Gaussian Mixture Model

Goal : Approximate the distribution using the data

<u>Gaussian mixture distribution</u>

The distribution has k clusters, cluster j has probability π_j.

$$\text{Prob}\{x \in \text{cluster } j\} = \pi_j , \quad j = 1, 2, \cdots, k, \text{ where } \sum_{j=1}^{k} \pi_j = 1$$

Define one-hot vectors, $r_j \in \{0,1\}^k$, where $j = 1, 2, \cdots, k$ to identify the clusters : $(r_j)_I = \delta_{jI}$, $I = 1, 2, \cdots, k$

$x \in$ cluster $j \iff r = r_j$, $j = 1, 2, \cdots, k$

Within each cluster, we assume the distribution is Gaussian $p(x \mid r = r_j) = P_g(x; z_j, \Sigma_j)$

where $P_g(x; z_j; \Sigma_j) = (2\pi)^{-d/2} |\Sigma_j|^{-\frac{1}{2}} e^{-\frac{1}{2}(x - z_j)^T \Sigma_j^{-1} (x - z_j)}$

- Law of total probability : $p(x) = \sum_{j=1}^{k} p(x \mid r = r_j)\, p(r_j) = \sum_{j=1}^{k} \pi_j P_g(x; z_j, \Sigma_j)$

parameter estimation (observed data $\{x_i\}_{i=1}^{N} \longrightarrow \{\pi_j, z_j, \Sigma_j\}_{j=1}^{k}$

Log-likelihood : $I(\{\pi_j, z_j, \Sigma_j\}) = \log p(x_1, \cdots, x_N)$

$$= \log \prod_{i=1}^{N} p(x_i) = \sum_{i=1}^{N} \log \Big[\sum_{j=1}^{k} \pi_j P_g(x_i; z_j, \Sigma_j) \Big]$$

$$\ell = \sum_{i=1}^{N} \log \Big[\sum_{j=1}^{k} \pi_j P_g(x_i; z_j, \Sigma_j) \Big]$$

$$0 = \nabla_{z_j} \ell = - \sum_{i=1}^{N} \left[\frac{\pi_j P_g(\underline{x_{ij}}, z_j, \Sigma_j)}{\sum_{e=1}^{k} \pi_e P_g(x_i; z_e, \Sigma_e)} \right] \cdot \Sigma_j^{-1} (\underline{x_i} - \underline{z_j})$$

$$\text{where } r_{ij} = \frac{\overset{\text{prior}: p(r=r_j)}{\pi_j} \cdot \overset{p(x_i \mid r = r_j) \text{ conditional}}{P_g(\underline{x_{ij}}, z_j, \Sigma_j)}}{\underbrace{\sum_{e=1}^{k} \pi_e P_g(x_i; z_e, \Sigma_e)}_{\sum_{e=1}^{k} p(r,\, x_i) = \underbrace{p(x_i)}_{\text{marginal}}}} = p(\underbrace{r = r_j}_{\text{posterior}} \mid \underline{x_i})$$

r_{ij} : Probability that $x_i \in$ cluster j after observing the data point.

$\left(\sum_{j=1}^{k} r_{ij} = 1 \right)$

Cont. $0 = \nabla_{z_j} l = -\sum_{i=1}^{N} r_{ij} \cdot \Sigma_j^{-1} (\underline{x_i} - \underline{z_j})$

$0 = \Sigma_j^{-1} \sum_{i=1}^{N} r_{ij} (\underline{x}_i - \underline{z_j})$

$0 = \sum_{i=1}^{N} r_{ij} \underline{x}_i - \left(\sum_{i=1}^{N} r_{ij} \right) \underline{z_j}$

$\Rightarrow \underline{z_j} = \dfrac{\sum^{N} r_{ij} x_i}{\sum_{i=1}^{N} r_{ij}}$ (weighted avearge of all samples)

↑ (effective # of sample in j-th cluster N_j)

Maximization conditions:

- Setting the derivative of l w.r.t. $\mathbf{z}_j$ to zero gives

$$\mathbf{z}_j = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_{ij} \mathbf{x}_i, \quad j = 1, \cdots, K$$

where

$$\gamma_{ij} = \frac{\pi_j p_g(\mathbf{x}_i; \mathbf{z}_j, \Sigma_j)}{\sum_{l=1}^{K} \pi_l p_g(\mathbf{x}_i; \mathbf{z}_l, \Sigma_l)} = p(\mathbf{r}_j \,|\, \mathbf{x}_i),$$

$$N_j = \sum_{i=1}^{N} \gamma_{ij}.$$

$$\Sigma_j = \frac{1}{N_j} \sum_{i=1}^{N} r_{ij} (x_i - z_j)(x_i - z_j)^T, \quad j = 1, \cdots, K$$

- $\begin{cases} \max_{\{\pi_j\}} l = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{k} \pi_j P_g(x_i; z_j, \Sigma_j) \right] \\ \text{s.t. } \sum_{j=1}^{k} \pi_j = 1 \end{cases}$

$L = l + \lambda \left(\sum_{j=1}^{k} \pi_j - 1 \right)$ λ: Lagrange multiplier

$$0 = \frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{N} \left[\frac{P_g(x_{ij}, z_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l P_g(x_i; z_l, \Sigma_l)} \right] + \lambda$$

ratio = 1

Multiply both sides by π_j;

$$\left[\sum_{j=1}^{k} \pi_j \right] \lambda = -\sum_{i=1}^{N} \frac{\sum_{j=1}^{k} \pi_j P_g(\underline{x}_i; z_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l P_g(\underline{x}_i; z_l; \Sigma_l)} = -\sum_{i=1}^{N} 1 = -N$$

$\Rightarrow \lambda = -N$

$$0 = \frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{N} \underbrace{\left[\frac{\pi_j P_g(x_{ij}, z_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l P_g(x_i; z_l, \Sigma_l)} \right]}_{r_{ij}} - N \pi_j$$

$$0 = \sum_{i=1}^{N} r_{ij} - N \pi_j \Rightarrow \pi_j = \frac{\sum_{i=1}^{N} r_{ij}}{N} = \frac{N_j}{N}, \quad j = 1, 2, \cdots, k$$

Summary

$$\begin{cases} z_j = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_{ij} x_i & (1) \\ \Sigma_j = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_{ij} (x_i - z_j)(x_i - z_j)^T & (2) \\ \pi_j = \frac{N_j}{N} & (3) \end{cases}$$

where

$$\begin{cases} \gamma_{ij} = \frac{\pi_j p_g(x_i; z_j, \Sigma_j)}{\sum_{l=1}^{K} \pi_l p_g(x_i; z_l, \Sigma_l)} = p(r_j | x_i) & (4) \\ N_j = \sum_{i=1}^{N} \gamma_{ij} & (5) \end{cases}$$

Algorithm:

- Given the dataset $\{\mathbf{x}_i\}_{i=1}^{N}$ in R^d.
- Initialize $\{\pi_j, \mathbf{z}_j, \Sigma_j\}_{j=1}^{K}$:

$$\pi_j = \frac{1}{K}, \quad \mathbf{z}_j \in R^d, \quad \Sigma_j \in R^{d \times d}, \quad j = 1, \cdots, K$$

- Repeat the following two steps until convergence:
  i) Update $\{\gamma_{ij}, N_j\}$ according to Eqs. (4), (5).
  ii) Update $\{\pi_j, \mathbf{z}_j, \Sigma_j\}$ according to Eqs. (1) – (3).

Clustering:

- Given a data point $\mathbf{x}$, compute the posterior probability $p(\mathbf{r} = \mathbf{r}_j \,|\, \mathbf{x})$ - the probability that the data belongs to cluster j after observing the sample.
- Bayes rule:

$$p(\mathbf{r} = \mathbf{r}_j \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, \mathbf{r} = \mathbf{r}_j) p(\mathbf{r} = \mathbf{r}_j)}{p(\mathbf{x})}$$

where

$$p(\mathbf{x} \,|\, \mathbf{r} = \mathbf{r}_j) = p_g(\mathbf{x}; \mathbf{z}_j, \Sigma_j) : \text{class conditional}$$
$$p(\mathbf{r} = \mathbf{r}_j) = \pi_j : \text{prior}$$
$$p(\mathbf{x}) = \sum_{l=1}^{K} \pi_l p_g(\mathbf{x}; \mathbf{z}_l, \Sigma_l) : \text{margin}$$

$\{\pi_j, z_j, \Sigma_j\} \rightarrow$ compute posterior probability $p(r = r_j | x)$

Assign x to the cluster with the largest probability

$k = \text{argmax}_j \, p(r = r_j | x)$