Clustering

GOALS of CLUSTERING

- The main goal of clustering is to sort data into groups
- Certain methods can remove background noise or train algorithms on how to sort data
- Help to identify characteristics of the different groups

DEFINITIONS

Unsupervised: the are no defined groups at the beginning sorting is done on just the data

Supervised: There is input knowledge about the clusters of the points, like from a training set

Centroid: the average or center point of a set of points within a cluster

Hard clustering: each data point is assigned to a single cluster, belongs only to this cluster

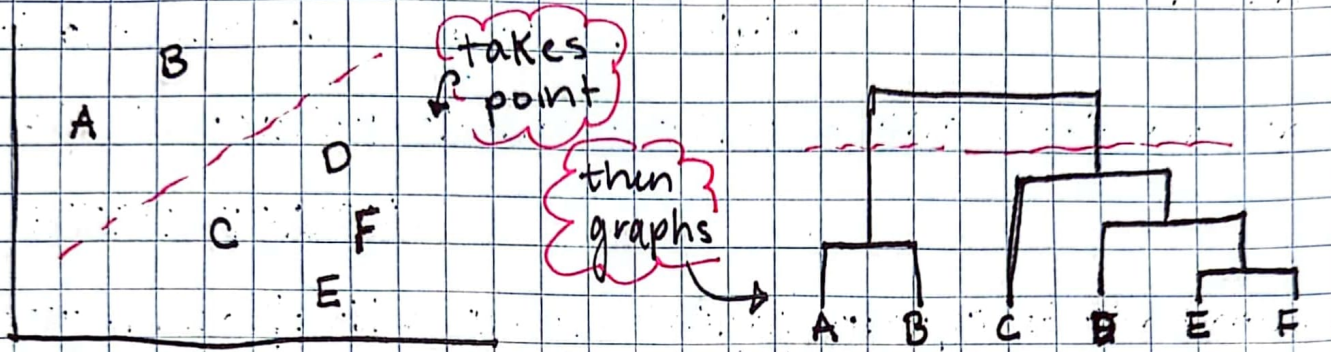
Soft Cluster: points may belong to multiple clusters w/ different probabilities

METHODS

TREE/DENDROGRAM

Procedure:

- each point starts in its own cluster
- clusters with the smallest distance are merged w/ the location being a the center
- clusters are merged until its 1 cluster



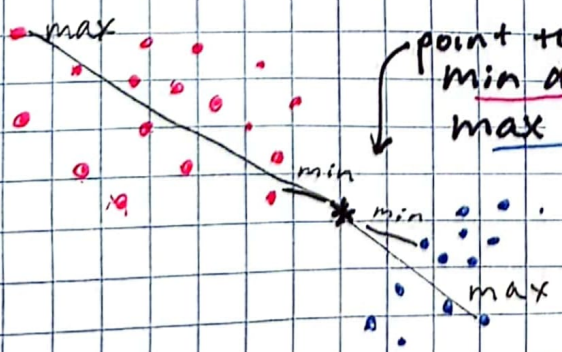
You can decide at what point the groups should stop merging.

There are different ways to define distance:

Single linkage: smallest distance between two points in the different clusters is used

Complete linkage: To sort a point the smallest maximum distance is used

Average linkage: uses the average distance of all the points in the cluster. Kind of a middle ground between single & complete.

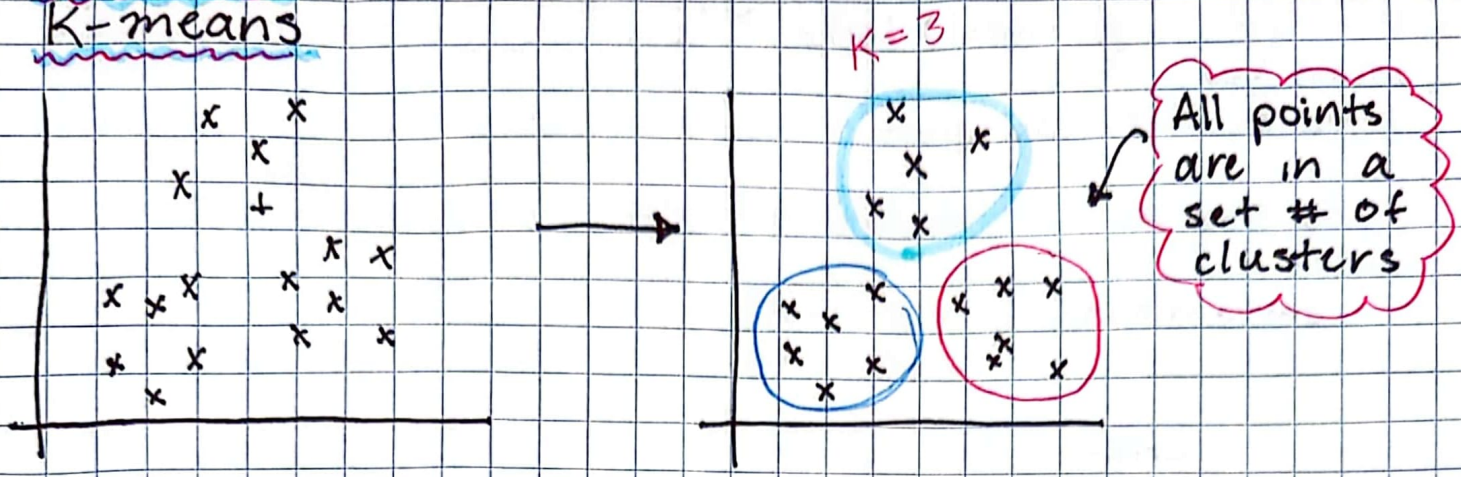


point to be sorted, pink has smallest min distance, but blue has smallest max distance

→ single

→ complete

K-means



K-means uses centroids to sort data into a pre-determined "K" number of groups.

Start by choosing the centroids or seed locations (K of them)

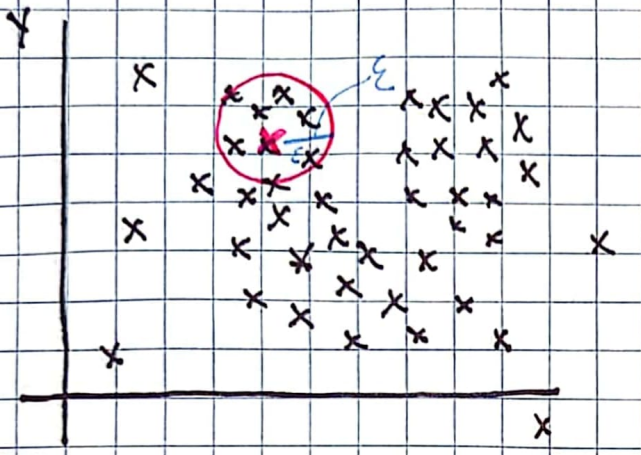
- Assign each point to the closest centroid
- Calculate the new centroids based on the average of the points in the cluster
- Reassign points based on new centroids
- Repeat until convergence

So you pick 3 Things:
- How many groups (K)
- The initial seed locations
- The definition of distance

Note: This method can yield different results based on the seed location. Not always correct.

DBSCAN (Density-Based Spatial Clustering of Applications w/ Norse)

- Useful for sorting clusters which may be hard to sort with other methods
- Not all of the points have to be put into a group



There are 2 parameters to define:

- ϵ : vector distance
- μ : # of data points

DBSCAN starts by seeing which points have μ points w/ in ϵ

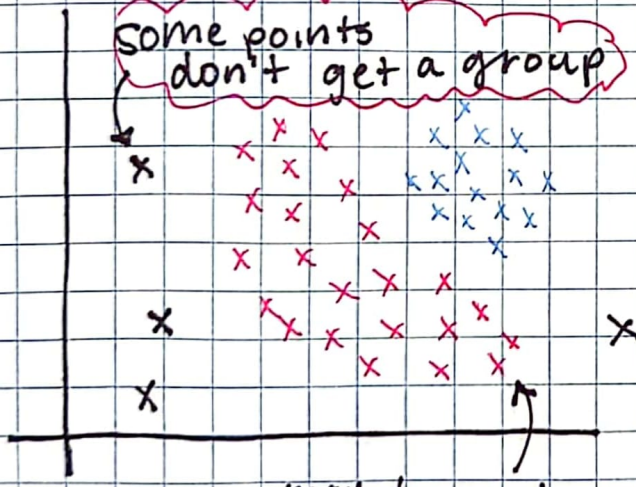
↓

Each point that meets that criteria is a core point

One core point is assigned a cluster & core points w/ in ϵ of the cluster are added to the cluster.

Non-core points w/ in ϵ of a core point to the cluster are added

Groups are created so each core point has a cluster (no core points are left out)



may have to vary μ & ϵ to get optimal results

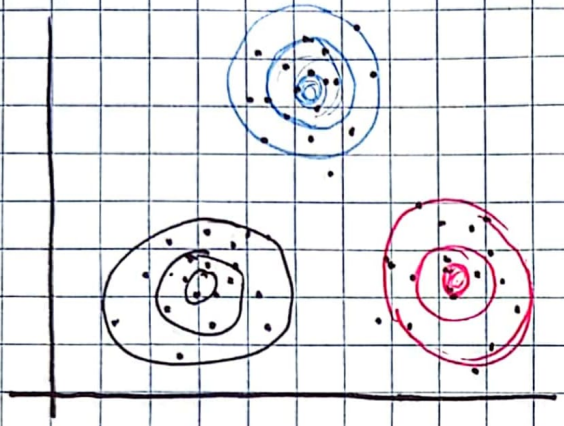
Other methods that are similar to DBSCAN:

- PRIM
- OPTICS
- BIRCH
- DENCLUE
- CHAMELEON

Gaussian Mixture Model

Instead of assigning points to a certain cluster, GMM gives the probability that a point is in a cluster

A gaussian is fit over each of the clusters.



This method relies on the EM algorithm

First an estimate of k & the cluster locations is given

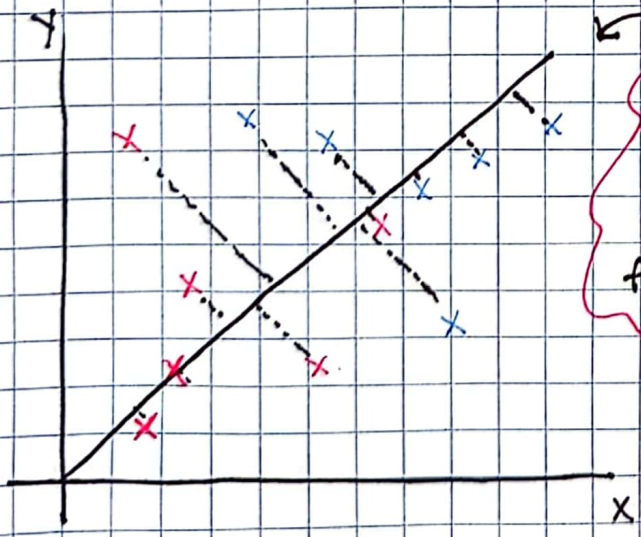
E: likelihood of each object belonging to each cluster is calculated

M: cluster parameters are optimized

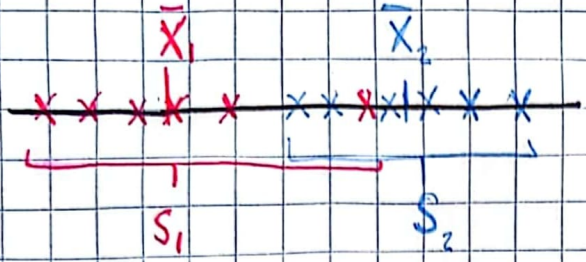
This continues until convergence

LDA (Linear Discriminant Analysis)

LDA tries to maximize the "separability" of the data It kinda acts similar to PCA, but instead of explaining variance it wants to explain the separation.



wants a new axis that best seperates the data so you use info from both x & y



$$a = \frac{\bar{X}_2 - \bar{X}_1}{S_1 + S_2}$$

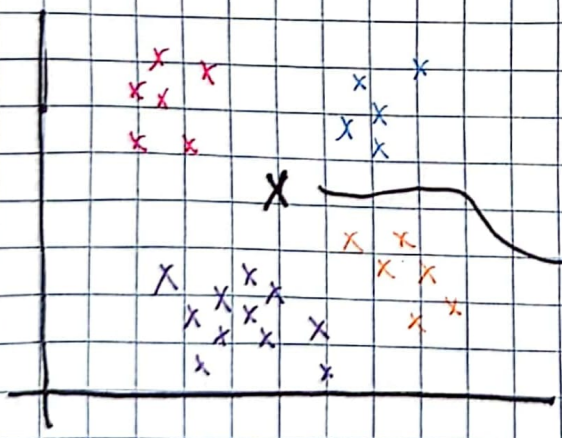
↳ maximum separation occurs for:

$$Sep = \frac{a^2(\bar{X}_2 - \bar{X}_1)^2}{a'(S_1 + S_2)a}$$

The axis is usually decided w/ training data and then used on the real data. This can also be done on higher dimensions.

For another method try: QDA

K-nn



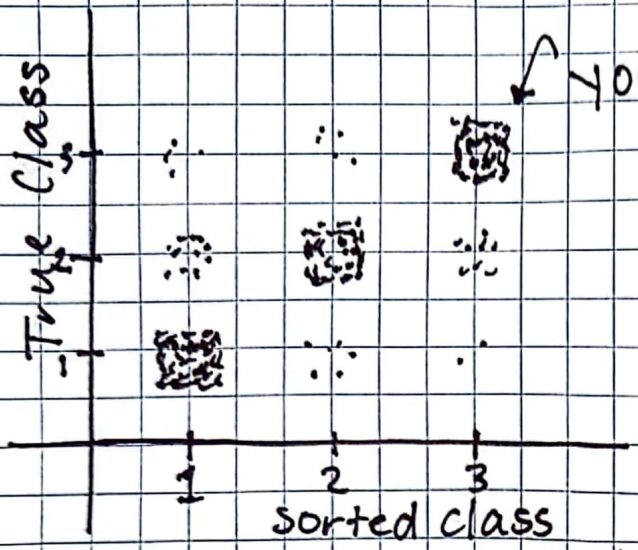
This method is used to assign points to already known groups

if we want to sort this point to a group we look at K nearest neighbor. For instance, if K=5, we look at the 5 closest points and see which cluster most of those points came from

The cluster w/ the most nearest neighbors is the cluster the point is assigned to

This method is often used in machine learning.

Also a note on the misclassification graphs



You hope that most of the data is on the diagnol