CSCI 567 HW 4

Amir Erfan Eshratifar
USC ID: 2875431498

1 Generative models

1. Because the samples are assumed to be i.i.d the likelihood function:

$$P(X = x_1, X = x_2, \dots, X = x_n|\theta) = P(X = x_1|\theta)P(X = x_2|\theta)\dots P(X = x_n|\theta) = \begin{cases} \frac{1}{\theta^n} & \forall x_i : x_i \in [0, \theta) \\ 0 & otherwise \end{cases}$$

$\theta^*_{MLE} \geq max(x_1, x_2, \dots, x_n)$ because all x_is must be included in $[0, \theta)$.

2. (a) By applying Bayes rule:

$$P(Z_n = k|x_n, \theta_1, \theta_2, \omega_1, \omega_2) = \frac{P(X = x_n|Z_n = k, \theta_1, \theta_2, \omega_1, \omega_2)P(Z_n = k|\theta_1, \theta_2, \omega_1, \omega_2)}{P(X = x_n|\theta_1, \theta_2, \omega_1, \omega_2)}$$

$$= \frac{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1}{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1 + \frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2}1[k == 1] + \frac{\frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2}{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1 + \frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2}1[k == 2]$$

(b) By θ, I mean $(\theta_1, \theta_2, \omega_1, \omega_2)$:

$$Q_q(\theta, \theta^{OLD}) = \sum_n \sum_k P(k|x_n; \theta^{OLD}) \log P(x_n, k|\theta)$$

$$= \sum_n P(k = 1|x_n; \theta^{OLD}) \log \frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1 + P(k = 2|x_n; \theta^{OLD}) \log\{\frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2\}$$

$$= \sum_n \{\frac{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1}{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1 + \frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2}1[k == 1] \log\{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1\}$$

$$+ \frac{\frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2}{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1 + \frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2}1[k == 1] \log\{\frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2\}\}$$

(c)

$$Q_q(\theta, \theta^{OLD}) = \sum_n \{P(k = 1|x_n; \theta^{OLD}) \log\{\frac{1}{\theta_1}1[0 < x_n \leq \theta_1]\omega_1\} + P(k = 2|x_n; \theta^{OLD}) \log\{\frac{1}{\theta_2}1[0 < x_n \leq \theta_2]\omega_2\}\}$$

Now we have to maximize $Q_q(\theta, \theta^{OLD})$:

$$\Rightarrow \theta_1 \geq max(x_1, x_2, \dots, x_n), \theta_2 \geq max(x_1, x_2, \dots, x_n)$$

2 Mixture Density Models

1.

$$P(x_b|x_a) = \frac{P(x_a, x_b)}{P(x_a)} = \frac{\sum_{k=1}^K \pi_k P(x_a, x_b|k)}{\sum_{k=1}^K P(x_a, k)} = \frac{\sum_{k=1}^K \pi_k P(x_a|k)P(x_b|x_a, k)}{\sum_{k=1}^K P(k)P(x_a|k)}$$

$$= \frac{\sum_{k=1}^K \pi_k P(x_a|k)P(x_b|x_a, k)}{\sum_{k=1}^K \pi_k P(x_a|k)} = \sum_{k=1}^K \frac{\pi_k P(x_a|k)}{\sum_{k=1}^K \pi_k P(x_a|k)} P(x_b|x_a, k)$$

$$\Rightarrow \lambda_k = \frac{\pi_k P(x_a|k)}{\sum_{k=1}^K \pi_k P(x_a|k)}$$

It is clear that λ_k is non-negative and:

$$\sum_{k=1}^K \lambda_k = \sum_{k=1}^K \frac{\pi_k P(x_a|k)}{\sum_{k=1}^K \pi_k P(x_a|k)} = 1$$

3 The connection between GMM and K-means

1. Assuming that there is only one minimum for $\left\|x_n - \mu_j\right\|^2$:

$$\begin{aligned}
\lim_{\sigma \to 0} \gamma(z_{nk}) &= \lim_{\sigma \to 0} \frac{\pi_k exp(-\|x_n - \mu_k\|^2 /(2\sigma^2))}{\sum_j \pi_j exp(-\left\|x_n - \mu_j\right\|^2 /(2\sigma^2))} \\
&= \lim_{\sigma \to 0} \frac{\pi_k exp(-\|x_n - \mu_k\|^2 /(2\sigma^2))}{\max_j \pi_j exp(-\left\|x_n - \mu_j\right\|^2 /(2\sigma^2))} \\
&= \lim_{\sigma \to 0} \frac{\pi_k exp(-\|x_n - \mu_k\|^2 /(2\sigma^2))}{\pi_j exp(\max_j(-\left\|x_n - \mu_j\right\|^2 /(2\sigma^2)))} \\
&= \lim_{\sigma \to 0} \frac{\pi_k exp(-\|x_n - \mu_k\|^2 /(2\sigma^2))}{\pi_j exp(-\min_j(\left\|x_n - \mu_j\right\|^2)/(2\sigma^2))} \\
&= \begin{cases} 1 & k = argmin_j \left\|x_n - \mu_j\right\|^2 \\ 0 & otherwise \end{cases}
\end{aligned}$$

$$\begin{aligned}
\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n, z_n = k) &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})[\log \pi_k + \log N(x_n|\mu_k, \sigma^2 I) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})[\log \pi_k + \log det(2\pi\Sigma)^{-0.5} + -0.5(x_n - \mu_k)^T(\sigma^2 I)^{-1}(x_n - \mu_k)] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})[const + -0.5/\sigma^2 \|x_n - \mu_k\|^2]
\end{aligned}$$

$$\begin{aligned}
\Rightarrow max_{\mu_k} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n, z_n = k) &= max_{\mu_k} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})[const + -0.5/\sigma^2 \|x_n - \mu_k\|^2] \\
&= max_{\mu_k} \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}[-0.5/\sigma^2 \|x_n - \mu_k\|^2] \\
&= min_{\mu_k} \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2
\end{aligned}$$

Therefore, if σ goes to infinity, the maximization of complete data log-likelihood of GMM model is equal to minimization of the cost function in k-means.

4 Naive Bayes

1.

$$\begin{aligned}
L = \log P(D) &= \sum_{n=1}^{N} log(P(Y_n = y_n) \prod_{d=1}^{D} P(X_{nd} = x_{nd}|Y_n = y_n)) \\
&= \sum_{c=1}^{C}\sum_{n=1}^{N} 1(y_n == c) \log P(Y_n = y_n) + \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{d=1}^{D} 1(y_n == c) \log P(X_{nd} = x_{nd}|Y_n = y_n) \\
&= \sum_{c=1}^{C}\sum_{n=1}^{N} 1(y_n == c) \log \pi_c + \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{d=1}^{D} 1(y_n == c)[-\frac{(x_{nd} - \mu_{cd})^2}{2\sigma_{cd}^2} - \log \sigma_{cd} - \log \sqrt{2\pi}]
\end{aligned}$$

2. Gradient with respect to μ_{cd}:

$$\frac{\partial \log P(D)}{\partial \mu_{cd}} = \sum_n 1(y_n == c)\frac{(x_{nd} - \mu_{cd})^2}{2\sigma_{cd}^2} = 0 \Rightarrow \mu_{cd}^* = \frac{\sum_{n=1}^N 1(y_n == c)x_{nd}}{\sum_{n=1}^N 1(y_n == c)}$$

Gradient with respect to σ_{cd}:

$$\frac{\partial \log P(D)}{\partial \sigma_{cd}} = \sum_n 1(y_n == c)\frac{(x_{nd} - \mu_{cd}^*)^2}{2\sigma_{cd}^3} - \frac{1}{\sigma_{cd}} = 0 \Rightarrow \sigma_{cd}^* = \sqrt{\frac{\sum_{n=1}^N 1(y_n == c)(x_{nd} - \mu_{nd}^*)^2}{\sum_{n=1}^N 1(y_n == c)}}$$

First we need apply the constraint $\sum_c \pi_c = 1$ using Lagrangian:

$$L(log(P(D), \lambda) = \sum_{c=1}^C \sum_{n=1}^N 1(y_n == c) \log \pi_c + \sum_{c=1}^C \sum_{n=1}^N \sum_{d=1}^D 1(y_n == c)[-\frac{(x_{nd} - \mu_{cd})^2}{2\sigma_{cd}^2} - \log \sigma_{cd} - \log \sqrt{2\pi}] + \lambda(1 - \sum_c \pi_c)$$

Gradient with respect to π_c:

$$\frac{\partial \log L(log(P(D), \lambda)}{\partial \pi_c} = \sum_n 1(y_n == c)/\pi_c - \lambda = 0 \Rightarrow \pi_c = \frac{1}{\lambda}\sum_n 1(y_n == c)$$

$$\sum_c \pi_c = 1 \Rightarrow \lambda = \sum_c \sum_n 1(y_n == c) = N$$

$$\Rightarrow \pi_c^* = \frac{1}{N}\sum_n 1(y_n == c)$$