Homework 1

Subject: Representative Learning

Submitted by: Sayan Chakraborty

Roll Number: EE18MTECH11030

Ans 1:

To derive the expression for the optimal decorrelating linear transform for a set of observations $X \in \mathbb{R}^{d \times N}$ where each row is assumed to be zero-mean.

To obtain the optimal decorrelating linear transform for $X \in \mathbb{R}^{d \times N}$ we use the idea of principal component analysis (PCA), since PCA has the distinction of being the optimal orthogonal transformation for keeping the subspace that has the largest variance.

Given, $X \in \mathbb{R}^{d \times N}$ as the set of observations with each row having mean zero.

Let the optimal decorrelating linear transform be

$$Y = PX \quad \text{—} (1)$$

The idea is to find the matrix P such that the covariance matrix of Y is diagonal, i.e, Y is decorrelated.

Let the covariance matrix of X be

$C_X = \frac{1}{N} X X^T$

The eigenvalue decomposition of C_{XX} is EDE^T where E is the matrix of eigenvectors where the eigenvectors are orthonormal, and D is a diagonal matrix with eigenvalues of C_{XX} in its diagonal entries.

Now,

$$C_{YY} = \frac{1}{N} Y Y^T \text{ is the covariance matrix of } Y$$

$$\Rightarrow C_{YY} = \frac{1}{N} PX(PX)^T$$

$$\Rightarrow C_{YY} = \frac{1}{N} PXX^T P^T$$

$$\Rightarrow C_{YY} = P C_{XX} P^T$$

$$\Rightarrow C_{YY} = P E D E^T P^T$$

Select $P = E^T$

then,

$$C_{YY} = E^T E D E^T E$$

$$\Rightarrow C_{YY} = D \quad (\because E^T E = I)$$

$\therefore$ by selecting the transformation matrix $P = E^T$ we see that C_{YY} can be diagonalized and hence Y is decorrelated.

∴ the decorrelating linear transformation as stated in ① can be achieved by selecting $P = E^T$.

Ans 2:

To derive partial derivatives of the log-likelihood functions of a Gaussian Mixture Model (GMM) with respect to each parameters. Also, find the locally optimal parameters in terms of the posterior probabilities and observations.

The GMM density is given by:

$$f(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x; \mu_k, \Sigma_k), \quad \text{—①}$$

$$\sum_{k=1}^{K} \pi_k = 1, \quad 0 \leq \pi_k \leq 1$$

The log-likelihood function of ① is given as:

$$\ell(x;\theta) = \sum_{n=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_n; \mu_k, \Sigma_k)\right] \quad \text{—②}$$

Due to log of the summation term in ②, it is difficult to obtain the parameter estimates. To ease the computation, we introduce latent random variables described as follows:

define, $\underline{z} = [0, \ldots, 1, \ldots 0]^T$ as one hot vector
$\underline{z}$ is zero except at the k^{th} location,
where, $\underline{z}$ is the latent random variable.

Define the prior probability as:

$$p(\underline{z}_k = 1) = \pi_k \quad \text{--- (3)}$$

then,
$$p(\underline{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \quad \text{--- (4)}$$

Now,
$$p(\underline{x} \mid \underline{z}_k = 1) = \mathcal{N}(\underline{x}; \mu_k, \Sigma_k) \quad \text{--- (5)}$$

then,
$$p(\underline{x} \mid \underline{z}) = \prod_{k=1}^{K} \mathcal{N}(\underline{x}; \mu_k, \Sigma_k)^{z_k} \quad \text{--- (6)}$$

Now,
we express $p(\underline{x})$ as the marginal distribution of $p(\underline{x}, \underline{z})$ as:

$$p(\underline{x}) = \sum_{\underline{z}} p(\underline{x}, \underline{z})$$
$$= \sum_{\underline{z}} p(\underline{z}) \cdot p(\underline{x} \mid \underline{z}) \quad \text{--- (7)}$$

Substituting (4) and (6) in (7), we get

$$p(\underline{x}) = \sum_{k=1}^{K} \left[\prod_{i=1}^{K} \pi_i^{z_i} \prod_{j=1}^{K} \mathcal{N}(\underline{x}; \mu_j, \Sigma_j)^{z_j} \right]$$
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\underline{x}; \mu_k, \Sigma_k)$$

Define the posteriori probability as:

$$p(z_K=1 \mid X) = \frac{p(X \mid z_K=1)\; p(z_K=1)}{p(X)}$$

$$= \frac{\pi_K \, N(X;\, \mu_K, \Sigma_K)}{\sum_{k=1}^{K} \pi_K \, N(X;\, \mu_K, \Sigma_K)}$$

$$= \gamma(z_K) \quad \text{——} \quad (8)$$

Similarly,

$$\gamma(z_{nK}) = p(z_K=1 \mid x_n) \quad \text{——} \quad (9)$$

We will use (8) to obtain locally optimal parameter estimates of (1)

(i) Find the partial derivatives

For μ_K

$$\frac{\partial \ell}{\partial \mu_K}$$

$$= \frac{\partial}{\partial \mu_K} \left[\sum_{n=1}^{N} \log \left[\sum_{K=1}^{K} \pi_K \, N(X_n;\, \mu_K, \Sigma_K) \right] \right]$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{K=1}^{K} \pi_K \, N(X_n;\, \mu_K, \Sigma_K)} \; \frac{\partial}{\partial \mu_K} \sum_{K=1}^{K} \pi_K \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x_n - \mu_K)^T \Sigma^{-1} (x_n - \mu_K)\right)$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}[x_n; \mu_k, \Sigma_k)]} \frac{\partial}{\partial \mu_k} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k)\right)$$

$$= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n, \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}[x_n; \mu_k, \Sigma_k)]} \frac{\partial}{\partial \mu_k} \left(-\frac{1}{2} (x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k)\right)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \mu_k} \left(-\frac{1}{2} (x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k)\right) \quad \left(\text{from } 8, 9\right)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \Sigma^{-1} (x_n - \mu_k)$$

$$= \Sigma^{-1} \left[\sum_{n=1}^{N} \gamma(z_{nk}) x_n - \sum_{n=1}^{N} \gamma(z_{nk}) \mu_k \right] \quad (10)$$

For π_k

In order to maximize (2) with respect to π_k we make use the fact that $\sum_{k=1}^{K} \pi_k = 1$. Now, using Lagrange multiplier λ we obtain

$$g = \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}[x_n; \mu_k, \Sigma_k) \right] + \lambda \left[\sum_{k=1}^{K} \pi_k - 1 \right] \quad (11)$$

Note that $\sum_{k=1}^{K} \pi_k = 1$ is used as a constraint.

Now,

$$\frac{\partial}{\partial \pi_k}\left\{\sum_{n=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)\right] + \lambda\left[\sum_{k=1}^{K}\pi_k - 1\right]\right\}$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)} \mathcal{N}(x_n; \mu_k, \Sigma_k) + \lambda \quad \text{--- (12)}$$

for Σ_k

$$\frac{\partial l}{\partial \Sigma_k}$$

$$= \frac{\partial}{\partial \Sigma_k}\left(\sum_{n=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)\right]\right)$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)} \frac{\partial}{\partial \Sigma_k}\left(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)\right)\right) \quad \text{--- (13)}$$

Now,

$$\frac{\partial}{\partial \Sigma_k}\left(\frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)\right)\right)$$

$$= -\frac{1}{2} \frac{2\pi|\Sigma_k|\Sigma_k^{-1}}{2\pi|\Sigma_k|\sqrt{2\pi|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)\right)$$

$$+ \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp\left(-\tfrac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)\right)\left(\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1}\right.$$

$$= -\frac{1}{2} \frac{2\pi |\Sigma_k| \Sigma_k^{-1} \exp\left(\frac{-1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right)}{2\pi |\Sigma_k| \sqrt{2\pi |\Sigma_k|}}$$

$$+ \frac{1}{2\sqrt{2\pi |\Sigma_k|}} \exp\left(-\tfrac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right) \left(\Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right)$$

$$= -\frac{1}{2} \frac{1}{\sqrt{2\pi |\Sigma_k|}} \exp\left(\frac{-1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right) . \left[\Sigma_k^{-1} - \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right]$$

$$= -\tfrac{1}{2} \mathcal{N}(x_n ; \mu_k, \Sigma_k) . \left[\Sigma_k^{-1} - \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right] \quad \text{--- (14)}$$

Substituting (14) in (13)

(13) $\Rightarrow$

$$\sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n ; \mu_k, \Sigma_k)} \pi_k \frac{\partial}{\partial \Sigma_k} \left(\sum_{k=1}^{K} \frac{1}{\sqrt{2\pi |\Sigma_k|}} \exp\left(\frac{-1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right)\right)$$

$$= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n ; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n ; \mu_k, \Sigma_k)} \left[\Sigma_k^{-1} - \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right]$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left[\Sigma_k^{-1} - \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right] \quad \text{--- (15)}$$

(89) Finding the locally optimal parameters

For μ_k

(10) $= 0$

$$\Rightarrow \Sigma^{-1}\left[\sum_{n=1}^{N} \gamma(z_{nk})\, \underline{x}_n - \sum_{n=1}^{N} \gamma(z_{nk})\, \mu_k\right] = 0$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk})\, \underline{x}_n - \sum_{n=1}^{N} \gamma(z_{nk})\, \mu_k = 0 \quad \left(\text{premultiplying by } \Sigma\right)$$

$$\Rightarrow \mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\, \underline{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\Rightarrow \boxed{\hat{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, \underline{x}_n},$$

where, $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

For π_k

(12) $= 0$

$$\Rightarrow \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)]} \mathcal{N}(x_n; \mu_k, \Sigma_k) + \lambda = 0$$

Multiplying both sides by π_k and summing over k, we have,

$$\Rightarrow \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)} + \sum_{k=1}^{K} \pi_k \lambda) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)} + \sum_{k=1}^{K} \pi_k \lambda) = 0$$

$$\Rightarrow \sum_{n=1}^{N} 1 + \lambda = 0$$

$$\Rightarrow \lambda = -N$$

Then, we have that,

$$\Rightarrow \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)]} \mathcal{N}(x_n; \mu_k, \Sigma_k) + \lambda = 0$$

Multiplying both sides by π_k

$$\Rightarrow \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)} - N\pi_k = 0$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) - N\pi_k = 0$$

$$\Rightarrow \boxed{\hat{\pi}_k = \frac{N_k}{N}}, \quad \text{where } N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

For Σ_K

$$(15) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) \left[\Sigma_k^{-1} - \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right] = 0$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) \left[\Sigma_k \Sigma_k^{-1} - \Sigma_k \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right] = 0$$

(premultiplying by Σ_K^{-1})

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) \left[I - I (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right] = 0$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) \left[\Sigma_K - (x_n - \mu_k)(x_n - \mu_k)^T I \right) = 0$$

(post multiplying by Σ_k)

$$\Rightarrow \Sigma_K = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\Rightarrow \boxed{\hat{\Sigma}_K = \frac{1}{N_K} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T},$$

where, $N_K = \sum_{n=1}^{N} \gamma(z_{nk})$

Ans: 3

The code for this is:

EE18MTECH11030_Question3_HW1.ipynb

Comments about the code:

There are two ways for feeding the data:

(i) You can use your own .xlsx file to feed data.

(ii) The code generates data (see section Generate data) and saves it in Data.xlsx file.

If you use your own data please comment the section 2. Generate data. You will then be asked to input your filename and mixture number (K).

If you want to use the data generated by the code, please use the excel filename as Data.xlsx.

Note that the data is randomly generated by the code by randomly selecting $\underline{\mu}$, $\underline{\pi}_i$, Σ. The number of observations N is hard coded as 10,000, K is hardcoded as 3, d varies between 2 to 3.

K = number of mixture, d = dimension of data

If required this parameters can be taken as user inputs to generate data.

Observations

(i) By varying μ, Σ, p_i it was observed that if μs for all mixtures were close the code performs better if K (number of mixtures) is selected as 1.

(ii) However, if μs for all mixtures vary largely, selecting K for estimation to be equal to the original number of mixtures the code performed well.

(iii) It is therefore necessary to visualize the original data first before selecting K for estimation.

(iv) Overall, the code performed well for randomly generated data.

Also note that sections 4.2 and 4.3 cannot be used if you use your own data.