Expectation Maximization

EM in general

- Suppose θ contains all parameters to estimate
- In MLE, we are to maximize $P(x;\theta)$ (evidence) on training data $D = \{x_1, \dots, x_n\}$. We assume directly optimizing $P(x;\theta)$ is difficult, but optimizing $P(x,z;\theta)$ is easy

$$\begin{aligned} P(x,z) &= P(z)P(x|z) \\ P(x) &= \sum_z P(x,z) \\ P(z|x) &= \frac{P(z)P(x|z)}{\sum_z P(z)P(x|z)} \end{aligned}$$

  However, $\sum_z P(z)P(x|z)$ is *intractable* due to sum in denominator
  Find a **variational distribution** $q(z|x)$ (which is *tractable*) to approximate $P(z|x)$
- Now we introduce $q(z|x)$ to approximate $P(z|x;\theta)$
  We have the following always holds:

$$\begin{aligned} \log P(x;\theta) &= \sum_z q(z|x) \log P(x;\theta) \\ &= \sum_z q(z|x) \big(\log P(x,z;\theta) - \log P(z|x;\theta)\big) \\ &= \sum_z q(z|x) \Big(\big(\log P(x,z;\theta) - \log q(z|x)\big) - \big(\log P(z|x;\theta) - \log q(z|x)\big)\Big) \\ &= \sum_z q(z|x) \log \frac{P(x,z;\theta)}{q(z|x)} - \sum_z q(z|x) \log \frac{P(z|x;\theta)}{q(z|x)} \end{aligned}$$

  - first half is called **Evidence Lower Bound (ELBO)**
  - second half is called **KL divergence** or KL$(q||p)$
    - measures divergence of p and q
    - $\sum_z q(z|x) \log \frac{P(z|x;\theta)}{q(z|x)} \leq \log \sum_z q(z|x) \frac{P(z|x;\theta)}{q(z|x)} = 0$ (Jensen's Inequality)
      $\Rightarrow \mathrm{KL}(q||p) \geq 0$, when $q = p$, takes 0
- EM is maximizing ELBO. Maximizing the lower bound also increase evidence

EM steps

- EM is a two-step iterative algorithm to *maximize ELBO*

$$L(q,\theta) = \sum_z q(z|x) \log \frac{P(x,z;\theta)}{q(z|x)}$$

1. **E-step**: Given θ fixed, optimize q
   Let's assume current $\theta = \theta^{old}$, ELBO $L(q, \theta^{old})$ is functional of q
   Also when θ^{old} is fixed, $\log P(x;\theta^{old})$ is a *constant*
   To maximize ELBO w.r.t $q \quad \Rightarrow \quad min\, \mathrm{KL}(q||p) \quad \Rightarrow \quad q(z|x) = P(z|x;\theta^{old})$
   That is, E-step just let $q(z|x)$ take the **posterior** $P(z|x;\theta^{old})$
2. **M-step**: Given q fixed, optimize θ
   Now we have $q(z|x) = P(z|x;\theta^{old})$. Substitute into ELBO:

$$
\begin{aligned}
L(q,\theta) &= \sum_z P(z|x;\theta^{old}) \log \frac{P(x,z;\theta)}{P(z|x;\theta^{old})} \\
&= \sum_z P(z|x;\theta^{old}) \log P(x,z;\theta) - \text{const} \\
&\Rightarrow \underset{\theta}{max}\, E_z \log P(x,z;\theta) \quad z \sim P(z|x;\theta^{old})
\end{aligned}
$$

3. Repeat 1 and 2 until convergence

EM for GMM

- Objective:

$$
\log P(x;\theta) = \sum_{i=1}^{n} \log\big(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k,\Sigma_k)\big)
$$

- **E-step** compute $q(z|x) = P(z|x;\mu,\Sigma,\pi)$
  Let $z_{ik} = \{0,1\}$ indicates x_i whether generated from cluster k

$$
\begin{aligned}
P(z,x;\mu,\Sigma,\pi) &= \prod_{i=1}^{n} P(x_i|z_i;\mu,\Sigma)P(z_i;\pi) \\
&= \prod_{i=1}^{n}\prod_{k=1}^{K} N(x_i|\mu_k,\Sigma_k)^{z_{ik}} \prod_{k=1}^{K} \pi_k^{z_{ik}} \\
&= \prod_{i=1}^{n}\prod_{k=1}^{K} \big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)^{z_{ik}} \\
P(z|x;\mu,\Sigma,\pi) &= \frac{1}{P(x)} \prod_{i=1}^{n}\prod_{k=1}^{K} \big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)^{z_{ik}} \\
&= \prod_{i=1}^{n} \Big(\frac{1}{P(x_i)} \prod_{k=1}^{K} \big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)^{z_{ik}}\Big) \\
&= \prod_{i=1}^{n} P(z_i|x_i;\mu,\Sigma,\pi)
\end{aligned}
$$

- **M-step**
  Objective:

$$
\begin{aligned}
&\underset{\mu,\Sigma,\pi}{max} E_z\big(\log P(x,z;\mu,\Sigma,\pi)\big) \quad z \sim P(z|x;\mu^{old},\Sigma^{old},\pi^{old}) \\
&= E_z\Big(z_{ik} \log\big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)\Big) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} E_z(z_{ik}) \log\big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)
\end{aligned}
$$

  We can compute $E_z(z_{ik})$

$$
\begin{aligned}
E_z(z_{ik}) &= 1 \cdot P(z_{ik}=1|x_i) + 0 \cdot P(z_{ik}=0|x_i) \\
&= \frac{P(z_{ik}=1,x_i)}{P(x_i)} \\
&= \frac{\pi_k N(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_k,\Sigma_k)} = \gamma_{ik}
\end{aligned}
$$

  Update the objective:

$$\underset{\mu,\Sigma,\pi}{max}\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}\log\big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)\quad \text{s.t.}\sum_{k=1}^{K}\pi_k=1$$

Use Lagrange function

$$L(\mu,\Sigma,\pi,\lambda)=\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}\log\big(\pi_k N(x_i|\mu_k,\Sigma_k)\big)+\lambda(1-\sum_{k=1}^{K}\pi_k)$$

- Compute π_k

$$\begin{aligned}\frac{\partial L}{\partial \pi_k}&=\sum_{i=1}^{n}\frac{\gamma_{ik}}{\pi_k}-\lambda=0\\ \pi_k&=\frac{\sum_{i=1}^{n}\gamma_{ik}}{\lambda}\\ \lambda\sum_{k=1}^{K}\pi_k&=\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}=n\quad\Rightarrow\quad\lambda=n\\ \pi_k&=\frac{\sum_{i=1}^{n}\gamma_{ik}}{n}\end{aligned}$$

- Compute μ_k and Σ_k
  We know that

$$N(x_i|\mu_k,\Sigma_k)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp\big(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\big)$$

  So

$$\begin{aligned}L(\mu,\Sigma,\pi,\lambda)&=\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}\big(\log\pi_k-\frac{d}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\big)\\ \frac{\partial L}{\partial\mu_k}&=\sum_{i=1}^{n}\gamma_{ik}\Sigma_k^{-1}(x_i-\mu_k)=0\\ \mu_k&=\frac{\sum_{i=1}^{n}\gamma_{ik}x_i}{\sum_{i=1}^{n}\gamma_{ik}}\\ \frac{\partial L}{\partial\Sigma_k}&=\sum_{i=1}^{n}\gamma_{ik}\frac{1}{|\Sigma_k|}\frac{\partial|\Sigma_k|}{\partial\Sigma_k}+\gamma_{ik}(x_i-\mu_k)(x_i-\mu_k)^T\\ &=-\sum_{i=1}^{n}\gamma_{ik}\Sigma_k+\sum_{i=1}^{n}\gamma_{ik}(x_i-\mu_k)(x_i-\mu_k)^T=0\\ \Sigma_k&=\frac{\sum_{i=1}^{n}\gamma_{ik}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\gamma_{ik}}\end{aligned}$$