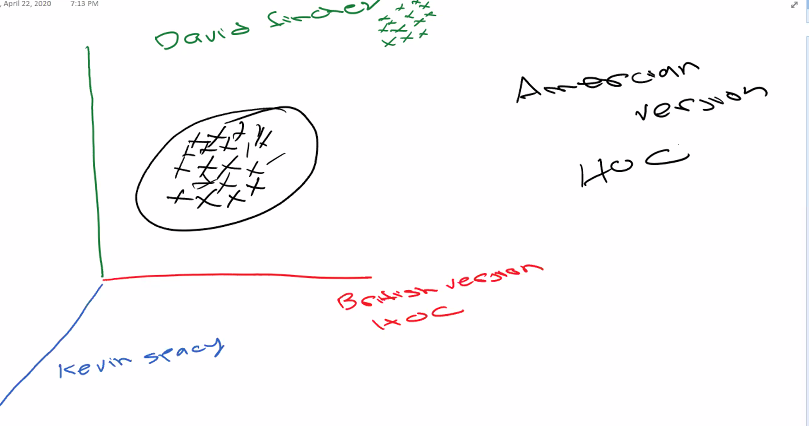
In unsupervised learning we have no idea what is the target label.

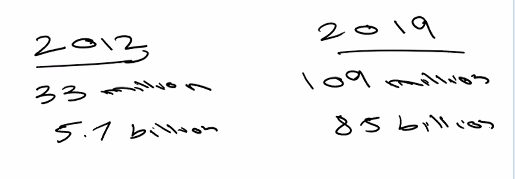
In clustering we divide into segments.

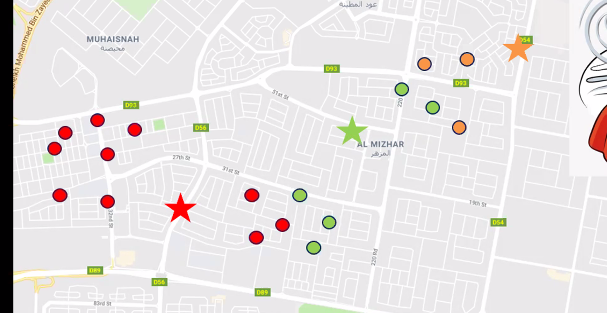
Shoaib likes sci fi and zubair likes action.

Netflix recommends sci fi movies to shoaib and action to zubair.

We find set of homogenous groups in clustering.



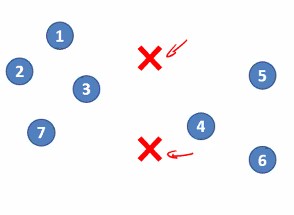


Newyorker pizza 

Green delivers to green red to red orange to orange.

We need these stars in the middle. Stars are the delivery centers.

K meam algorithm draws two centroids.



Centroids always drop on random locations.

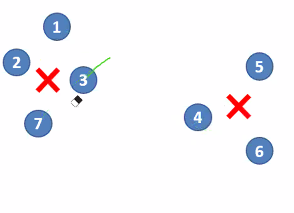
Every data point will pick closest centroid.

Second step optimize

Centroids will move so that quadratic distance will be minimum

Center 1 will go to A

Center 2 will go to B



Process of reassignment and reoptimizing will continue until point will not move.

No of centroids is equal to no of clusters.

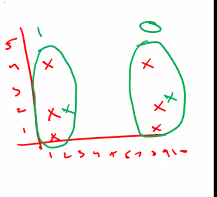
Cost function is the difference of every point from centroid.

Cost function or error will be more for if we use 2 instead of 3.

N-init is no of initialization. Centroid drop can change the cluster.k mean repeats the cluster drops 10 times.The one the comes in majority is considered.

Max iteration repeats reassignment and optimizing 300 times.It wont complete 300 if cost function aligns before It will exit before 300.

11 1 00 0 are the cluster names. Array shows that first three points belong to cluster 1 and last three points belong to cluster 0.



Center are the location of centroids.

Our objective would be to have min no of clusters and min cost function. Trade off between clusters and cost function.

If there is no clear elbow we use another method.

Silhouette score within cluster centroid distance less but for clusters distance should be more.

Within cluster datapoints should be close and should be far apart from other clusters.

We will use both elbow method and Silhouette score.

Performance metric is the silhouette score.

If we are using distance algorithm we need to scale the data always.(min max scalar)

We cant use assignment when using inplace=True..

Labels are predictions. Array of clusters.

If we don’t normalize the big ones will dominate and some clusters can be made wrong.

When we view the data we do not use the scaled we use actual we only do scaling for algorithm.

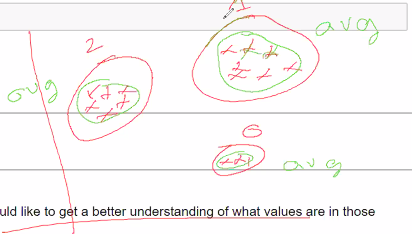
E=historical data T=Kmeans. P=silhouette score

Cost function=distance of centroid from all data points.

.inertia\_ shows cost function.

[0] contains k and [1] contains cost function

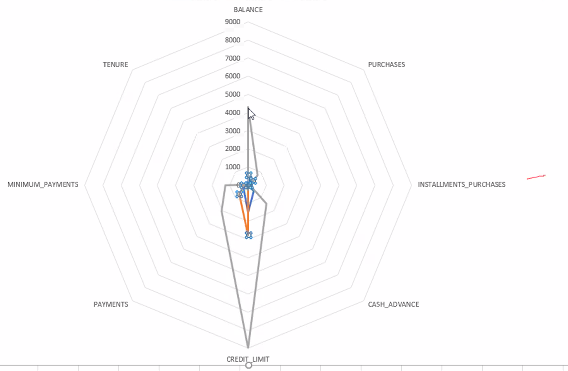
We start from two since 1 has no value of b.

Now we have to find the basis for clusters and homogeneity.

We are calculating avg to better understand what is in the these clusters.Since the same type of people are in these clusters.

Sales spent more than budget and marketing spent less etc. radar plot

You can make the radar plot in excel.



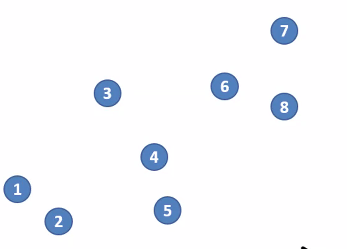
There is balance two times because to complete the circle you have to end where you start.

1-uniform dataset 2-2 crescent data set.

K mean cannot separate the data properly.

K mean can only for no 4 and 6 not for the rest. K mean is best for 6

We can see how the clusters get segmented in dendogram.



Every data point considered cluster

It will make one big cluster

Then it will create cluster of data points that are near to each other.

Sub and parent clusters will be made.

Single link when we join two clusters we will consider nearest point.

The process will continue until only one cluster is left.one big cluster.

We view and cut according to the requirement.

We will consider the farthest point from the nearest cluster.

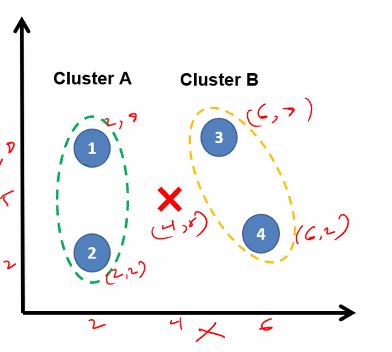
Then we will consider the two lines. 1 4 lines and 3 5 lines. We consider which line is shortest.

Disadvantage if data points are close and we are considering farthest point ideally yellow should have gone to brown but it still goes to green that is limitation.

In average we consider every point distance than take average.then we compare two numbers (averages with each other) smallest num selected

Ward method best.

First off we decided middle point.



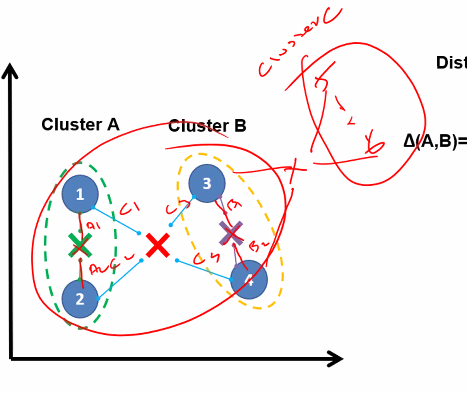
6+2/2=4

2+8=10/2=5 so (4,5) is center.

Center will be calculated on the basis of all the four points. All x will be added and all y will be added than averaged.

We calculate the distance of every point from center.

We minus the variance between the clusters.



The value we get at the end is compared and than the cluster B will join the best cluster in this case A.

Green and purple are the variance in the clusters.

Hierarchical clustering usage.

Blacks had same behavior to democrats so they join democrats.

It is also used in evolution trees.

We use adjusted rand score mostly.

Blue behavior was similar to red.they were democrats but they were siding with republicans.

We are splitting using grey line for two clusters 600 value we can use any value according to the no of clusters that you want. The colors show how many clusters will be made.evey cluster will have different color.

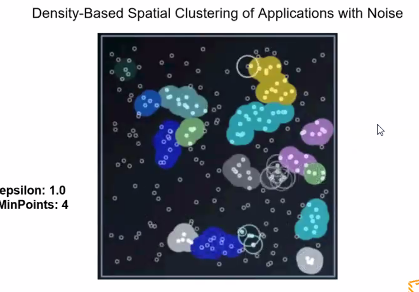
Y is the ground truth it is not training.

We are deciding which party is supporting Which bill.

Pink=support and black=not support republicans and democrats have opposite behavior.

Very pink means no vote cast. White is independent.

Carper is the wrong one



DBSCAN decides how many clusters should be in data.

Dense data points are created into clusters. The points that are not part of clusters are noise they are not dense.

It starts one by one.

Episolon is radius and diameter.

No neighbor on first point so noise.

Min no of points for cluster should be 5 so 5 neighbours only 1 so second point also noise.

Similarly third is also noise

Four is also noise. 5th point has all requirements so cluster is made.

Limitation if too many points it may draw a huge cluster.

-1 noise in this case.

Image segmentation and object recognition

3 blobs same score ,distributed best k mean,Best DBSCAN

Sillouette score less since a value is more so it has less sillouete score.

So we need new evaluation metric DBCV we use DBCV for DBSCAN.

Density within cluster should be high but it should low between clusters for DBCV.

Similar shape and dense are blobs

Data frame to values conversion.

Epsilon value should be controlled not more not less.

Epsilon=1.2 min\_samples=4( 3 clusters)

Episolon=1 min\_samples=15(2 clusters)

Dataset\_make has many type of shapes.

Colors are being decided by dbscan and kmeans