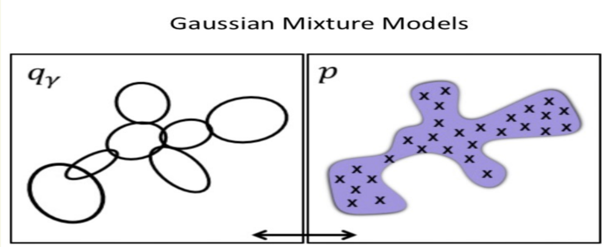
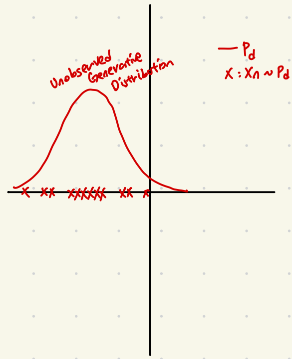




Reminder

Gaussian Distribution : $P(x;\mu,\sigma) = \mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Assume we have N i.i.d samples from Gaussian distribution, $\{X_n \sim P_d\}_{n=1}^{N}$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \mu)^2$$

1. Maximum Likelihood : ★ likelihood: Measure of how well a statistical model explains the observed data.

$$\arg\max_{\mu,\sigma} p(x_1, x_2, \dots, x_N; \mu, \sigma) = \prod_{n=1}^{N} p(x_n;\mu,\sigma)$$

2. Maximum log-likelihood:

$$\arg\max_{\mu,\sigma} \log\left(\prod_{n=1}^{N} p(x_1, x_2, \dots, x_N)\right) = \sum_{n=1}^{N} \log(p(x_n;\mu,\sigma))$$

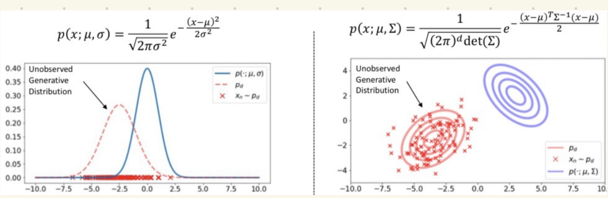
3. Minimizing Negative Log likelihood:

$$\arg\min_{\mu,\sigma} \underbrace{\sum_{n=1}^{N} \frac{\log(2\pi\sigma^2)}{2} + \frac{(x_n-\mu)^2}{2\sigma^2}}_{L}$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \mu^* = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow \sigma^* = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu^*)^2$$

Multivariate Gaussian



$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \mu^* = \frac{1}{N}\sum_{n=1}^{N} x_n$$

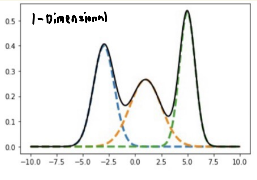
$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow \sigma^* = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu^*)^2$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \mu^* = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow \Sigma^* = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu^*)(x_n - \mu^*)^T$$

★ Maximizing log-likelihood is equivalent to minimizing KL divergence

Mixture of Gaussian



$$p(x;[(a_k,\mu_k,\sigma_k)]_{k=1}^{K}) = \sum_{k=1}^{K} a_k N(x;\mu_k,\sigma_k) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

where $a_k \geq 0$, $\sum_{k=1}^{K} a_k = 1$, and $\sigma_k \geq 0$.

$\mu \in \mathbb{R}^{K\times dim}$, $\mu_k \in \mathbb{R}^{1\times dim}$

$\sigma \in \mathbb{R}^{K\times dim}$, $\sigma_k \in \mathbb{R}^{1\times dim}$

$N(x_n;\mu_k,\sigma_k) \in \mathbb{R}^{N\times dim}$

Log-likelihood : $\log\left(\prod_{n=1}^{N}\left(\sum_{k=1}^{K} a_k N(x_n;\mu_k,\sigma_k)\right)\right) = \sum_{n=1}^{N}\log\left(\sum_{k=1}^{K} a_k N(x_n;\mu_k,\sigma_k)\right)$

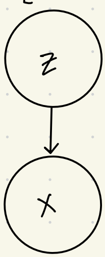
log(sum(·)) doesn't have a closed form solution

↳ How do we deal with this? : Expectation Maximization

Expectation Maximization - Origins

indicator of Gaussian

$Z = [z_1, \dots, z_K]$



$p(Z) = a_k$

$P(X|Z)$

$P(z_k=1) = a_k$

$P(x|z_k=1) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$

$$P(x) = \sum_{k=1}^{K} P(x|z_k=1)\underbrace{P(z_k=1)}_{a_k} = \underbrace{\sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}}_{a_k \cdot N(x_n;\mu_k,\sigma_k)}$$

ex) $0.3\, N(x_n;\mu_1,\sigma_1) + 0.7\cdot N(x_n;\mu_2,\sigma_2)$

Since log-likelihood of GMM doesn't have a closed-form solution, we need gradient of likelihood (for optimizer, such as Gradient Descent)

$\frac{d}{d\theta}\log p(x) = \frac{d}{d\theta}\log\left(\sum_z p(z,x)\right)$

where $\theta : \{\mu_1,\dots,\mu_K, \sigma_1,\dots,\sigma_K\}$

$$\dots = \sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)p(z)) = \sum_z p(z|x)\frac{d}{d\theta}\underbrace{\log(p(x|z))}_{\text{we can optimize } \mu,\sigma} + \sum_z p(z|x)\frac{d}{d\theta}\underbrace{\log(p(z))}_{\text{we can optimize } a_k}$$

And, $p(z|x) = \frac{p(x|z)p(z)}{\sum_{z'} p(x|z')p(z')}$: Is soft assignment of data x to each Gaussian

Expectation Maximization - Algorithm

- Expectation step : for fixed parameters $[(a_k,\mu_k,\sigma_k)]_{k=1}^{K}$ compute $r_n^k = p(z_k=1|x_n)$ for each sample

$$r_n^k = \frac{a_k N(x_n;\mu_k,\sigma_k)}{\sum_i a_i N(x_n;\mu_i,\sigma_i)}$$

- Maximization step : for fixed r_n^k solve the maximum log-likelihood to obtain optimal parameters :
  - Mixture Coefficients : $a_k = \frac{N_k}{N}$ for $N_k = \sum_n r_n^k$
  - Means : $\mu_k = \frac{1}{N_k}\sum_n r_n^k x_n$
  - Variances : $\sigma_k^2 = \frac{1}{N_k}\sum_n r_n^k (x_n-\mu_k)^2$
  - Covariances : $\Sigma_k = \frac{1}{N_k}\sum_n \overset{2100\times 1}{r_n^k}\underbrace{\overset{2\times 2100}{(x_n-\mu_k)}}_{2100\times 2}\underbrace{\overset{2100\times 2}{(x_n-\mu_k)^T}}_{2\times 2100}$: 2100×2

EM-GMM Algorithm Pseudocode (Input: X, K=3)

- Initialize K Gaussian distributions and their mixture coefficients

  $p(x|z_k=1) = \mathcal{N}(\mu_k, \Sigma_k)$

  $p(z_k=1) = a_k = \frac{1}{K}$

- Iterate until convergence

  1. For fixed GMM parameters, Calculate Soft Assignments

     ★ Soft Assignment : $p(z|x) = \dfrac{p(x|z)\,p(z)}{\sum_{z'} p(x|z')\,p(z')}$

     $$p(z_k=1) = \frac{p(x|z_k=1)\,p(z_k=1)}{\sum_{i=1}^{K} p(x|z_i=1)\,p(z_i=1)}$$

  2. For fixed soft assignments, calculate GMM parameters

     - $\mu_k = \dfrac{\sum_{n=1}^{N} p(z_k=1|x_n)\,x_n}{\sum_{n=1}^{N} p(z_k=1|x_n)}$

     - $\Sigma_k = \dfrac{\sum_{n=1}^{N} p(z_k=1|x_n)\,(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{N} p(z_k=1|x_n)}$

     - $a_k = \dfrac{\sum_{n=1}^{N} p(z_k=1|x_n)}{N}$