Week-5

Multivariate Gaussian.

$$g(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{+\frac{1}{2}((x-\mu)^T \Sigma^{-1} (x-\mu))}$$

GMM

$$\lambda = \sum_{k=1}^{K} w_k g_k (x, \mu_k, \Sigma_k)$$

$$\sum_{k=1}^{K} w_k = 1$$

How to train a GMM

EM - Expectation Maximization

↓ Statistics ↓ fitting GMM to given data

1) Iterative Procedure

In each Iteration → compute statistics → Means (wts) → σ

Objective: maximizing representation of data of X by λ

Let there be λ_1 & λ_2

$$P(X/\lambda_1) \& P(X/\lambda_2)$$

$$P(X/\lambda_2) > P(X, \lambda_1)$$

As λ_2 is later to λ_1

Page No.
Date

Convergence

$P(X/\lambda_2) - P(X/\lambda_1) \leq$ Threshold.

To begin with $X = \{x_1 \dots x_M\}$
+ Intial Model λ - GMM

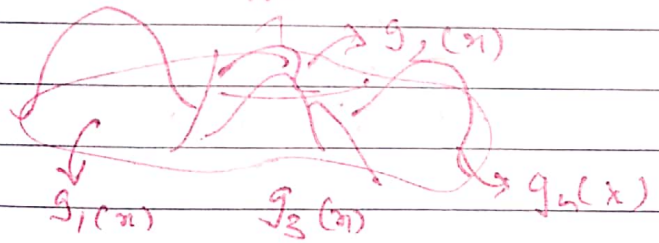
Ways to Generate Initial Model → Random
↓ k-means clustering algo

"k" → M data.
partition ↓ non-overlaping
k_n partition.
} Random.

k-means

~~clus~~ cluster data points into "k" non-overlaping data.

$$\lambda = \sum_{k=1}^{K} w_k g_k(x_i, \mu_k, \Sigma_k)$$



$$\lambda = \sum_{i}^{k} w_k g_k(x, \mu_k, \Sigma_k)$$

* EA

$X = \{x_1 \dots x_M\}$

$P(k/x_i, \lambda)$ = x_i belonging to k^{th} Mixture model in λ model.

$$= \frac{w_k g_k(x_i/\mu_k, \Sigma_k)}{\sum_{i=1}^{k} w_m g_m(x_i/\mu_m, \Sigma_M)}$$

$$\bar{\omega}_k = \frac{1}{M}\sum_{i=1}^{M} P(k/x_i, \lambda)$$

$$\bar{\mu}_k = \frac{\sum_{i=1}^{M} P(k/x_i, \lambda) \cdot x_i}{\sum_{i=1}^{M} P(k/x_i, \lambda)}$$

$$\bar{\Sigma}_k = \frac{\sum_{i=1}^{M} P(k/x_i, \lambda)\,(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{M} P(k/x_i, \lambda)}$$

$$\bar{\lambda} = \sum_{k=1}^{K} \bar{w}\, g_k(x_i, \bar{\mu}_k, \bar{\Sigma}_k)$$

$$P(X/\bar{\lambda}) = P(X, \bar{\lambda}) \geq \text{Threshold}$$

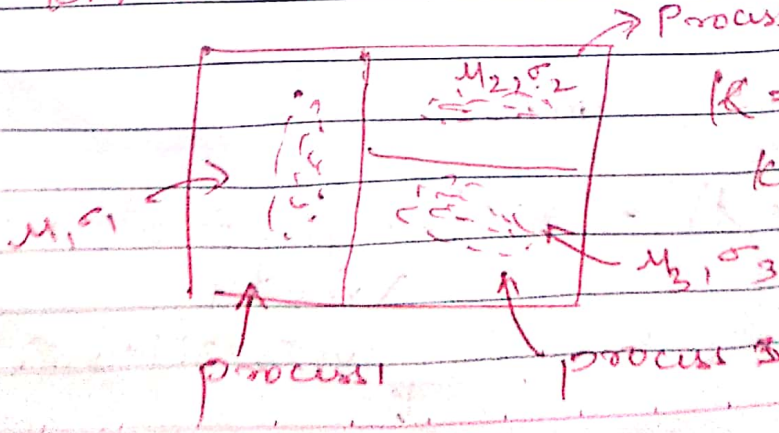
$$\lambda = \bar{\lambda}$$

$$P(X/\bar{\lambda}) = \frac{1}{M}\sum_{i=1}^{M} P(x_i/\bar{\lambda})$$

$$\max_i P(x_i/\bar{\lambda})$$

Clustering by GMM

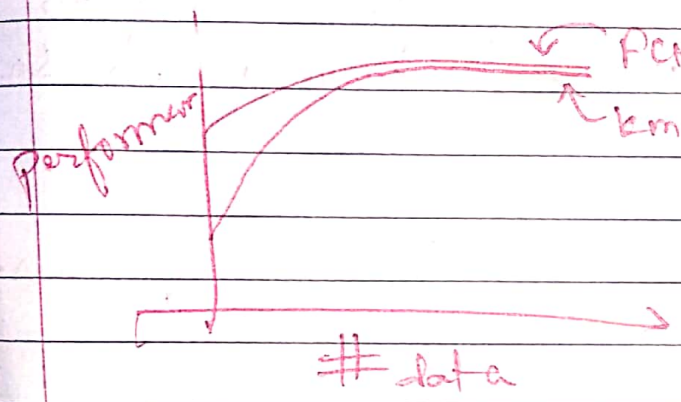
kmeans -> FCM -> GMM



FCM Advantages over K-means.

Mean vector in FCM & are more stable as it gives more data for mean vector Calculation.

GMM : takes care spread (or) of data So better fit than FCM.



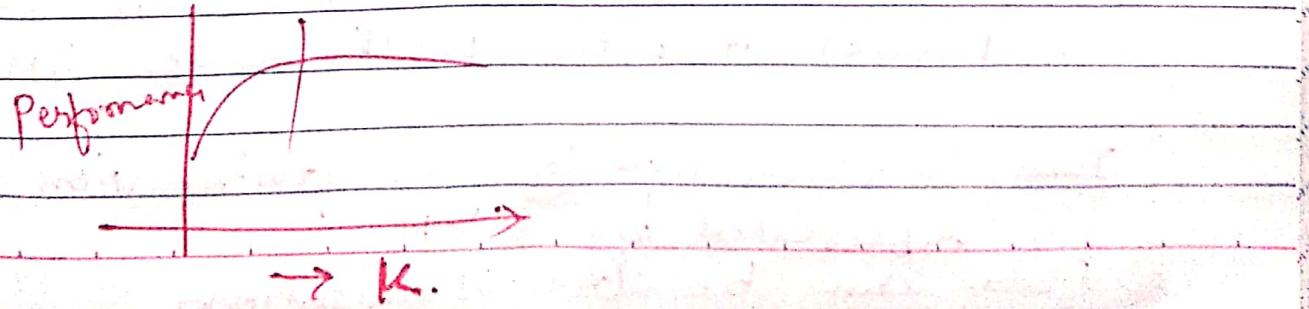
FCM performs better in case of less data.

When we go for GMM we have μ, σ^2 (variance)

μ - First order statistics (FOS)
σ^2 - Second order (SOS)
as GMM has SOS it is better representative of model

data requirement Grows in exponential order SOS requires much more data than FOS.

How to chose 'K' in GMM



Cluster purity:-

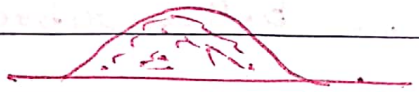
Depending value of k that many clusters are created, we also have reference which point belongs to which cluster

ideally if there are 3 cluster, when we put k = 3, data points belonging to particular cluster (reference) falls to that cluster then cluster purity is said to 100%

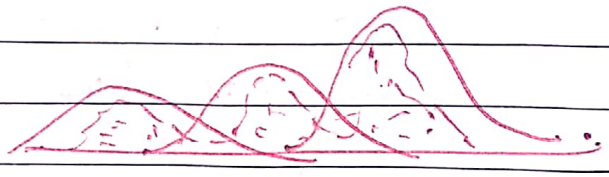
performance is measured in terms of cluster purity.

~~How to set~~

Why Gaussians are used so extensively.



A Lone Gaussian & a fit Gaussian



A mixture of Gaussian model data can be fit. (K=3)

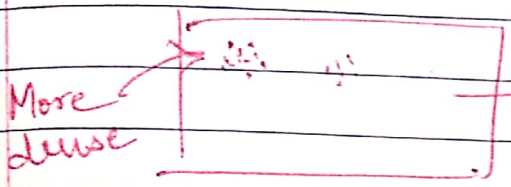
So a non-Gaussian distribution can be model it with multivariate GMM.

So theoritically all distribution can be represented by GMM

⇒ Data can be Gaussianisation

"Density" Based clustering

↓ literal meaning



→ concept of cluster

↓

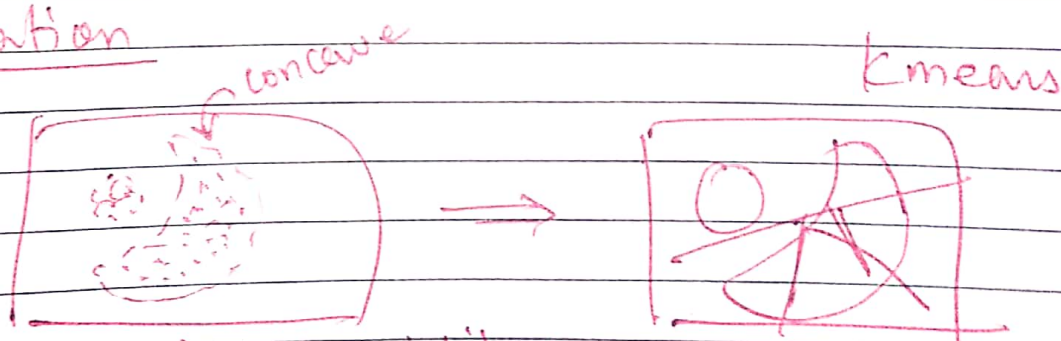
correlated it with density.

More data points in neighbours tells us about presence of a ~~data~~ cluster

Denseness is also attribute (or exploited) to come up with a type of clustering

DBSCAN (Density based spacial clustering Application & noise)

Motivation



In kmeans, FCM, GMM there is no means to reject outliers

in Density based we could argue that it is not dense enough to be a cluster so an outlier

→ Prone to outlier : outlier are assign to Cluster, after that they pull mean vector towards them which makes their cluster prone to far point than near points.

DBSCAN Can

DBSCAN Density based Can find out any arbitary shape cluster without being affected by noise.

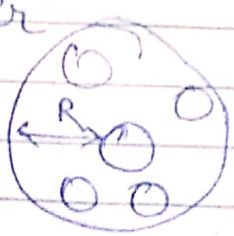
Density Based Spatial Clustering of Application 4 Noise

R: Radius of Neighbour hood
M: Min number of neighbour

Intitution / Idea.

If a point is in a cluster it Should be near to lots of other points in that cluster

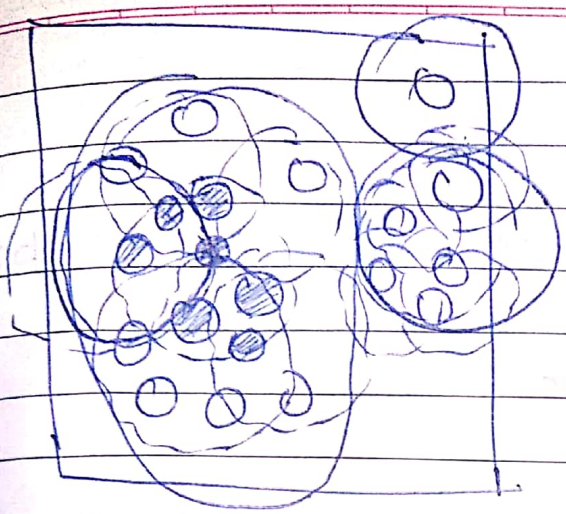
R



If a taking any point if in a given Radius 'R' it have more than M neighbour we call it Dense area or cluster.

3 types of Data point

→ Core
→ Border
→ Outlier



R = 2
M = 6.

Core : If there are M points or more in a given radius.

Border If there are points but less than M it is border

Outlier : if points in neighbourhood is zero

~~Soft Constraints~~

Core points surrounded by border points generate arbitory shape.

Disadvantage.

⇒ Need to select 2 parameter 'R' & 'M'

Advantage.

⇒ Arbitary shaped cluster
⇒ Robust to Classifier
⇒ Doesn't require specification of clusters.