Gaussian Mixture Model:

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

Tyes of clustering:

Hard - clusters don't overlap
Elements either belong to cluster or not

soft clustering - clusters may overlap
Strength of association between clusters and instances.

Mixture Model:

Probabilistic way of doing soft Clustering

Each Cluster: Generative model (Gaussian or multinormal)

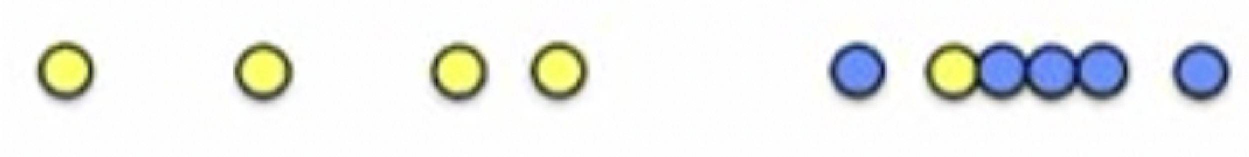
Parameters(ex:mean,covariance are unknown)

Expectation-Maximization:
Automatically discovers all parameters for k sources.

Ex: **Mixture Model in 1D:**

Observations (x1,xn) points.
K=2 gaussian with unknown mean and variance.

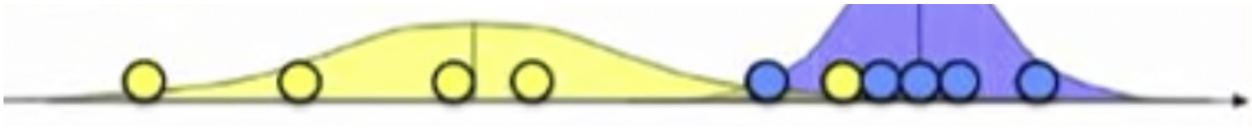
Estimation is trivial if we know the source of observations



If we already know the K=2 and already know which points belong to which K, then the only thing to do is estimate the mean and variance for blue and yellow.

$$\mu_b = \frac{x_1 + x_2 + \ldots + x_{n_b}}{n_b}$$

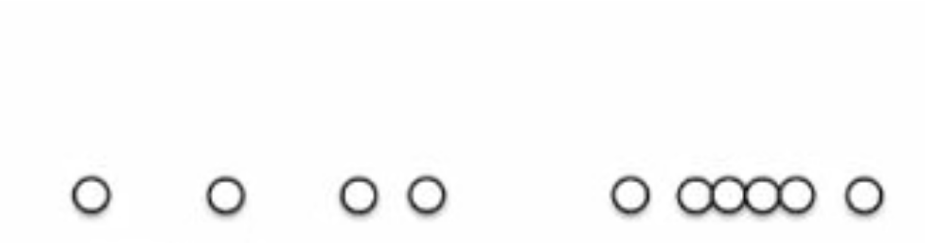
$$\sigma_b^2 = \frac{(x_1 - \mu_1)^2 + \ldots + (x_n - \mu_n)^2}{n_b}$$



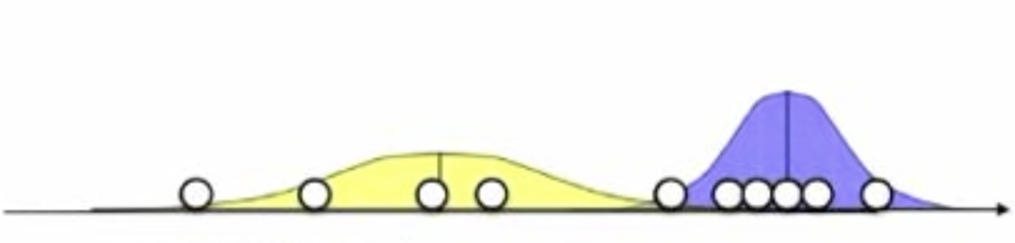
If It is done if we would know which point came from which cluster.

What if we don't know the source?

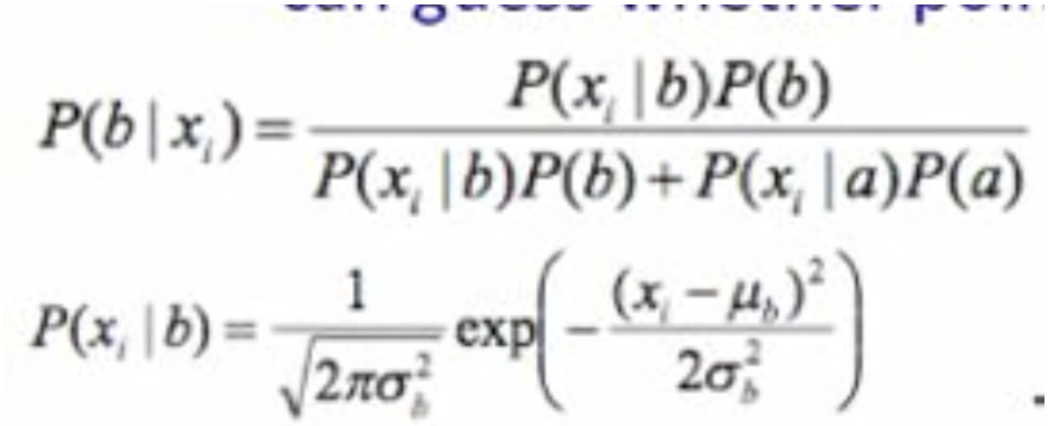
We don't know which belongs to yellow and which belongs to blue.



But if we knew the parameters of the gaussian (mean, variance),
Then we can guess more likely which point belongs to yellow and which point belongs to blue.



We can do this by the Bayes rule and the gaussian formula.



$$P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right).$$

Once you compute the blue, yellow is just **1-bi.**

Expectation-Maximization:

Chicken and egg problem:

- Need means and variances of the 2 clusters to guess the source of points.
- Need to know the source to estimate mean and variance.

EM algorithm:

- Starts with 2 randomly placed gaussian.
- Then, for each point (P(b|xi) does it look like it came from yellow or it came from the blue?
- Adjust the mean and variance to fit points assigned to them.
- Iterate until convergence.

Choosing the covariance type

- *covariance_type="**diag**", which means that the size of the cluster along each dimension can be set independently, with the resulting ellipse constrained to align with the axes.*

- *A slightly simpler and faster model is covariance_type="**spherical**", which constrains the shape of the cluster such that all dimensions are equal. The resulting clustering will have similar characteristics to that of k-means, though it is not entirely equivalent.*
- *A more complicated and computationally expensive model (especially as the number of dimensions grows) is to use covariance_type="**full**", which allows each cluster to be modeled as an ellipse with arbitrary orientation*

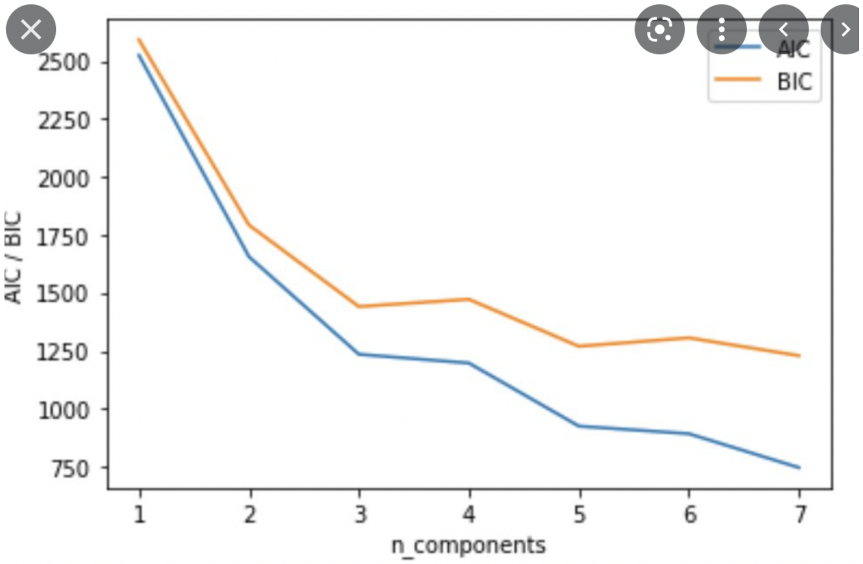
Diff between Kmeans and EM algorithm?

Kmeans does hard clustering, it takes a point and puts it into one cluster or the other. Whereas the EM algorithm computes the probability it goes to the blue or the yellow cluster. It does soft clustering.

Probabilities between (0 and 1)

Determine GMM Clusters with BIC:

The fact that GMM is a generative model gives us a natural means of determining the optimal number of components for a given dataset.
A generative model is inherently a probability distribution for the dataset, and so we can simply evaluate the *likelihood* of the data under the model, using cross-validation to avoid over-fitting.
Another means of correcting for over-fitting is to adjust the model likelihoods using some analytic criterion such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Scikit-Learn's GMM estimator actually includes built-in methods that compute both of these, and so it is very easy to operate on this approach.



From Wiki,

BIC: In statistics, the Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

Hierarchical Clustering:

Types:
Agglomerative
Divisive

In agglomerative clustering, each data point is considered a cluster.
At each iteration, similar clusters are merged with other clusters until one cluster or k clusters are formed.

The basic algorithm of Agglomerative is straightforward.

- Compute the proximity matrix
- Let each data point be a cluster
- Repeat: Merge the two closest clusters and update the proximity matrix
- Until only a single cluster remains

Key operation is the computation of the proximity of two clusters

The Hierarchical Clustering Technique can be visualized using a Dendrogram.

A **Dendrogram** is a tree-like diagram that records the sequences of merges or splits.

In Divisive Clustering:
It is exactly the opposite of Agglomerative clustering.

In Divisive Hierarchical clustering, we consider all the data points as a single cluster and in each iteration, we separate the data points from the cluster which are not similar
Each data point that is separated is considered as an individual cluster. In the end, we'll be left with n clusters.

As we're dividing the single clusters into n clusters, it is named Divisive Hierarchical clustering.

HOW DO WE CALCULATE THE SIMILARITY BETWEEN TWO CLUSTERS?
Calculating the similarity between two clusters is important to merge or divide the clusters.

Single link(Min)
Average link
Complete Link(Max)
Distance Between Centroids
Ward's Method

Space and Time complexity;
Space complexity = $O(n^2)$
Time complexity = $O(n^3)$

Silhouette Score:

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: *Means clusters are well apart from each other and clearly distinguished.*
0: *Means clusters are indifferent, or we can say that the distance between clusters is not significant.*
-1: *Means clusters are assigned in the wrong way.*

Silhouette Score = (b-a)/max(a,b)
a= average intra-cluster distance i.e the average distance between each point within a cluster.
b= average inter-cluster distance i.e the average distance between all clusters.

What is the Facebook algorithm?

Let's look at how the algorithm works based on the article we read https://blog.hootsuite.com/facebook-algorithm/.

The Facebook algorithm determines which posts people see and in what order they see them every time they check their Facebook feed. scores posts and then arranges them in descending, non-chronological order of interest for each individual user. We dont know the exact nuances but We do know that, like all social media recommendation algorithms, one of its goals is to keep users on the platform so that they see more advertisements.

In fact, Facebook faced heat in 2021 because the algorithm was prioritizing controversial content. Controversy often gets the highest engagement and can even trigger “compulsive use” of the platform.

According to Facebook, the algorithm is all about assisting users in "discovering new content and connecting with the stories they care about the most," while "keeping spam and misleading content at bay." As you'll see below, recent Facebook algorithm changes have aimed to address content and privacy concerns.

Fb Algorithm History:

2009*: Facebook first algorithm to bump posts with the most Likes to the top of the feed.*

2015: Facebook begins de-ranking Pages that post excessively promotional content and introduced the "See First" feature, which allows users to specify whether they want a Page's posts to be prioritized in their feed.

2016: FB adds a “time spent” ranking signal to measure a post’s value based on the amount of time users spent with it even if user does not like the post.

2017: Now Fb started Weighting reactions (Ex: ❤️😍😒😩😭😠😠)
Then another ranking was added for video: completion rate(videos which are watched till the end without skipping)

2018: The FB new algorithm prioritizes “posts that spark conversations and meaningful interactions

2019: Fb prioritizes “high-quality, original video” that has watching time more than one minute especially video which has high attention and are more than 3 min long and also prioritize content from "close friends"

2020: fb shares certain algorithmic insights to help users understand how it distributes content and allows them to take control of their data to provide greater feedback to the algorithm.

2021: three main ranking signals:
Who posted it
Type of content
nteractions with the post

Does the social media platforms "care" if they are spreading misinformation?

When it comes to social media, the most popular products are from the meta family (Facebook, Instagram, WhatsApp, etc..). Despite the fact that there are other platforms with an equal number of users, such as Twitter, Snapchat, and TikTok, we will focus on Facebook for our use case.

Popularity and recommendations in social media are based on the most important aspect of how the algorithm works and what key factors it considers before delivering it to users.

The goal of the algorithm is to provide users with relevant material that is tailored to customer needs, ensuring that they remain interested and committed to the platform.

The algorithms are set up in such a way that if the majority of your friends **like, share, or comment on a feed,** it is given higher attention. Then there are feeds with a **higher engagement rate**, **locaiton popularity**, **interest** as well as a variety of other factors.

Now coming to the **misinformation**

False information has become common on social media, particularly during election seasons. According to research, false news peaked on social media during the 2012 and 2016 presidential elections, and a bipartisan Senate committee found that the Russian government used Facebook, Instagram, and Twitter to spread false information and conspiracy theories, as well as stoke divisions, before and after the 2016 election.

False rumors spread faster and wider than true information because we as people believe in controversial and high engaging contet.

Also, people spreading false information are distracted and does not care about the truth.

One of the Important factors of misinformation is the role of **echo chambers in the spread of misinformation**. When echo chambers and the extent of homophily are limited, misinformation does not spread very far.

"The engagement effect can lead to endogenous echo chambers as documented by Levy (2020) for Facebook"

"We as users are not concerned with the source/truth; rather, we are concerned with the material that is intriguing and engaging, and only in the latter sentence are we concerned with the truth"

According to my analysis, during the early stages of social media evolution, in order to enhance user engagement, the algorithms didn't take into account a lot of aspects, but now, thanks to a slew of new legislation and a slew of other external reasons, social media is very concerned about wrong information.

The only way to limit the spread of false information is to tighten regulations and try to filter or stop it at the source.

There has already been a lot of work done to delete phony accounts and fake posts, but people who are intending to create fake news cannot be prevented even if we take steps to stop spreading misinformation unless we are certain about the source.

So, do social media platforms "care" if they are disseminating false information?

It depends; it is a blend of truth to some extent but not entirely truthful.