GAUSSIAN MIXTURE MODEL (GMM)

→ Gaussian mixture model are a probablistic model for representing normal distributed subpopulation within an overall population.

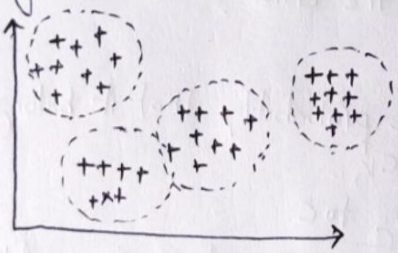
→ Mixture modds in general don't require which subpopulation a data point belongs to, allowing the model to learn the subpopulation automatically.

→ For example, in modeling human being data, height is typically modelled as normal distribution for each gender with a mean approximately 5'10" for males and 5'5" for females. Given only height data and not gender assignment for each data point, distribution of all height would follow the sum of two scaled (different variance) normal distribution.

→ A model making this assumption, the parameter follow normal distribution.

→ GMM assume there are a certain number of Gaussian distributions and each of these distribution represent a cluster. Hence a Gaussian Mixture model tends to group the data points belong to a single distribution together.

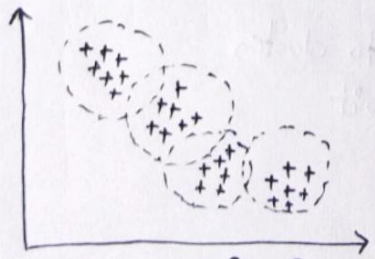
Motivating GMM : Weakness of K Means.



- In this case, data is well seperated data.
- K means algorithm can quickly label those clusters.
- There might be slightly overlap between two clusters assignment of point between them.

- Unfortunately, K means has no measure of probability.

→ K means model is that it places a circle at the center of each clusters, the radius acts as a hard cutoff for cluster assignment, any point outside the circle is not considered a member of cluster.



→ By eye, we can recognized that transformed clusters are non-circular thus circular clusters would be poor fit.

→ K means try to force-fit data into four circular clusters, resulting circles overlap.

→ GMM address this issue, since it is probablistic model, it find possible probablistic cluster assignment.

What is Expectation - Maximization? (EM)

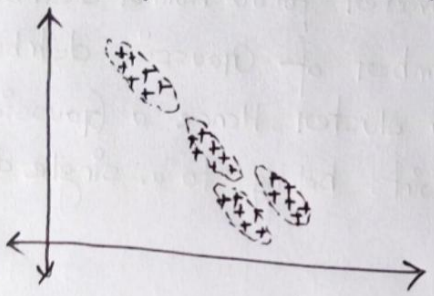
→ EM is a stahstical algorithim for finding the right model parameters. We typically use EM when the data has missing values or in other word when the data is incomplete.

Broadly, EM algorithim has two steps:

i) E-step → In this step, available data is used to estimate (guess) the value of the missing variables.

ii) M-step → Based on the estimated values generated in E-step the complete data is used to update the parameters.

After applying GMM Clustering



→ Clusters are organised diagonally.

→ Gaussian mixture model worked better in this case (non-spherical data)

→ K means only considers mean to update the centroid while GMM takes into account the mean as well as variance of the data.

Normal EM follows the steps –

E-step → For each point x_i, calculate the probability that it belongs to cluster divide by distribution $c_1, c_2, \cdots, c_k$.

$$\text{E-step} = \frac{\text{Probability } x_i \text{ belong to } c}{\text{Sum of probability } x_i \text{ belong to } c_1, c_2, \cdots, c_k}$$

Value will be high when point is assigned to righ cluster & lower otherwise.

M step →

New density is defined by ratio of the number of points in the cluster & total number of points.

$$= \frac{\text{No. of points assigned to cluster}}{\text{Total number of point}}$$