1. Expectation Maximization

EM algorithm is useful for the model containing latent variables Z when the maximum likelihood is hard to derive from the observed data Y. We can write the maximum likelihood of Y like following

$$\arg\max_{\theta} \mathcal{L}(Y;\theta) = \arg\max_{\theta} log(p(Y;\theta))$$

The Expectation Maximization rewrites the question as the following

$$\arg\max_{\theta}\ log\int_Z p(Y,Z;\theta)dZ$$

Thus, we can derive the EM with an approximation $q(Z;\gamma)$ for $p(Z|Y)$ to avoid evaluating such complex distribution directly

$$= \arg\max_{\theta}\ log\int_Z \frac{q(Z;\gamma)}{q(Z;\gamma)}p(Y,Z;\theta)dZ$$

$$= \arg\max_{\theta}\ log\,\mathbb{E}_q[\frac{p(Y,Z;\theta)}{q(Z;\gamma)}]$$

Since the log function is concave, $log(\mathbb{E}_p[X]) \geq \mathbb{E}_p[log(X)]$ with Jensen's inequality.

$$\geq \arg\max_{\theta}\ \mathbb{E}_q[log(\frac{p(Y,Z;\theta)}{q(Z;\gamma)})]$$

$$= \arg\max_{\theta}\ \int_Z q(Z;\gamma)log\ p(Y,Z;\theta)dZ - \int_Z q(Z;\gamma)log\ q(Z;\gamma)dZ$$

$$= \arg\max_{\theta}\ \int_Z q(Z;\gamma)log\ p(Y,Z;\theta)dZ - H_q[Z]$$

Where $H_q[Z]$ is the entropy of Z over distribution q

So far, we can express the EM algorithm in a simpler way as

Iterate until θ converge

- E Step

  Evaluate $q(Z;\gamma) = p(Z|Y)$

- M Step

  $\arg\max_{\theta}\ \int_Z q(Z;\gamma)log\ p(Y,Z;\theta)dZ$

---

2. EM In General Form

Actually, we can represent the EM algorithm with variational lower bound $\mathcal{L}(\theta,\gamma)$

$$\mathcal{L}(\theta,\gamma) = \mathbb{E}_q[log(\frac{p(Y,Z;\theta)}{q(Z;\gamma)})]$$

$$= \int_Z q(Z;\gamma) log \, \frac{p(Y,Z;\theta)}{q(Z;\gamma)} dZ$$

$$= -\int_Z q(Z;\gamma) log \, \frac{q(Z;\gamma)}{p(Z|Y)p(Y;\theta)} dZ$$

$$= log \, p(Y;\theta) - \int_Z q(Z;\gamma) \, log \, \frac{q(Z;\gamma)}{p(Z|Y)} dZ$$

$$= log \, p(Y;\theta) - KL[q(Z;\gamma)||p(Z|Y)] \tag{5}$$

Thus

$$\max_\theta \mathcal{L}(Y;\theta) \geq \arg\max_{\theta,\gamma} \mathcal{L}(\theta,\gamma)$$

With KKT, the constrained optimization problem can be solve with Lagrange multiplier

$$\arg\max_{\theta,\gamma} \mathcal{L}(\theta,\gamma) = \arg\max_{\theta,\gamma} log \, p(Y;\theta) - \beta KL[q(Z;\gamma)||p(Z|Y)]$$

Since we've known the KL-divergence is always greater or equal to 0, when $KL[q(Z;\gamma)||p(Z|Y)] = 0$, the result of EM algorithm will be equal to the maximum likelihood $\mathcal{L}(\theta,\gamma) = \mathcal{L}(Y;\theta)$. In the mean time, minimizing the KL-divergence is actually find the best approximation $q(Z;\gamma)$ for $p(Z|Y)$.

Thus, we can also represent the EM algorithm as

Iterate until θ converge

- E Step at k-th iteration

  $\gamma_{k+1} = \arg\max_\gamma \mathcal{L}(\theta_k, \gamma_k)$

- M Step at k-th iteration

  $\theta_{k+1} = \arg\max_\theta \mathcal{L}(\theta_k, \gamma_{k+1})$

---

3. Variational Bayesian Expectation Maximization

In EM, we approximate a posterior $p(Y,Z;\theta)$ without any prior over the parameters θ. Variational Bayesian Expectation Maximization(VBEM) defines a prior $p(\theta;\lambda)$ over the parameters. Thus, VBEM approximates the bayesian model $p(Y,Z,\theta;\lambda) = p(Y,Z|\theta)p(\theta;\lambda)$. Then, we can define a lower bound on the log marginal likelihood

$$log \, p(Y) = log \int_{Z,\theta} p(Y,Z,\theta;\lambda) dZ d\theta$$

$$= log \int_{Z,\theta} q(Z,\theta;\phi^Z,\phi^\theta) \frac{p(Y,Z|\theta)p(\theta;\lambda)}{q(Z,\theta;\phi^Z,\phi^\theta)} dZ d\theta$$

With mean field theory, we factorize q into a joint distribution $q(Z,\theta;\phi^Z,\phi^\theta) = q(Z;\phi^Z)q(\theta;\phi^\theta)$. Thus, the equation can be rewritten as

$$= log \int_{Z,\theta} q(Z;\phi^Z)q(\theta;\phi^\theta)\frac{p(Y,Z|\theta)p(\theta;\lambda)}{q(Z;\phi^Z)q(\theta;\phi^\theta)}dZd\theta$$

$$= log\ \mathbb{E}_{q(Z;\phi^Z)q(\theta;\phi^\theta)}[\frac{p(Y,Z|\theta)p(\theta;\lambda)}{q(Z;\phi^Z)q(\theta;\phi^\theta)}]$$

Since the log function is concave, $log(\mathbb{E}_p[X]) \geq \mathbb{E}_p[log(X)]$ with Jensen's inequality

$$\geq \mathbb{E}_{q(Z;\phi^Z)q(\theta;\phi^\theta)}[log\ \frac{p(Y,Z|\theta)p(\theta;\lambda)}{q(Z;\phi^Z)q(\theta;\phi^\theta)}]$$

Thus, we get the ELBO $\mathcal{L}(\phi^Z, \phi^\theta)$

$$\mathcal{L}(\phi^Z, \phi^\theta) = \mathbb{E}_{q(Z;\phi^Z)q(\theta;\phi^\theta)}[log\ \frac{p(Y,Z|\theta)p(\theta;\lambda)}{q(Z;\phi^Z)q(\theta;\phi^\theta)}]$$

Recall that we need to solve $\arg\max_{\phi^Z}\ \mathcal{L}(\phi^Z, \phi^\theta)$ and $\arg\max_{\phi^\theta}\ \mathcal{L}(\phi^Z, \phi^\theta)$ separately in E-step and M-step. Thus, we can derive

$$\frac{d}{d\phi^Z}\mathcal{L}(\phi^Z, \phi^\theta) = 0$$

$$\frac{d}{d\phi^\theta}\mathcal{L}(\phi^Z, \phi^\theta) = 0$$

Then, we can derive further

$$\frac{d}{dq(Z;\phi^Z)}\mathcal{L}(\phi^Z, \phi^\theta)$$

$$= \frac{d}{dq(Z;\phi^Z)}\int_{Z,\theta} q(Z;\phi^Z)q(\theta;\phi^\theta)log\frac{p(Y,Z|\theta)p(\theta;\lambda)}{q(Z;\phi^Z)q(\theta;\phi^\theta)}dZd\theta$$

$$= \int_{Z,\theta} q(\theta;\phi^\theta)log\ p(Y,Z|\theta)p(\theta;\lambda)dZd\theta - \int_{Z,\theta} q(\theta;\phi^\theta)log\ q(\theta;\phi^\theta)dZd\theta$$

$$- \int_{Z,\theta} q(\theta;\phi^\theta)log\ q(Z;\phi^Z)dZd\theta - \int_{Z,\theta} q(Z;\phi^Z)q(\theta;\phi^\theta)\frac{1}{q(Z;\phi^Z)}dZd\theta$$

$$= \mathbb{E}_{q(\theta;\phi^\theta)}[log\ p(Y,Z|\theta) + log\ p(\theta;\lambda) - log\ q(\theta;\phi^\theta) - \mathbb{E}_{q(Z;\phi^Z)}[log\ q(Z;\phi^Z)] - 1]$$

$$-\frac{d}{dq(Z;\phi^Z)}\int_{Z,\theta} q(Z;\phi^Z)q(\theta;\phi^\theta)log\ q(Z;\phi^Z)dZd\theta$$

Variational Bayesian EM Algorithm

Iterate until $\mathcal{L}(\phi^Z, \phi^\theta)$ converge

- E Step: Update the variational distribution on Z

  $q(Z;\phi^Z) \propto e^{(\mathbb{E}_{q(\theta;\phi^\theta)}[log\ p(Y,Z,\theta)])}$

- M Step: Update the variational distribution on θ

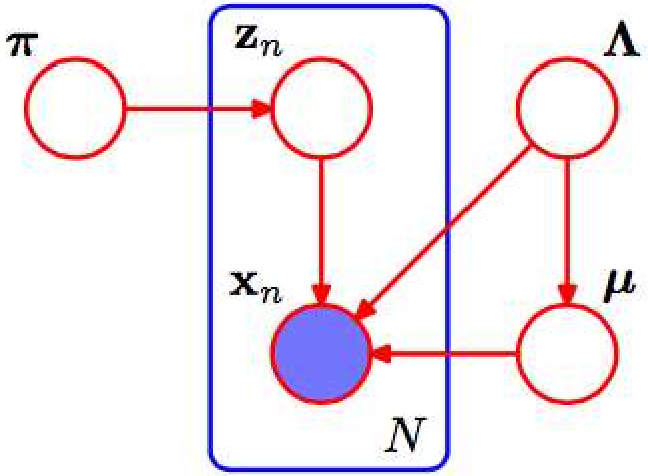
  $q(\theta;\phi^\theta) \propto e^{(\mathbb{E}_{q(Z;\phi^Z)}[log\ p(Y,Z,\theta)])}$

---

3. Variational Bayesian Gaussian Mixture Model

Graphical Model

Gaussian Mixture Model & Clustering



The variational Bayesian Gaussian mixture model(VB-GMM) can be represented as the above graphical model. We see each data point as a Gaussian mixture distribution with K components. We also denote the number of data points as N. Each x_n is a Gaussian mixture distribution with a weight π_n corresponds to a data point. z_n is an one-hot latent variable that indicates which cluster(component) does the data point belongs to. Finally, A component k follows the Gaussian distribution with mean μ_k and covariance matrix Λ_k. $\Lambda = \{\Lambda_1, ..., \Lambda_K\}$ and $\mu = \{\mu_1, ..., \mu_K\}$ are vectors denote the parameters of Gaussian mixture distribution.

Thus, the joint distribution of the VB-GMM is

$$p(X, Z, \pi, \mu, \Lambda) = p(X|Z, \pi, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)$$

$p(X|Z, \pi, \mu, \Lambda)$ denotes the Gaussian mixture model given on the latent variables and parameters. $p(Z|\pi)$ denotes the latent variables. As for priors, $p(\pi)$ denotes the prior distribution on the latent variables Z and $p(\mu|\Lambda)p(\Lambda)$ denotes the priors distribution on the Gaussian distribution X.

Gaussian Mixture Model

Suppose each data point $x_n \in \mathbb{R}^D$ has dimension D. We define the latent variables $Z = \{z_1, ..., z_N\}, Z \in \mathbb{R}^{N \times K}$, where $z_i = \{z_{i1}, ..., z_{iK}\}, z_i \in \mathbb{R}^K, z_{ij} \in \{0, 1\}$. Each z_i is a vector containing k binary variables. z_i can be seen as an one-hot encoding that indicates which cluster belongs to. As for $\pi \in \mathbb{R}^K$, π is the weight of the Gaussian mixture model of each component.

$$p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

Then, we define the components of the Gaussian mixture model. Each component follows Gaussian distribution and is parametrized by the mean μ_k and covariance matrix Λ_k^{-1}. Thus, the conditional distribution of the observed data $X \in \mathbb{R}^{N \times D}$, given the variables Z, μ, Λ is

$$p(X|Z,\mu,\Lambda) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}$$

where data X contains N data points and D dimensions, parameter $\mu \in \mathbb{R}^K, \mu = \{\mu_1, ..., \mu_K\}$ and $\Lambda \in \mathbb{R}^{K\times D\times D}, \Lambda_k \in \mathbb{R}^{D\times D}, \Lambda = \{\Lambda_1, ..., \Lambda_K\}$ are the mean and the covariance matrix of each component of Gaussian mixture model.

Dirichlet Distribution

Next, we introduce another prior over the parameters. We choose the symmetric Dirichlet distribution over the mixing proportions π. Support $x_1, ..., x_K$ where $x_i \in (0,1)$ and $\sum_{i=1}^{K} x_i = 1, K > 2$ with parameters $\alpha_1, ..., \alpha_K > 0$

$$X \sim \mathcal{D}ir(\alpha) = \frac{1}{B(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where the Beta function $B(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}$ and α and X are a set of random variables that $\alpha = \{\alpha_1, ..., \alpha_K\}$ and $X = \{X_1, ..., X_K\}$. Note that x_i is a sample value generated by X_i.

Expectation

The mean of the Dirichlet distribution is

$$E[X_i] = \frac{\alpha_i}{\sum_{k=1}^{K}\alpha_k}$$

$$E[ln\ X_i] = \psi(\alpha_i) - \psi(\sum_{k=1}^{K}\alpha_k)$$

where ψ is **digamma** function

$$\psi(x) = \frac{d}{dx}ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)} \approx ln(x) - \frac{1}{2x}$$

Symmetric Dirichlet distribution

In order to reduce the number of initial parameters, we use **Symmetric Dirichlet distribution** which is a special form of Dirichlet distribution that defined as the following

$$X \sim \mathcal{S}ymm\mathcal{D}ir(\alpha_0) = \frac{\Gamma(\alpha_0 K)}{\Gamma(\alpha_0)^K}\prod_{i=1}^{K} x_i^{\alpha_0 - 1} = f(x_1, ..., x_{K-1}; \alpha_0)$$

where $X = \{X_1, ..., X_{K-1}\}$. The α parameter of the symmetric Dirichlet is a scalar which means all the elements α_i of the α are the same $\alpha = \{\alpha_0, ..., \alpha_0\}$.

With Gaussian Mixture Model

Thus, we can model the distribution of the weights of Gaussian mixture model as a symmetric Dirichlet distribution.

$$p(\pi) = \mathcal{D}ir(\pi|\alpha_0) = \frac{1}{B(\alpha_0)} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$$

Gaussian Wishart Distribution

If a normal distribution whose parameters follow the Wishart distribution. It is called **Gaussian-Wishart distribution**. Support $\mu \in \mathbb{R}^D$ and $\Lambda \in \mathbb{R}^{D \times D}$, they are generated from Gaussian-Wishart distribution which is defined as

$$(\mu, \Lambda) \sim \mathcal{NW}(\mu_0, \lambda, W, \nu) = \mathcal{N}(\mu|\mu_0, (\lambda\Lambda)^{-1})\mathcal{W}(\Lambda|W, \nu)$$

where $\mu_0 \in \mathbb{R}^D$ is the location, $W \in \mathbb{R}^{D \times D}$ represent the scale matrix, $\lambda \in \mathbb{R}, \lambda > 0$, and $\nu \in \mathbb{R}, \nu > D - 1$.

Posterior

After making n observations $\{x_1, ..., x_n\}$ with mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, the posterior distribution of the parameters is

$$(\mu, \Lambda) \sim \mathcal{NW}(\mu_n, \lambda_n, W_n, \nu_n)$$

where

$$\lambda_n = \lambda + n$$

$$\mu_n = \frac{\lambda\mu_0 + n\bar{x}}{\lambda + n}$$

$$\nu_n = \nu + n$$

$$W_n^{-1} = W_0^{-1} + \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top + \frac{n\lambda}{n + \lambda} (\bar{x} - \mu_0)(\bar{x} - \mu_0)^\top$$

With Gaussian Mixture Model

$$p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|W_0, \nu_0)$$

E-Step

E-Step aims to update the variational distribution on latent variables Z

$$ln\ q(Z; \phi^Z) \propto \mathbb{E}_{q(\theta;\phi^\theta)}[log\ p(Y, Z, \theta)]$$

Thus, we can derive

$$ln\ q(Z) \propto \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(X, Z, \pi, \mu, \Lambda)]$$

$$= \mathbb{E}_\pi[ln\ p(Z|\pi)] + \mathbb{E}_{\mu,\Lambda}[ln\ p(X|Z, \mu, \Lambda)] + \mathbb{E}_{\pi,\mu,\Lambda}[ln\ p(\pi, \mu, \Lambda)]$$

$$= \mathbb{E}_\pi[ln\ p(Z|\pi)] + \mathbb{E}_{\mu,\Lambda}[ln\ p(X|Z, \mu, \Lambda)] + C$$

where

$$\mathbb{E}_{\pi}[ln\ p(Z|\pi)] = \mathbb{E}_{\pi}\Big[ln\ \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}\Big]$$

$$= \mathbb{E}_{\pi}\Big[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ ln\ \pi_k\Big]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ \mathbb{E}_{\pi}[ln\ \pi_k]$$

and

$$\mathbb{E}_{\mu,\Lambda}[ln\ p(X|Z,\mu,\Lambda)] = \mathbb{E}_{\mu,\Lambda}\Big[ln\ \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}\Big]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ \mathbb{E}_{\mu_k,\Lambda_k}\Big[ln\ \frac{e^{-\frac{1}{2}(x_n-\mu_k)^\top\Lambda(x_n-\mu_k)}}{\sqrt{(2\pi)^D det(\Lambda_k^{-1})}}\Big]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ \mathbb{E}_{\mu_k,\Lambda_k}\Big[-\frac{1}{2}(x_n-\mu_k)^\top\Lambda(x_n-\mu_k) - \frac{1}{2}ln((2\pi)^D det(\Lambda_k^{-1}))\Big]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\Big(-\frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\Big[(x_n-\mu_k)^\top\Lambda(x_n-\mu_k)\Big] - \frac{D}{2}ln\ 2\pi + \mathbb{E}_{\Lambda_k}\Big[ln\ det(\Lambda_k)\Big]\Big)$$

Due to simplification, let

$$ln\ \rho_{nk} = \mathbb{E}_{\pi}[ln\ \pi_k] - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\Big[(x_n-\mu_k)^\top\Lambda(x_n-\mu_k)\Big] - \frac{D}{2}ln\ 2\pi + \mathbb{E}_{\Lambda_k}\Big[ln\ det(\Lambda_k)\Big]$$

Thus,

$$ln\ q(Z) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} ln\ \rho_{nk} + C$$

In order to normalize the factor of ρ_{nk}, we divide the ρ_{nk} by $\sum_{j=1}^{K}\rho_{nj}$ and obtain the r_{nk}.

$$ln\ q(Z) \propto \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} ln\ r_{nk}, \text{where } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}$$

For convenience, we also define some useful variables.

$$N_k = \sum_{n=1}^{N} r_{nk}, \quad \bar{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}x_n, \quad S_k = \frac{1}{N_k} r_{nk}(x_n-\bar{x}_k)(x_n-\bar{x}_k)^\top$$

M-Step

E-Step aims to update the variational distribution on variables θ

$$ln\ q(\theta;\phi^{\theta}) \propto \mathbb{E}_{q(Z;\phi^Z)}[log\ p(Y,Z,\theta)]$$

Thus, we can derive

$$ln\ q(\pi, \mu, \Lambda) \propto \mathbb{E}_Z[ln\ p(X, Z, \pi, \mu, \Lambda)]$$

$$= \mathbb{E}_Z[ln\ p(X|Z, \pi, \mu, \Lambda)] + \mathbb{E}_Z[ln\ p(Z|\pi)] + \mathbb{E}_Z[ln\ p(\pi)] + \mathbb{E}_Z[ln\ p(\mu, \Lambda)]$$

Absolutely, we assume the joint distribution of parameters follows **mean field theorem** that the parameters of each component are independent $q(\pi, \mu, \Lambda) = q(\pi) \prod_{i=1}^{N} q(\mu_i, \Lambda_i)$. With it, the problem would be easier to solve.

Dirichlet Distribution

Gaussian-Wishart Distribution