Mixture Model

(Expectation maximization Algorithm)

Pradip Das, ISI Kolkata
pradipdas_2021@iitkalumni.org

Log-likelihood function

z : Latent space variables
x : Observed variables
Then the log-likelihood function is given by

$$\begin{aligned} l(\theta) &= \log(P(x|\theta)) \\ &= \log(\sum_z P(x,z|\theta)) \end{aligned} \tag{1}$$

Let $q(z|x,\theta)$ is an arbitrary density defined over z. From (1)

$$l(\theta) = \log(\sum_z q(z|x,\theta)\frac{P(x,z|\theta)}{q(z|x,\theta)}) \tag{2}$$

As $\log(x)$ is concave function then by Jensen's inequality we have

$$l(\theta) \geq \sum_z q(z|x,\theta)\log(\frac{P(x,z|\theta)}{q(z|x,\theta)}) \tag{3}$$

Instead of directly maximizing $l(\theta)$ we can maximize the lower bound.
Now,

$$\begin{aligned} l(\theta) &\geq \sum_z q(z|x,\theta)\log(\frac{P(x,z|\theta)}{q(z|x,\theta)}) \\ &= \sum_z q(z|x,\theta)\log(P(x,z|\theta)) - \sum_z q(z|x,\theta)\log(q(z|x,\theta)) \\ &= Q(\theta|\theta^{(t)}) + H(q) \end{aligned}$$

Where, $Q(\theta|\theta^{(t)}) = E_{q(z|x,\theta^{(t)})}(\log(P(x,z|\theta))$.

EM Algorithm

E-step:
Compute

$$Q(\theta|\theta^{(t)}) = E_{q(z|x,\theta^{(t)})}(\log(P(x,z|\theta))$$

M-step:

$$\theta^{(t+1)} = \arg\ \max_\theta\ E_{q(z|x,\theta^{(t)})}(\log(P(x,z|\theta))$$

Gaussian Mixture Model (GMM)

A Gaussian mixture is a function that is comprised of several Gaussian each identified by $k \in \{1, 2, .., K\}$ where K is the number of cluster in the data. Each Gaussian in mixture is comprised of following parameter:

1. A mean μ that defines its center.
2. A covariance Σ that defines its width.
3. A mixing probability π that defines how big or small a Gaussian function will be.

Let, π_k be the mixing coefficient of k-th cluster with condition $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$.
Let, $P(z_{nk} = 1|x_n)$ denotes the probability of x_n from Gaussian k then,

$$\pi_k = P(z_k = 1)$$

Hence,

$$\begin{aligned} P(x) &= \sum_{k=1}^{K} P(x|z_k = 1)P(z_k = 1) \\ &= \sum_{k=1}^{K} \pi_k P_k(x) \end{aligned}$$

Let us define $\theta = \{\mu_k, \pi_k, \Sigma_k : k = 1, 2, ..., K\}$
Then,

$$P(x_n|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

Let the dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ are iid drawn from an unknown distribution $P(x)$. Our objective is to find a good approximation of this unknown distribution $P(x)$ by means of a GMM with K mixture components.
So,

$$\begin{aligned} P(\mathcal{X}|\theta) &= \prod_{i=1}^{N} P(x_i|\theta) \\ &= \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right) \end{aligned}$$

Now we have to optimize the cost function $L = \log(P(\mathcal{X}|\theta))$ w.r.t. π_k, μ_k and Σ_k.

Bernoulli Mixture Model

Now we introduce the latent variables for the EM algorithm. Let $x^{(i)} \in \{0, 1\}^D$, $X = \{x^{(i)}\}_{i=1,...n}$. Let $z^{(i)} \in \{0, 1\}^K$ be an indicator vector, such that $z_k^{(i)} = 1$ if $x^{(i)}$ was drawn from Bernoulli(p^k) and 0 otherwise. Let $Z = \{z^{(i)}\}_{i=1,...n}$, $\{p^1, ..., p^k\} = p$ and a distribution

$\pi(k)$, over the selection of which set of Bernoulli parameters p^k is chosen. Then,

$$P(z^{(i)}|\pi) = \prod_{k=1}^{K} \pi(k)^{z_k^{(i)}}$$

And

$$P(x^{(i)}|z^{(i)}, p, \pi) = \prod_{k=1}^{K} P(x^{(i)}|p^k)^{z_k^{(i)}}$$

Therefore,

$$\begin{aligned} P(X, Z|\pi, p) &= \prod_{i=1}^{n} P(x^{(i)}, z^{(i)}|\pi, p) \\ &= \prod_{i=1}^{n} P(z^{(i)}|\pi) \cdot P(x^{(i)}|z^{(i)}, p, \pi) \\ &= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \pi(k)^{z_k^{(i)}} \prod_{k=1}^{K} P(x^{(i)}|p^k)^{z_k^{(i)}} \right] \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi(k) P(x^{(i)}|p^k) \right]^{z_k^{(i)}} \end{aligned} \tag{4}$$

Let,

$$\begin{aligned} \eta(z_k^{(i)}) &= E(z_k^{(i)}|x^{(i)}, p, \pi) \\ &= 1 \cdot P(z_k^{(i)} = 1|x^{(i)}, p, \pi) + 0 \cdot P(z_k^{(i)} = 0|x^{(i)}, p, \pi) \\ &= \frac{P(x^{(i)}|z_k^{(i)} = 1, p, \pi) P(z_k^{(i)} = 1|\pi, p)}{\sum_j P(x^{(i)}|z_j^{(i)} = 1, p, \pi) P(z_j^{(i)} = 1|\pi, p)} \\ &= \frac{\pi(k) \prod_{d=1}^{D} (p_d^{(k)})^{x_d^{(i)}} (1 - p_d^{(k)})^{1-x_d^{(i)}}}{\sum_j \pi(j) \prod_{d=1}^{D} (p_d^{(j)})^{x_d^{(i)}} (1 - p_d^{(j)})^{1-x_d^{(i)}}} \end{aligned}$$

Now from (4) we hove,

$$P(X, Z|\widetilde{\pi}, \widetilde{p}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\widetilde{\pi}(k) P(x^{(i)}|\widetilde{p}^k) \right]^{z_k^{(i)}}$$

Then

$$\begin{aligned} \log P(X, Z|\widetilde{\pi}, \widetilde{p}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_k^{(i)} \left[\log \widetilde{\pi}(k) + \log P(x^{(i)}|\widetilde{p}^k) \right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_k^{(i)} \left[\log \widetilde{\pi}(k) + \log \left(\prod_{d=1}^{D} (\widetilde{p}_d^k)^{x_d^{(i)}} (1 - \widetilde{p}_d^k)^{1-x_d^{(i)}} \right) \right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_k^{(i)} \left[\log \widetilde{\pi}(k) + \sum_{d=1}^{D} x_d^{(i)} \log(\widetilde{p}_d^k) + (1 - x_d^{(i)}) \log(1 - \widetilde{p}_d^k) \right] \end{aligned}$$

Therefore taking expectation on above equation:

$$
\begin{aligned}
E\left(\log P(X,Z|\widetilde{\pi},\widetilde{p})|X,p,\pi\right) &= E\left(\sum_{i=1}^{n}\sum_{k=1}^{K} z_k^{(i)}\left[\log\widetilde{\pi}(k) + \sum_{d=1}^{D} x_d^{(i)}\log(\widetilde{p}_d^k) + (1-x_d^{(i)})\log(1-\widetilde{p}_d^k)\right]\right)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} E(z_k^{(i)})\left[\log\widetilde{\pi}(k) + \sum_{d=1}^{D} x_d^{(i)}\log(\widetilde{p}_d^k) + (1-x_d^{(i)})\log(1-\widetilde{p}_d^k)\right]\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \eta(z_k^{(i)})\left[\log\widetilde{\pi}(k) + \sum_{d=1}^{D} x_d^{(i)}\log(\widetilde{p}_d^k) + (1-x_d^{(i)})\log(1-\widetilde{p}_d^k)\right]
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
Q(\widetilde{\theta}|\theta) &= E\left(\log P(X,Z|\widetilde{\pi},\widetilde{p})|X,p,\pi\right)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \eta(z_k^{(i)})\left[\log\widetilde{\pi}(k) + \sum_{d=1}^{D} x_d^{(i)}\log(\widetilde{p}_d^k) + (1-x_d^{(i)})\log(1-\widetilde{p}_d^k)\right]
\end{aligned}
$$

Then to find the optimal $\widetilde{p}^k$ and $\widetilde{\pi}(k)$ we have to differentiate $Q(\widetilde{\theta}|\theta)$.

$$
\begin{aligned}
\frac{\partial Q(\widetilde{\theta}|\theta)}{\partial \widetilde{p}_d^k} = 0 &\implies \sum_{i=1}^{n}\eta(z_k^{(i)})\left[\frac{x_d^{(i)}}{\widetilde{p}_d^k} - \frac{1-x_d^{(i)}}{1-\widetilde{p}_d^k}\right] = 0\\
&\implies \widetilde{p}_d^k = \frac{\sum_{i=1}^{n}\eta(z_k^{(i)})x_d^{(i)}}{\sum_{i=1}^{n}\eta(z_k^{(i)})}
\end{aligned}
$$

Thus,

$$
\widetilde{p}^k = \frac{\sum_{i=1}^{n}\eta(z_k^{(i)})x^{(i)}}{\sum_{i=1}^{n}\eta(z_k^{(i)})}
$$

Again to maximize E-step w.r.t $\widetilde{\pi}(k)$ we have to maximize $\sum_{i=1}^{n}\sum_{k=1}^{K}\eta(z_k^{(i)})\log\widetilde{\pi}(k)$ subject to $\sum_{k=1}^{K}\widetilde{\pi}(k) = 1$ Let define,

$$
\mathcal{L} = \sum_{i=1}^{n}\sum_{k=1}^{K}\eta(z_k^{(i)})\log\widetilde{\pi}(k) - \lambda\left(\sum_{k=1}^{K}\widetilde{\pi}(k) - 1\right) \tag{5}
$$

Then

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \widetilde{\pi}(k)} = 0 &\implies \sum_{i=1}^{n}\eta(z_k^{(i)})\frac{1}{\widetilde{\pi}(k)} - \lambda = 0\\
&\implies \widetilde{\pi}(k) = \frac{\sum_{i=1}^{n}\eta(z_k^{(i)})}{\lambda}
\end{aligned}
$$

Therefore from (5) we have

$$
\mathcal{L}(\lambda) = \sum_{i=1}^{n}\sum_{k=1}^{K}\eta(z_k^{(i)})\left(\log\left(\sum_{i=1}^{n}\eta(z_k^{(i)})\right) - \log(\lambda)\right) - \left(\sum_{k=1}^{K}\sum_{i=1}^{n}\eta(z_k^{(i)}) - 1\right)
$$

Then

$$\frac{d\mathcal{L}}{d\lambda} = 0 \implies \lambda = \sum_{k=1}^{K} \sum_{i=1}^{n} \eta(z_k^{(i)})$$

Therefore

$$\widetilde{\pi}(k) = \frac{\sum_{i=1}^{n} \eta(z_k^{(i)})}{\sum_{k=1}^{K} \sum_{i=1}^{n} \eta(z_k^{(i)})}$$