EE5601 - Representation learning

HW1

Name:- Bandi Adarsh
Roll:- CE17btech11010

1) Expression for optimal decorrelating linear transform

Our set of observations $\longrightarrow X \in R^{d \times N}$

and each row in the dataset has zero mean

→ This is nothing but our principal component analysis (PCA)

→ Let us consider $Y = PX$ — ①

where; P is the applied "linear transformation" such that 'Y' is a diagonal matrix.

W.K.T, $C_{XX} = \frac{1}{N} XX^T$

⇒ Now, $C_{YY} = \frac{1}{N}(YY^T) = \frac{1}{N}(PX)(PX^T)$

$$= P\left(\frac{XX^T}{N}\right)P^T$$

⇒ $C_{YY} = P C_{XX} P^T$ where; C_{XX} is a symmetric matrix

⇒ $C_{XX} = E(D)(E^T)$

where; $E \longrightarrow$ eigen vector matrix ⇒ $E = \begin{bmatrix} | & | & & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & & | \end{bmatrix}$

and $D = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$ → diagonal elements are eigenvalues of matrix

Also; $EE^T = I$ — ③

Now, from eqⁿ ② ⇒ $C_{YY} = P(E \cdot D \cdot E^T) P^T$

$$\Rightarrow C_{YY} = (PE)\, D\, (PE)^T \quad - ④$$

Suppose let $P = E^T$

$$\Rightarrow C_{YY} = (E^T \cdot E)\, D\, (E^T \cdot E)^T$$

$$= (I)(D)(I) = D \quad (\text{from eq}^n\ ③)$$

$$\Rightarrow C_{YY} = D \longrightarrow \text{a diagonal matrix} \qquad D = \begin{bmatrix} \lambda_1 & 0 & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

So, as C_{YY} equals to a diagonal matrix. So, our data after transformation as been optimally decorrelated. So, our assumption of "$P = E^T$" is true.

∴ For optimal decorrelating, our linear transform i's matrix $\boxed{P = E^T}$ → where; E is eigen vector matrix of X

(2) Expressions for partial derivatives of log likelihood function of GMM

The GMM distribution : $P(\underline{x}) = \sum_{k=1}^{K} w_k \, \mathcal{N}(\underline{x}, \mu_k, \Sigma_k)$

where; $w_k \rightarrow$ mixture weights and $\sum_{k=1}^{K} w_k = 1$

let $\underline{z}$ be our latent variable. let the joint distribution of GMM be $P(\underline{x}, \underline{z})$

Now, $P(\underline{x}, \underline{z}) = \sum_{z} P(\underline{x}, \underline{z}) = \sum_{\underline{z}} P(\underline{z}) \cdot P(\underline{x}/\underline{z})$

If our $\underline{z} \rightarrow$ hot vector $\Rightarrow P(\underline{x}/\underline{z}_k) = \prod_{k=1}^{K} \mathcal{N}(\mu_k, x_*, \Sigma_k)^{z_k}$ — (1)

The likelihood function $\rightarrow L(\underline{x}, \underline{\theta}) = \prod_{i=1}^{N} P(\underline{x}, w_k, \mu_k, \Sigma_k)$

where; Parameter $(\underline{\theta}) = \{\mu_k, w_k, \Sigma_k\}$

$\Rightarrow \log(L(x, \theta)) = \sum_{i=1}^{N} \log(P(\underline{x}, \mu; w, \Sigma))$

Substitute $P(\underline{x}) = \sum_{j=1}^{K} w_j \cdot \mathcal{N}(\underline{x}, \mu_j, \Sigma_j)$

$\Rightarrow \log(L(x, \theta)) = \sum_{i=1}^{N} \left(\sum_{j=1}^{K} w_j \cdot \mathcal{N}(\underline{x}, \mu_j, \Sigma_j) \right)$

Now partial differentiate w.r.t each parameter

$\Rightarrow \dfrac{\partial}{\partial w_j}(\log L) = \sum_{i=1}^{N} \left(\dfrac{\mathcal{N}(\underline{x}, \mu_j, \Sigma_j)}{\sum_{j=1}^{K} w_j \cdot \mathcal{N}(x, \mu_j, \Sigma_j)} \right)$

$$\Rightarrow \frac{\partial}{\partial \mu_j}(\log L) = \sum_{i=1}^{N} \left(\frac{\omega_j \cdot \frac{\partial}{\partial \mu_j} \mathcal{N}(x_i, \mu_j, \Sigma_j)}{\sum_{j=1}^{k} \omega_j \cdot \mathcal{N}(x_i, \mu_j, \Sigma_j)} \right)$$

and

$$\Rightarrow \frac{\partial}{\partial \sigma_j}(\log L) = \sum_{i=1}^{N} \left(\frac{\omega_j \cdot \frac{\partial}{\partial \sigma_j} \mathcal{N}(x_i, \mu_j, \Sigma_j)}{\sum_{j=1}^{k} \omega_j \cdot \mathcal{N}(x_i, \mu_j, \Sigma_j)} \right)$$

The multivariate Gaussian distribution

$$\hookrightarrow \mathcal{N}(x_i, \mu_j, \Sigma_j) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where; $|\Sigma|$ = det of covariance matrix

$$\Rightarrow \frac{\partial}{\partial \mu_i}(\mathcal{N}) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \cdot \left(2(x-\mu)^T \Sigma^{-1}\right)$$

Consider, $P(z_k = 1/x_i) \rightarrow$ prob that x_i comes from k^{th} gaussian

$P(x_i/z_k = 1) \rightarrow$ probability that x_i comes from $\mathcal{N}(\mu_k, \Sigma_k$

$$\Rightarrow = \sqrt{2\pi}\left(-\frac{1}{2}|\Sigma_k|^{-1/2} \Sigma_k^{-1}\right) \mathcal{N}(x_i, \mu_k, \Sigma_k)$$

$$+ \left[|2\pi\Sigma_k|^{-1/2} \mathcal{N}(x, \mu_k, \Sigma_k) \frac{1}{2} \Sigma^{-1}(x_i-\mu)(x_i-\mu)^T \Sigma^{-1}\right)$$

$$= -\frac{1}{2}\left(\Sigma_k - \Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)\Sigma_k^{-1}\right) \mathcal{N}(x_i, \mu_k, \Sigma_k)$$

Now, $\frac{d}{d\Sigma_k}(\log L) = \sum_{i=1}^{N} \gamma(z_k^i)\left(\Sigma_k^{-1} - \Sigma_k^{-1}(x_i - \mu_k)(x_i-\mu_k)^T \Sigma_k^{-1}\right) = 0$

$$\Rightarrow \sum_{i=1}^{N} \gamma(z_k^i) = \sum_{i=1}^{N} \gamma(z_k^i)\left((x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1}\right)$$

$$\Rightarrow \boxed{\Sigma_k = \sum_{i=1}^{N} \frac{\gamma(z_k^i)(x_i - \mu_k)(x_i - \mu_k)^T}{N_k}}$$

where; $N_k = \sum_{i=1}^{N} \gamma(z_k^i)$

If we solve for $\frac{d}{d\omega_k}(\log L) = 0$ and $\sum^{K} \omega_k = 1$

$$\Rightarrow \sum_{i=1}^{N}\left(\frac{N(x_i, \mu_k, \Sigma_k)}{\sum \omega_k N(x_i, \mu_k, \Sigma_k)}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\gamma(z_k^i)}{\omega_k} = 0 \quad \text{and} \quad \sum_{k=1}^{K} \omega_k = 1$$

By using Lagrange multipliers:

$$\ln(P(X \mid \mu, \Sigma, \omega)) + \lambda\left(\sum_{k=1}^{K} \pi_{k} - 1\right) = 0$$

$$\Rightarrow \sum \frac{N(x_i, \mu_k, \Sigma_k)}{\sum \omega_j N(x_i, \mu_j, \Sigma_k)} + \lambda = 0$$

$\Rightarrow$ Multiply by ω_k,

$$\lambda \omega_k + \sum_{i=1}^{N} \frac{\omega_k N(x_i, \mu_k, \Sigma_k)}{\sum \omega_j N(x, \mu, \Sigma)} = 0$$

Summing over 'k' on both sides

$$\Rightarrow \lambda \sum_{k=1}^{K} w_k + \sum_{i=1}^{N} \frac{\sum_{k=1}^{K} w_k \cdot \mathcal{N}}{\sum_{k=1}^{K} w_k \cdot \mathcal{N}(x_i, \mu_k, \Sigma_k)} = 0$$

$$\Rightarrow \lambda(1) + N = 0 \Rightarrow \lambda = -N$$

So,
$$\sum_{i=1}^{N} \frac{\mathcal{N}(x_i, \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \cdot \mathcal{N}(x_i, \mu_j, \Sigma_j)} - N = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\gamma(z_k^i)}{w_k} = N \Rightarrow w_k = \sum_{i=1}^{N} \frac{\gamma(z_k^i)}{N}$$

$$\Rightarrow \boxed{w_k = \frac{N_k}{N}}$$

So, our parameters (local optimality) are :

$$\boxed{\mu_k = \frac{\sum_{i=1}^{N} \gamma(z_{ki})\, x_i}{\sum_{i=1}^{N} \gamma(z_{ki})}}$$

$$\boxed{\Sigma_k = \sum_{i=1}^{N} \left(\frac{\gamma(z_{ki})\, (x_i - \mu_k)(x_i - \mu_k)^T}{N_k} \right)}$$

$$\boxed{w_k = \frac{N_k}{N}}$$