"Clustering using GMM"

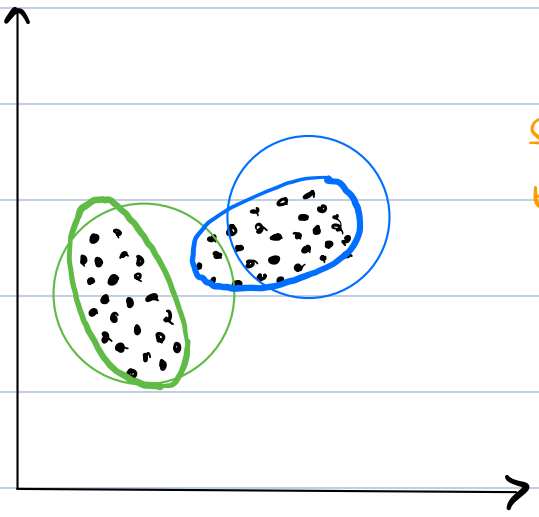
'k-means

1. Initialization of centroids
2. Data points Assignment
3. Re-calculation of centroids
4. Till Convergence

} Repeated

How to decide 'k' ?

1. Domain knowledge
2. Silhoutte Score
3. Elbow Curve

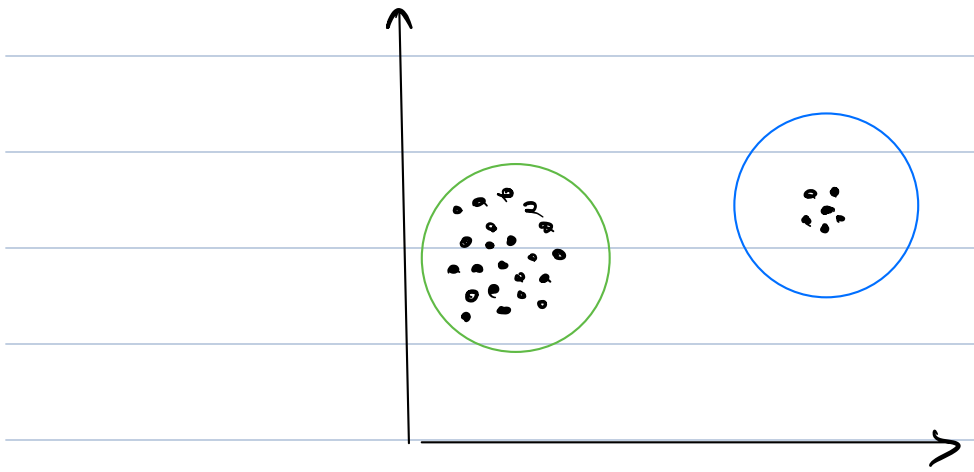


Spherical clusters → k means

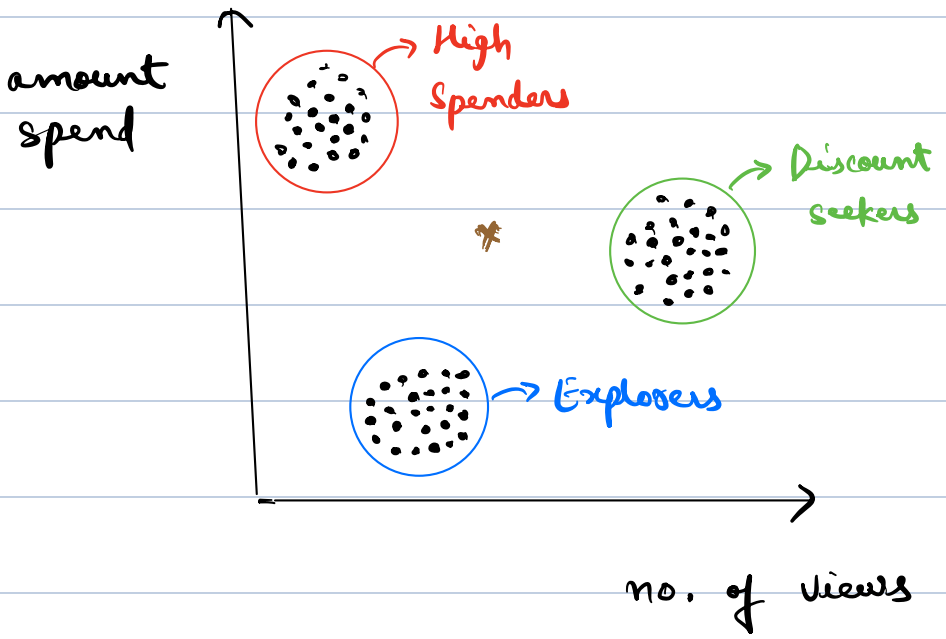
Elliptical clusters → GMM

Limitations of k-means:

1. Always given spherical / globular clusters.
2. Always gives almost equal sized clusters.
3. Hard Clustering : For each data point only 1 cluster is assigned.



"ecommerce data"



x → New customer

(Wealthy but also price-conscious) → 60% → High spenders

→ 40% → Discount seekers

↳ Calculate probabilities to see where a data point belongs

[Soft Clustering]

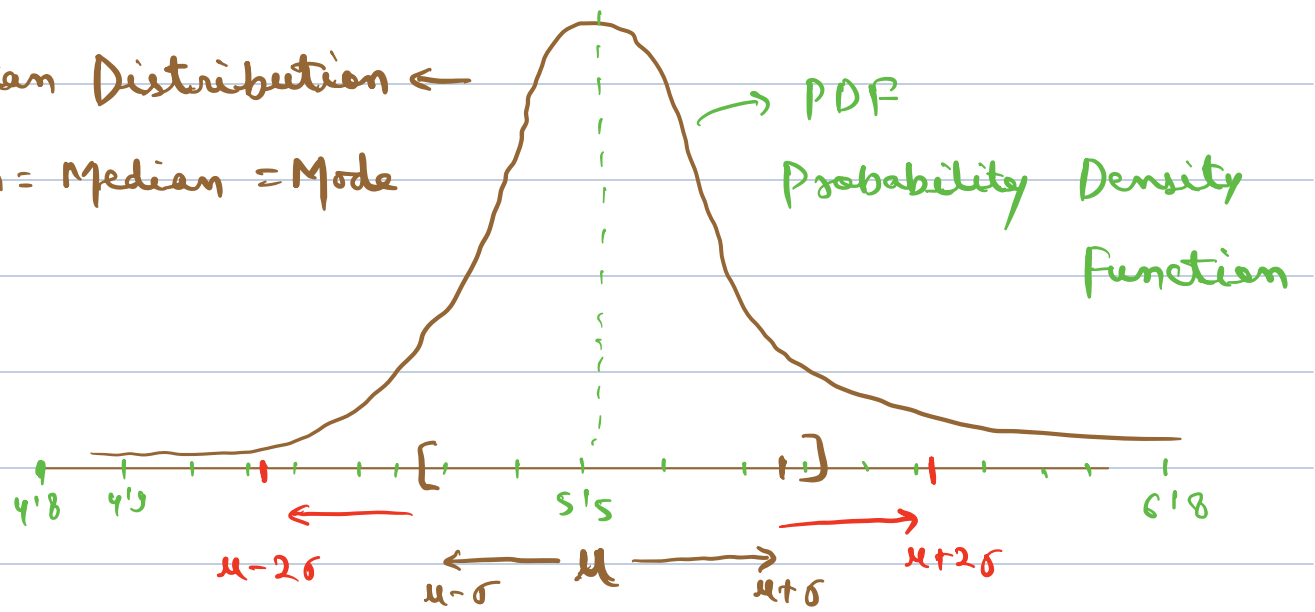
Example: We take a survey of heights of Indian citizens.

Gaussian Distribution ←

→ Mean = Median = Mode

→ PDF

Probability Density Function



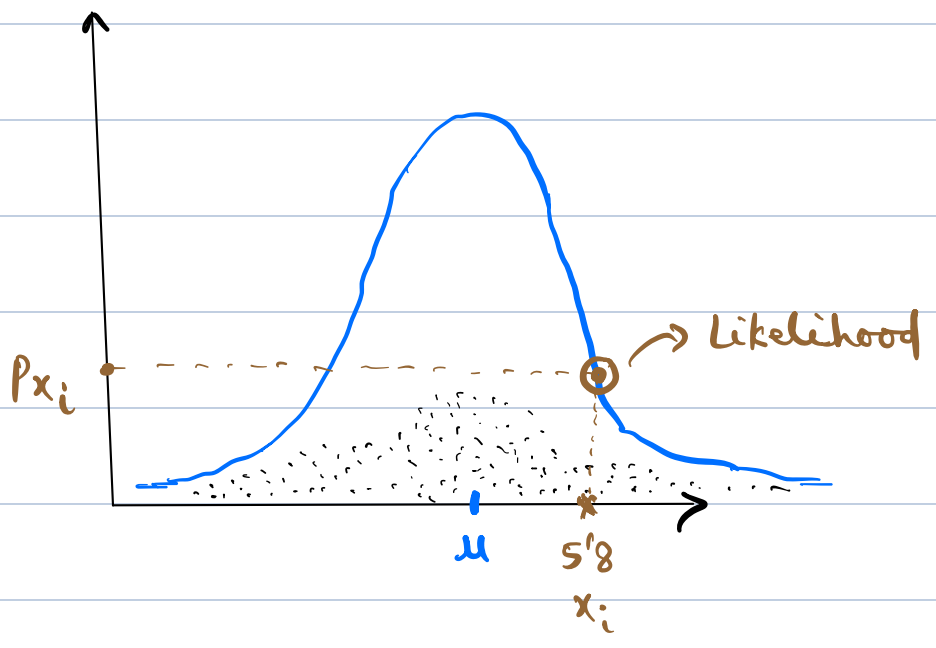
On average, height of Indian citizen is 5'5

Empirical formula :

$\mu \pm \sigma$ → 68% of data ✓

$\mu \pm 2\sigma$ → 95% of data ✓

$\mu \pm 3\sigma$ → 99% of data



$$PDF = p(x : \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

Data → μ and σ

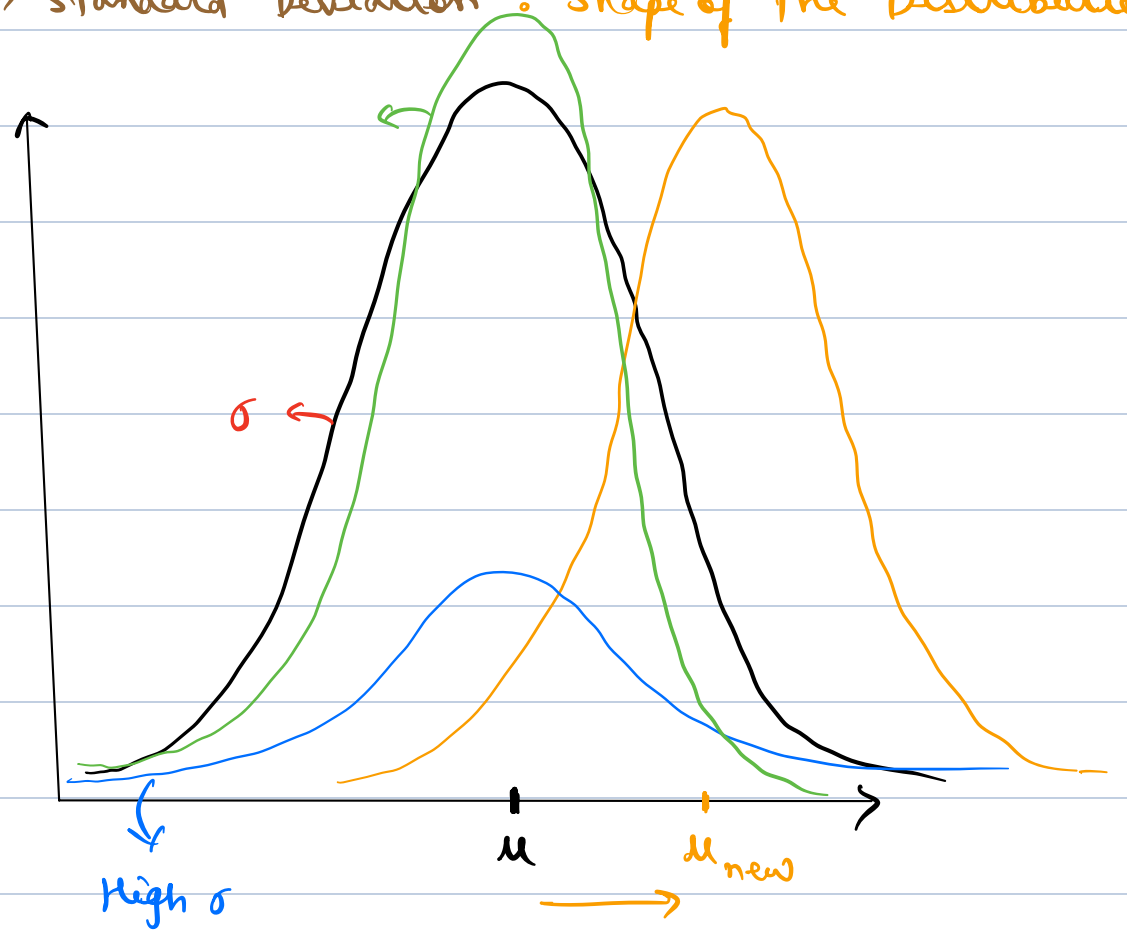
Data Point → x

"1-D Gaussian"

- PARAMETERS

→ Mean : Position of the Distribution

→ σ → Standard Deviation : Shape of the Distribution



$$PDF = p(x : \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

2-D Gaussian !

↳ 2 Distributions

→ μ_1

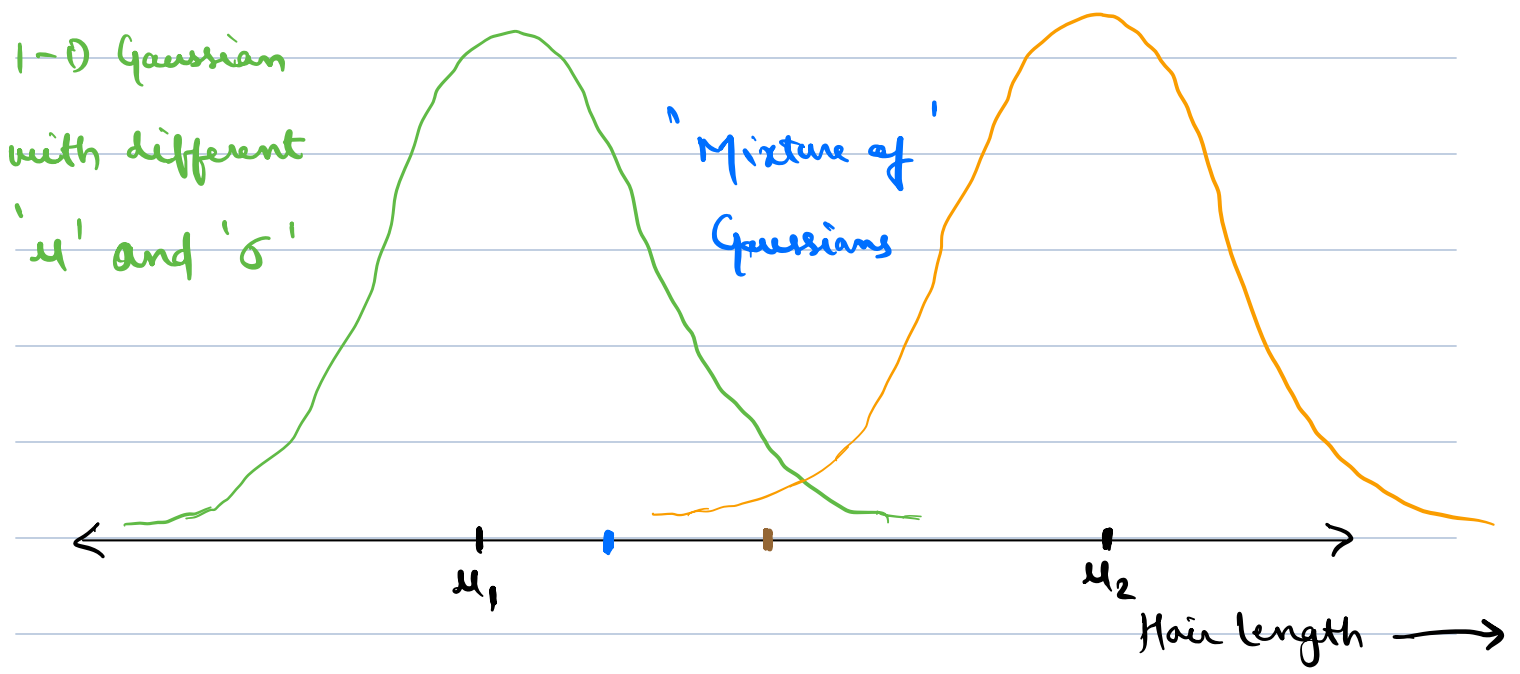
Position → Means → μ_1, μ_2

Shape → S.D of Distributions → σ_1, σ_2

* Covariance → Orientation / Angle of the Ellipse

Ex:

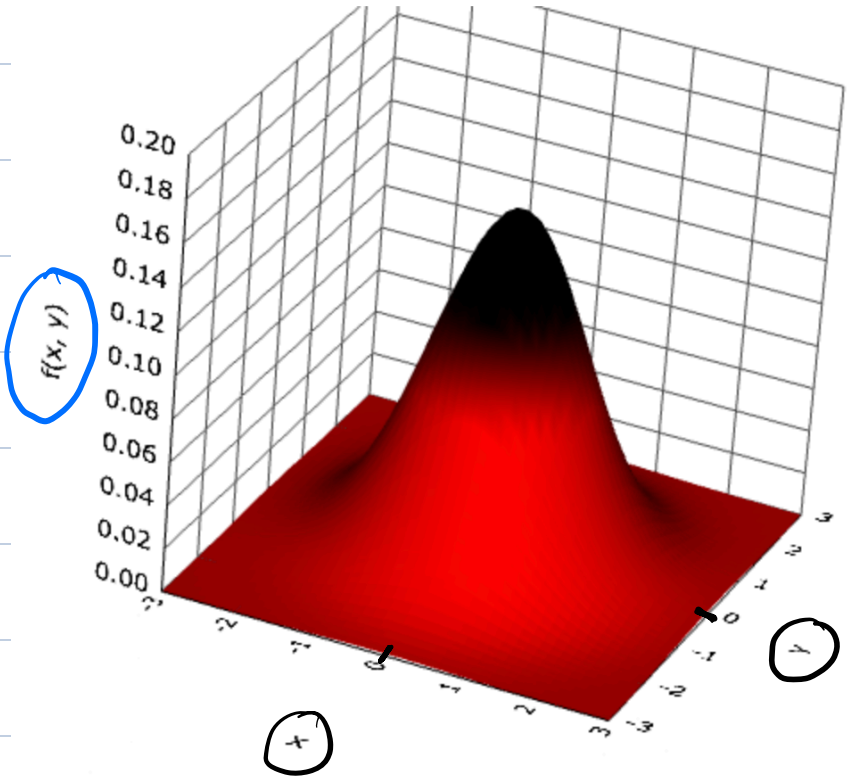
Survey was done → to find hair length of all Indian citizens

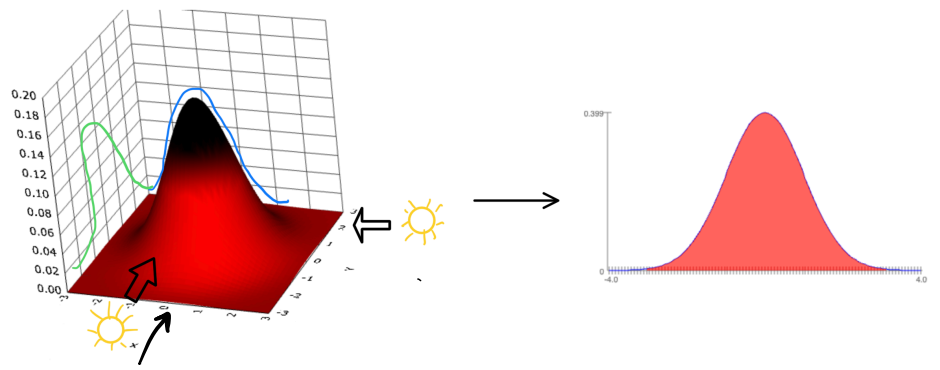


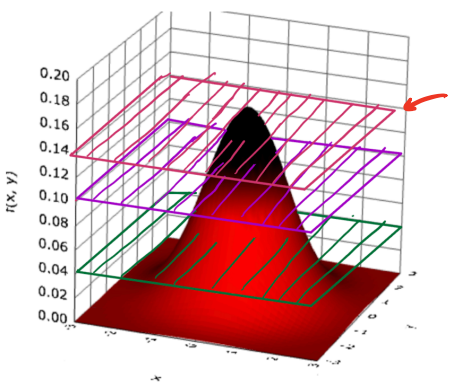
Uni-Modal → 1 guassian

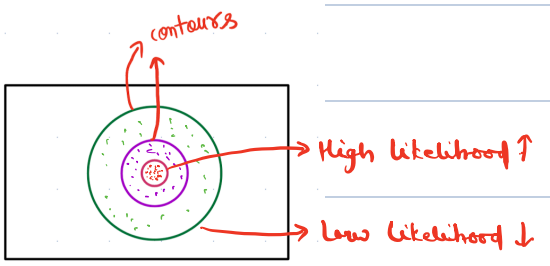
Bi-Modal → 2 guassians

Multi-Modal → Multi-gaussians









Steps of GMM :

1. Initialize 'k' guassians
2. Expectation Step

↳ Calculate likelihood / probability

③ Maximization Step

↳ Update mean and standard deviation

↳ Weighted avg. approach

④ Till the convergence.

Sample Calculation:

$$D = \{x_1, x_2, x_3, \ldots\ldots x_n\}$$

Step 1 : Choose 'k' random points as mean.

Initialize Means $\longrightarrow \mu_1 = x_1$,

$\longrightarrow \mu_2 = x_3$

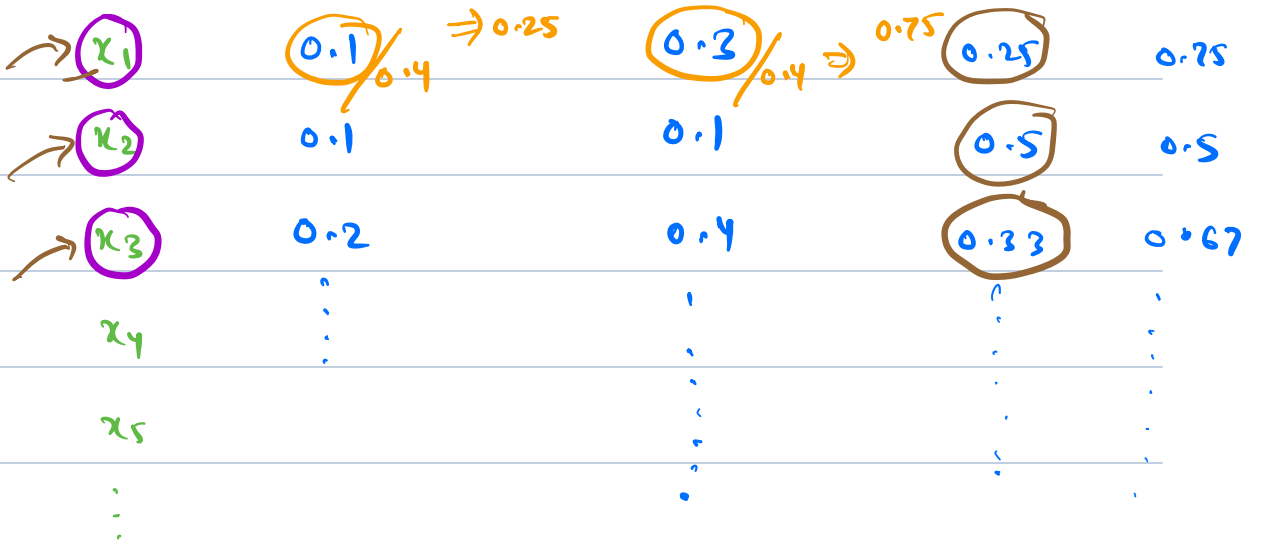
Initialize the variance

$$\sigma_1 = \frac{1}{n}\sum(x_i - \mu_1)^2$$

$$\sigma_2 = \frac{1}{n}\sum(x_i - \mu_2)^2$$

Step-2 : Expectation Step

$p(x, \mu_1)$ $p(x, \mu_2)$ $p_n(x, \mu_1)$ $p_n(x, \mu_2)$



Normalization of Probabilities

↳ so that sum of all prob. is equal to 1

③ Maximization

likelihood of data point in μ_1

$$\mu_{1\,new} = \frac{\sum_{i=1}^{n} p(x,\mu_1) \cdot x_i}{\sum_{i=1}^{n} p(x,\mu_i)}$$

x_i → data point

$\sum_{i=1}^{n} p(x,\mu_i)$ → Weighted average

$x_1 = 10$

$x_2 = 20$

$x_3 = 30$

$$\mu_{1\,new} = \frac{10 \times 0.25 + 20 \times 0.5 + 30 \times 0.33}{0.25 + 0.5 + 0.33}$$

$$\mu_{2\,new} =$$

Variance → σ^2

Variance

$$\sigma^2 = \frac{\sum p(\mu_{x_i}) (x_i - \mu)^2}{\sum p(\mu_{x_i})}$$

$\sum p(\mu_{x_i})$ → Weighted average of likelihood

GMM

Repeats Step 2 and Step 3 till Convergence.

k-means	GMM
① Hard Clustering → Distance	① Soft Clustering → Probability
② Spherical Clusters	② Elliptical Clusters
③ Equal sized Clusters	③ Size can vary
④ Use Centroids	④ Use Mean and Variance