Gaussian Mixture Model.
using EM method.

Why do we need gaussian model. (Advantages)

1) It is very similar to the normal distribution
2) It is easier to do mathematical manipulations and calculations using a gaussian model as compared to the other ones.
   → we can differentiate infinite times.

Sometimes the nature of the distributions is not gaussian; but we divide it into many models assuming each of the individual model is gaussian. This is called the gaussian mixture model or GMM.

Univariate Gaussian Distribution (1D)

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ ← mean, σ ← covariance

Multi variate gaussian Distribution:

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

x: this is a vector with dimension of x; μ: mean; Σ: covariance

To apply GMM we need to estimate (Σ, μ) of a distribution.

One method → Maximum Likelihood estimation. (ML)

* ML Method

1.) Considers the log of Gaussian distribution

$$\ln P(X|\mu, \Sigma) = -\frac{1}{2}\ln 2\pi|\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

2.) We take its derivative and equate it to zero. ie.

$$\frac{\partial \ln p(x|\mu, \Sigma)}{\partial \mu} = 0 \quad , \quad \frac{\partial \ln p(x|\mu, \Sigma)}{\partial \Sigma} = 0$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad , \quad \Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

where N is the number of samples or data points.

Gaussian Mixture.

* it is the linear superposition of Gaussians !.

$$p(x) = \sum_{k=1}^{K} \pi_k \, N(x \mid \mu_k, \Sigma_k)$$

- K: no. of gaussians
- π_k: mixing coeff : weightage for each gaussian distance.
- $N(x \mid \mu_k, \Sigma_k)$: normal multivariate gaussian distribution

* Normalization and positivity require:

$$0 \le \pi_k \le 1 \; ; \quad \sum_{k=1}^{K} \pi_k = 1$$

if we consider log-likelihood

$$\ln p(x \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln p(x_n) = \sum_{n=1}^{N} \ln \left[\pi_k N(x_n \mid \mu_k, \Sigma_k) \right]$$

Maximum likelihood doesn't work here as there is no closed form solution.

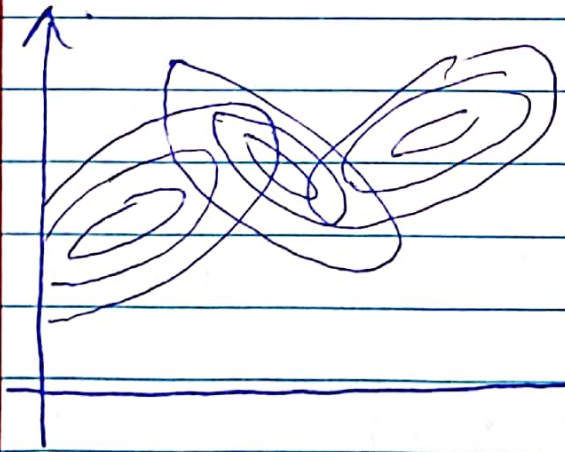
So for calculating parameters we can use the Expectation Maximization technique.

or the EM technique.

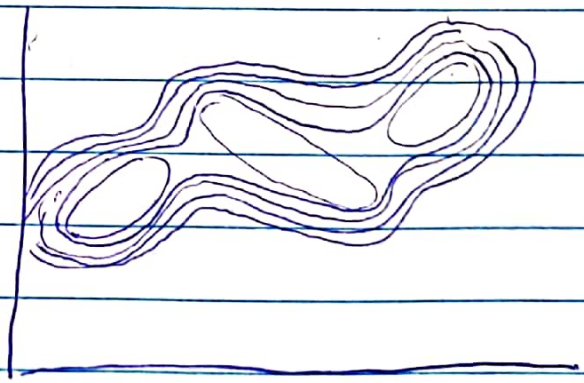
Extra

⇒ Mixture of 3 gaussians Example:

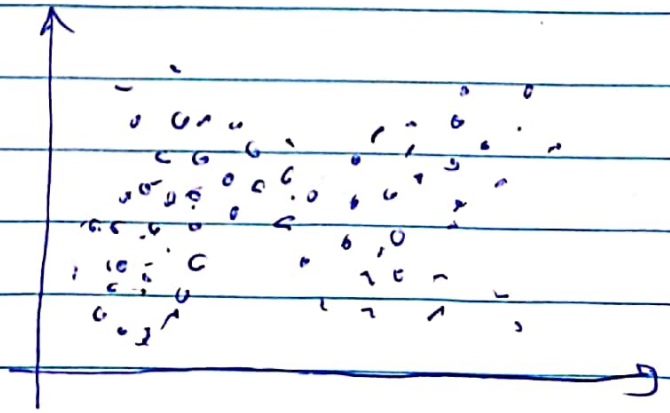
Let us represent the contour plot of 3 gaussians put together



This is the combined Contours plot of the sam



Sampled data.



Latent variable : posterior prob.

* We can think of the mixing coeff as priors probabilities for the components

* For a given value of 'k', we can evaluate the corresponding posterior probabilities, called responsibilities.

From Bayes Rule.

$$\gamma_k(x) = p(k|x) = \frac{p(k)p(x|k)}{p(x)}$$

latent variable

$$= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)} \quad \text{where } \pi_k = \frac{N_k}{N}$$

Interpret N_k as the effective no. of points assigned to cluster K.

What does EM algorithm do?

* EM algorithm is an iterative optimization technique which is operated locally to find set of parameters.

which all parmeters do we need to estimate?

→ we need to estimate many parameters lets say we are planning to use K gaussian surfaces then we need to predict μ_K, Σ_K and π_K

μ_K ← mean matrix
Σ_K ↓ covariance matrix
π_K ↓ This is a scalar.

for each gaussian.
hence in total we need to ~~guess~~. $K \times 3$ parmeters.

* There are two steps in EM

~~1) Expectation step~~

1.) Estimation step:
for a given parameter values we can compute the expected values of the latent variable

2.) Maximization step:
updates the parameters of our model based on the latent variable calculated using ML method.

Algorithm for GMM:

Given a gaussian mixture model the goal is to maximize the likelihood function with respect to the parameters compromizing the means and covariances of the components and the mixing coeffs.

1) Initialize the means μ_j; Covariances Σ_j and mixing coeff π_j; and evaluate the initial value of the log likelihood the starting points can be completely random.
also we can initialize them at some known points such as the mean of the total dataset. etc.

* The idea is that closer we are to the final solution, faster will be our solution.

2) E Step, Evaluate the responsibilities using the current parameter values.

$$\gamma_k(x) = \frac{\pi_k N(x|\mu_k \Sigma_k)}{\sum_{k=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$$

summing up the gaussians with the estimated parameters.

3.) M. Step : Re estimate the parameters using the current responsibilities
Till now we have calculated γ_j's

$$\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n) x_n}{\sum_{n=1}^{N} \gamma_j(x_n)}$$ } same as ML

$$\Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)(x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^{N} \gamma_j(x_n)}$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(x_n)$$

4.) Evaluate log likelihood

$$\ln(p(X|\mu, \Sigma, \pi)) = \sum_{n=1}^{N} \ln\left[\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)\right]$$

if there is no convergence return to step 2.
otherwise we have reached our final answer.

We need to repeat the steps 2, 3, & 4 until we reach convergence.

Again There can be more that one methods to check convergence

1) If the difference between the consequtive estimations of the log likelihood is below a certain threshold we can say that the estimation converged.

Another method may be to do the same but by using the parameters π_j, Σ_j, μ_j instead of log likelihood.

* EXTRA

What is γ_k (Latent function)?

→ The mathematical definition of γ_k is:

$$\gamma_k = \begin{cases} 1 & x \in \text{gaussian } k \\ 0 & \text{otherwise} \end{cases}$$

It can be considered as a type of probability density fuction. It behaves like the dirac delta function of K-means.

* EXTRA

How to pick the value of K in GMM

* One method is to pick that value of K which minimizes the value of

$$L = \log(P(x_1, x_2 \ldots x_n)) = \sum_{i=1}^{n} \log \sum_{k=1}^{k} P(x_i|k)P(k)$$

using brute force.

→ but this will result in a value of K that is equal to ~~number~~ number of training data points.

* Another option is to split the data points into training(T) and validation(V) set.
→ and then for each K: fit the parameters of T and measure the likelihood of V. based on the above results we can select a K which will give min. likelihood.
But again, in this method as well we may end up with a very large no. for K.

There are many methods of selecting a K which involve matrices and complex calculations.
According to Occam's Razor if we have two explainations to a certain problem in datasets we must always select the explaination that is simpler that way there are less chances of complications.

Hence: Occam's Razor: pick "simplest" of all models that fit. just
Now instead of using likelihood we penalize it as the complexity of the model increases.
→ Akaike info. criterion (AIC): $-L+K$
→ Bayesian info. criterion (BIC): $-L+\frac{1}{2}K\ln(N)$

here $L \rightarrow$ is the likelihood;
$K \rightarrow$ is no. of gaussians.

These are two popular types of penalties.
Instead of maximizing L, minimize $-L$ and add the defined penalty.

Finally now that we have defined this penalized model we can get K by the two methods discussed before.