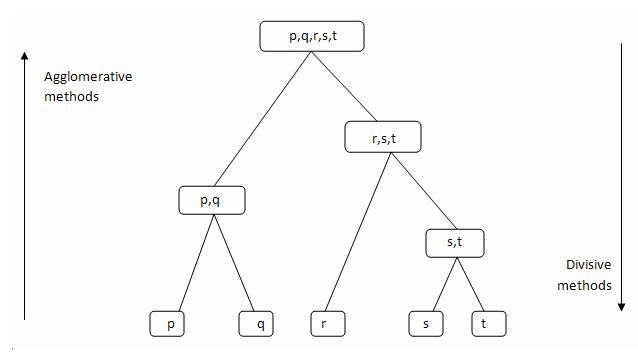
* Choosing K manually.
* Being dependent on initial values.
* Clustering data of varying sizes and density.
* Clustering outliers.
* Scaling with number of dimensions.

**Hierarchical clustering**

There are two types of hierarchical clustering algorithms:

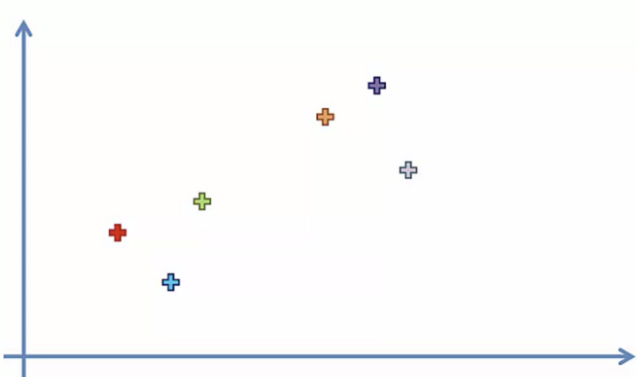
**Agglomerative** — Bottom up approach. Start with many small clusters and merge them together to create bigger clusters.

**Divisive** — Top down approach. Start with a single cluster than break it up into smaller clusters.

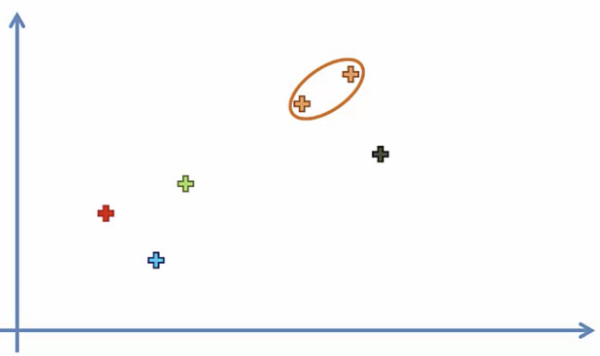


How it works?

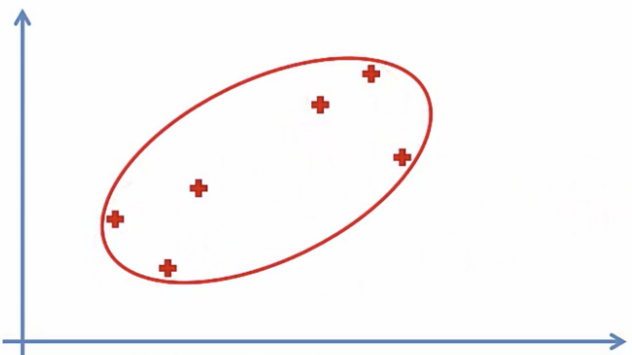
Step 1: Make each data point a cluster



Step 2: Take the two closest clusters and make them one cluster.

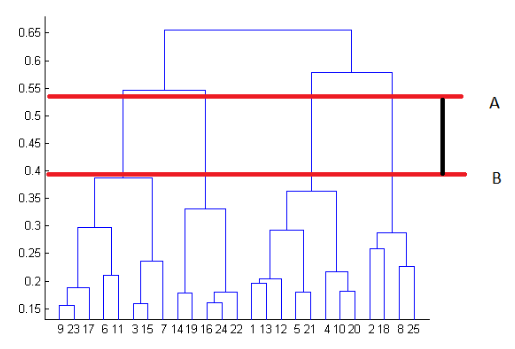


Step 3: Repeat step 2 until there is only one cluster



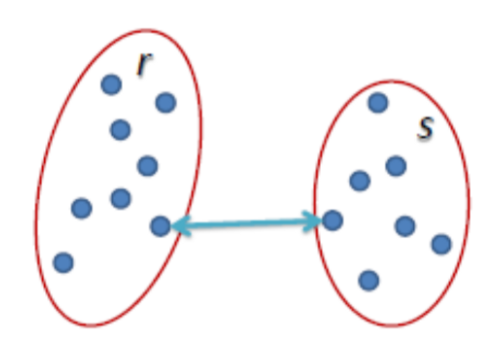
**Dendograms:**

We can use a dendrogram to visualize the history of groupings and figure out the optimal number of clusters. Determine the largest vertical distance that doesn’t intersect any of the other clusters. Draw a horizontal line at both extremities. The optimal number of clusters is equal to the number of vertical lines going through the horizontal line.

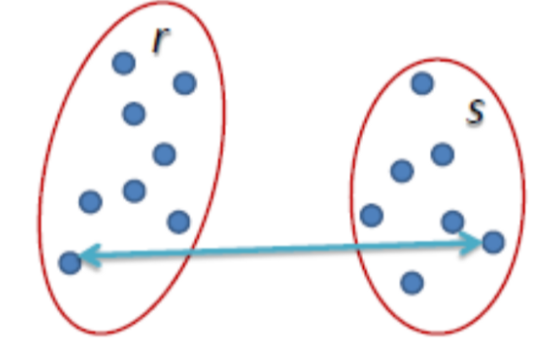


**Linking criteria:**

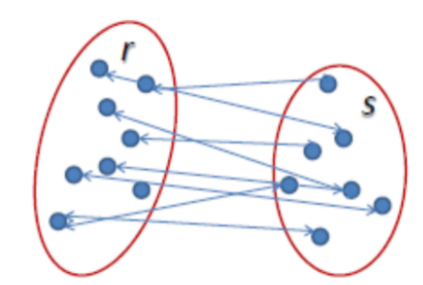
**Single Linkage:** The distance between two clusters is the shortest distance between two points in each cluster.



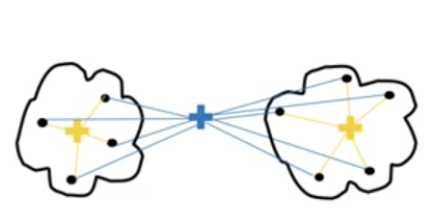
**Complete Linkage**: The distance between two clusters is the longest distance between two points in each cluster.



**Average Linkage:** The distance between clusters is the average distance between each point in one cluster to every point in other cluster**.**



**Ward Linkage**: The distance between clusters is the sum of squared differences within all clusters.



**Pros**

* No assumption on number of clusters (i.e. k-means)
* May correspond to meaningful taxonomies

**Cons**

* Once a decision is made to combine two clusters, it can’t be undone
* Too slow for large data sets, O(𝑛2 log(𝑛))