Homework 4

CSCI 567: Machine Learning

Seyed Mohammad Asghari Pari

USCID: 5788788474

Generative models

Solution to Question 1.1:

Since $x_1, x_2, \ldots, x_N$ are independent, we have

$$P(x_1, x_2, \ldots, x_N; \theta) = \prod_{n=1}^{N} P(x_n; \theta) = \prod_{n=1}^{N} \frac{1}{\theta} \mathbf{1}[0 < x_n \leq \theta]. \tag{1}$$

Then,

$$\begin{aligned} l(\theta) = \log P(x_1, x_2, \ldots, x_N; \theta) &= \sum_{n=1}^{N} \log \frac{1}{\theta} \mathbf{1}[0 < x_n \leq \theta] \\ &= \begin{cases} -\sum_{n=1}^{N} \log \theta = -N \log \theta : \theta \geq \max_i x_i \\ -\infty \qquad\qquad\qquad\qquad\quad : \text{o.w.} \end{cases} \end{aligned} \tag{2}$$

Now, by taking derivative w.r.t. θ, we get $\frac{\partial l(\theta)}{\partial \theta} = -\frac{N}{\theta}$ which is decreasing in θ. Hence, the maximum is achieved by choosing $\theta^{ML} = \max_n x_n$.

Solution to Question 1.2:

- By definition of conditional expectation, we have for $k \in \{1, 2\}$

$$\begin{aligned} P(k|x_n; \theta_1, \theta_2, \omega_1, \omega_2) &= \frac{P(x_n, k; \theta_1, \theta_2, \omega_1, \omega_2)}{P(x_n; \theta_1, \theta_2, \omega_1, \omega_2)} \\ &= \frac{P(x_n|k; \theta_1, \theta_2, \omega_1, \omega_2) P(k; \theta_1, \theta_2, \omega_1, \omega_2)}{\sum_{k' \in \{1,2\}} P(x_n|k'; \theta_1, \theta_2, \omega_1, \omega_2) P(k'; \theta_1, \theta_2, \omega_1, \omega_2)} \\ &= \frac{U(X = x_n|\theta_k)\omega_k}{U(X = x_n|\theta_1)\omega_1 + U(X = x_n|\theta_2)\omega_2}. \end{aligned} \tag{3}$$

- Let $\boldsymbol{\theta} = \{\theta_1, \theta_2, \omega_1, \omega_2\}$, then

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD}) = \sum_{n=1}^{N} \sum_{k \in \{1,2\}} P(k|x_n; \boldsymbol{\theta}^{OLD}) \log P(x_n, k|\boldsymbol{\theta}). \tag{4}$$

  Note that from (3), we have

$$\begin{aligned} P(k|x_n; \boldsymbol{\theta}^{OLD}) &= P(k|x_n; \theta_1^{OLD}, \theta_2^{OLD}, \omega_1^{OLD}, \omega_2^{OLD}) \\ &= \frac{U(X = x_n|\theta_k^{OLD})\omega_k^{OLD}}{U(X = x_n|\theta_1^{OLD})\omega_1^{OLD} + U(X = x_n|\theta_2^{OLD})\omega_2^{OLD}}. \end{aligned} \tag{5}$$

Further,

$$
\begin{aligned}
P(x_n, k|\boldsymbol{\theta}) &= P(x_n, k; \theta_1, \theta_2, \omega_1, \omega_2) = P(x_n|k; \theta_1, \theta_2, \omega_1, \omega_2)P(k; \theta_1, \theta_2, \omega_1, \omega_2) \\
&= U(X = x_n|\theta_k)\omega_k. \qquad (6)
\end{aligned}
$$

Substituting (5) and (6) in (4), we get,

$$
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD}) = \sum_{n=1}^{N} \sum_{k\in\{1,2\}} \frac{U(X = x_n|\theta_k^{OLD})\omega_k^{OLD}}{U(X = x_n|\theta_1^{OLD})\omega_1^{OLD} + U(X = x_n|\theta_2^{OLD})\omega_2^{OLD}} \log U(X = x_n|\theta_k)\omega_k. \qquad (7)
$$

Note that since $\theta_2^{OLD} \geq \max_n x_n$, we have $U(X = x_n|\theta_2^{OLD}) = \dfrac{1}{\theta_2^{OLD}}$ for all $n = 1, \ldots, N$. Furthermore, since $\min_n x_n \leq \theta_1^{OLD} \leq \max_n x_n$, if $x_n > \theta_1^{OLD}$, we have $U(X = x_n|\theta_1^{OLD}) = 0$. Considering these, (7) can be written as

$$
\begin{aligned}
&Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD}) \\
&= \sum_{n:x_n \leq \theta_1^{OLD}} \frac{U(X = x_n|\theta_1^{OLD})\omega_1^{OLD}}{U(X = x_n|\theta_1^{OLD})\omega_1^{OLD} + \frac{1}{\theta_2^{OLD}}\omega_2^{OLD}} \log U(X = x_n|\theta_1)\omega_1 \\
&+ \sum_{n} \frac{\frac{1}{\theta_2^{OLD}}\omega_2^{OLD}}{U(X = x_n|\theta_1^{OLD})\omega_1^{OLD} + \frac{1}{\theta_2^{OLD}}\omega_2^{OLD}} \log U(X = x_n|\theta_2)\omega_2. \qquad (8)
\end{aligned}
$$

- Note that from (8), if $x_n > \theta_1$ for any n that $x_n \leq \theta_1^{OLD}$, then $U(X = x_n|\theta_1) = 0$ and $\log U(X = x_n|\theta_1)\omega_1 = -\infty$, and hence $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD}) = -\infty$. Furthermore, from (8), if $x_n > \theta_2$ for any $n = 1, \ldots, N$, then $U(X = x_n|\theta_2) = 0$ and $\log U(X = x_n|\theta_2)\omega_2 = -\infty$, and hence $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD}) = -\infty$. Now, we use these facts to find $\boldsymbol{\theta}^{NEW}$ such that $= \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD})$. Note that $\dfrac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD})}{\partial \theta_1} = \sum_{n:x_n \leq \theta_1^{OLD}} P_{OLD}(1|x_n)(\dfrac{-1}{\theta_1})$ which is decreasing in θ_1. Hence, the maximum is achieved by choosing

$$
\theta_1^{NEW} = \max_{n:x_n \leq \theta_1^{OLD}} x_n. \qquad (9)
$$

Furthermore, $\dfrac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{OLD})}{\partial \theta_2} = \sum_n P_{OLD}(2|x_n)(\dfrac{-1}{\theta_2})$ which is decreasing in θ_2. Hence, the maximum is achieved by choosing

$$
\theta_2^{NEW} = \max_n x_n. \qquad (10)
$$

Mixture density models

Solution to Question 2.1:

By definition of conditional expectation, we have

$$P(\boldsymbol{x}_b|\boldsymbol{x}_a) = \frac{P(\boldsymbol{x}_b, \boldsymbol{x}_a)}{P(\boldsymbol{x}_a)} = \frac{\sum_{k=1}^{K} P(\boldsymbol{x}_b, \boldsymbol{x}_a, k)}{P(\boldsymbol{x}_a)}$$

$$= \frac{\sum_{k=1}^{K} P(\boldsymbol{x}_b|\boldsymbol{x}_a, k) P(\boldsymbol{x}_a, k)}{P(\boldsymbol{x}_a)} = \sum_{k=1}^{K} \frac{P(\boldsymbol{x}_a, k)}{P(\boldsymbol{x}_a)} P(\boldsymbol{x}_b|\boldsymbol{x}_a, k). \tag{11}$$

We define $\lambda_k = \frac{P(\boldsymbol{x}_a, k)}{P(\boldsymbol{x}_a)}$. Note that $\sum_{k=1}^{K} \lambda_k = \frac{\sum_{k=1}^{K} P(\boldsymbol{x}_a, k)}{P(\boldsymbol{x}_a)} = \frac{P(\boldsymbol{x}_a)}{P(\boldsymbol{x}_a)} = 1$. Further, λ_k can be written in terms of π_k and $P(\boldsymbol{x}_a|k)$ as follows,

$$\lambda_k = \frac{P(\boldsymbol{x}_a, k)}{P(\boldsymbol{x}_a)} = \frac{P(\boldsymbol{x}_a|k) P(k)}{\sum_{k'=1}^{K} P(\boldsymbol{x}_a|k') P(k')} = \frac{P(\boldsymbol{x}_a|k) \pi_k}{\sum_{k'=1}^{K} P(\boldsymbol{x}_a|k') \pi_{k'}}. \tag{12}$$

The connection between GMM and K-means

Solution to Question 3.1:

$$\gamma(z_{nk}) = \frac{\pi_k \exp(\frac{-||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2}{2\sigma^2})}{\sum_{j=1}^K \pi_j \exp(\frac{-||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2})} = \frac{1}{\sum_{j=1}^K \frac{\pi_j}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2})}. \quad (13)$$

Now, we consider two cases,

- If $k = \operatorname{argmin}_{k'} ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2$: in this case,

$$\begin{aligned} &||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 < ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2, \quad \forall k' \neq k \\ \Longrightarrow &||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2 < 0, \quad \forall k' \neq k \\ \Longrightarrow &\lim_{\sigma \to 0} \frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2}{2\sigma^2} = -\infty, \quad \forall k' \neq k \\ \Longrightarrow &\lim_{\sigma \to 0} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2}{2\sigma^2}) = 0, \quad \forall k' \neq k. \end{aligned} \quad (14)$$

  Note that (13) can be written as,

$$\begin{aligned} \gamma(z_{nk}) &= \frac{1}{\sum_{j=1}^K \frac{\pi_j}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2})} \\ &= \frac{1}{1 + \sum_{j \neq k} \frac{\pi_j}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2})}. \end{aligned} \quad (15)$$

  Now, using (14), $\lim_{\sigma \to 0} \gamma(z_{nk})$ from (15) can be written as,

$$\lim_{\sigma \to 0} \gamma(z_{nk}) = 1. \quad (16)$$

  Since in this case, $k = \operatorname{argmin}_{k'} ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2$, we have $r_{nk} = 1$. Hence, $\lim_{\sigma \to 0} \gamma(z_{nk}) = r_{nk}$.

- If $k \neq \operatorname{argmin}_{k'} ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2$: in this case,

$$\begin{aligned} &\exists l \neq k \quad ||\boldsymbol{x}_n - \boldsymbol{\mu}_l||^2 < ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2, \\ \Longrightarrow &\exists l \neq k \quad ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_l||^2 > 0, \\ \Longrightarrow &\exists l \neq k \quad \lim_{\sigma \to 0} \frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_l||^2}{2\sigma^2} = +\infty, \\ \Longrightarrow &\exists l \neq k \quad \lim_{\sigma \to 0} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_l||^2}{2\sigma^2}) = +\infty. \end{aligned} \quad (17)$$

  Note that (13) can be written as,

$$\begin{aligned} \gamma(z_{nk}) &= \frac{1}{\sum_{j=1}^K \frac{\pi_j}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2})} \\ &= \frac{1}{\frac{\pi_l}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_l||^2}{2\sigma^2}) + \sum_{j \neq l} \frac{\pi_j}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2})}. \end{aligned} \quad (18)$$

  Note that since $\exp(\cdot)$ is always non-negative, we have $\sum_{j \neq l} \frac{\pi_j}{\pi_k} \exp(\frac{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2}{2\sigma^2}) \geq 0$, and hence using (17), $\lim_{\sigma \to 0} \gamma(z_{nk})$ from (18) can be written as,

$$\lim_{\sigma \to 0} \gamma(z_{nk}) = \frac{1}{+\infty} = 0. \quad (19)$$

Since in this case, $k \neq \text{argmin}_{k'} ||\boldsymbol{x}_n - \boldsymbol{\mu}_{k'}||^2$, we have $r_{nk} = 0$. Hence, $\lim_{\sigma \to 0} \gamma(z_{nk}) = r_{nk}$.

Therefore, we showed that $\lim_{\sigma \to 0} \gamma(z_{nk}) = r_{nk}$. Now, using this, in the limit $\sigma \to 0$, we can write,

$$
\begin{aligned}
&\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})[\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{I})] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}[\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{I})] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}[\log \pi_k - \frac{1}{2\sigma^2} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 - \frac{1}{2} \log(2\pi)^D det(\sigma^2 \boldsymbol{I})]. \qquad (20)
\end{aligned}
$$

Note that the first term and the third term in (20) are independent of $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, therefore,

$$
\max_{\{\boldsymbol{\mu}_k\}_{k=1}^{K}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})[\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_n, \sigma^2 \boldsymbol{I})] \Leftrightarrow \min_{\{\boldsymbol{\mu}_k\}_{k=1}^{K}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2. \qquad (21)
$$

Naive Bayes

Solution to Question 4.1:

$$
\begin{aligned}
\log P(\mathcal{D}) &= \log \prod_{n=1}^{N} P(X = \boldsymbol{x}_n, Y = y_n; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) \\
&= \log \prod_{n=1}^{N} P(X = \boldsymbol{x}_n | Y = y_n; \boldsymbol{\mu}, \boldsymbol{\sigma}) P(Y = y_n; \boldsymbol{\pi}) \\
&= \sum_{n=1}^{N} [\log P(Y = y_n; \boldsymbol{\pi}) + \log P(X = \boldsymbol{x}_n | Y = y_n; \boldsymbol{\mu}, \boldsymbol{\sigma})] \\
&= \sum_{n=1}^{N} [\log P(Y = y_n; \boldsymbol{\pi}) + \log \prod_{d=1}^{D} P(X_d = x_{nd} | Y = y_n; \boldsymbol{\mu}, \boldsymbol{\sigma})] \\
&= \sum_{n=1}^{N} \log P(Y = y_n; \boldsymbol{\pi}) + \sum_{n=1}^{N} \sum_{d=1}^{D} \log P(X_d = x_{nd} | Y = y_n; \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&= \sum_{c=1}^{C} \sum_{n:y_n=c} \log P(Y = y_n; \boldsymbol{\pi}) + \sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} \log P(X_d = x_{nd} | Y = y_n; \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&= \sum_{c=1}^{C} \sum_{n:y_n=c} \log P(Y = c; \boldsymbol{\pi}) + \sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} \log P(X_d = x_{nd} | Y = c; \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&= \sum_{c=1}^{C} \sum_{n:y_n=c} \log \pi_c + \sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} [-\frac{1}{2} \log 2\pi\sigma_{cd}^2 - \frac{1}{2\sigma_{cd}^2}(x_{nd} - \mu_{cd})^2] \\
&= \sum_{c=1}^{C} \log \pi_c \times (\sum_{n:y_n=c} 1) + \sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} [-\frac{1}{2} \log 2\pi\sigma_{cd}^2 - \frac{1}{2\sigma_{cd}^2}(x_{nd} - \mu_{cd})^2]. \qquad (22)
\end{aligned}
$$

Solution to Question 4.2:

Now, we want to find $\pi_c, \mu_{cd}, \sigma_{cd}$ for all $c = 1, \ldots, C$ and $d = 1, \ldots, D$ that maximize (22).

- **Finding π_c^*:** Note that only the first term of (22) depends on π_c. Let $\alpha_c = \sum_{n:y_n=c} 1$ and note that $\sum_{c=1}^{C} \alpha_c = \sum_{c=1}^{C} \sum_{n:y_n=c} 1 = N$. Then we have the following optimization problem:

$$
\begin{aligned}
&\max_{\{\pi_c\}_{c=1}^{C}} \sum_{c=1}^{C} \alpha_c \log \pi_c \\
&\text{s.t.} \quad \sum_{c=1}^{C} \pi_c = 1. \qquad (23)
\end{aligned}
$$

If we find the Lagrangian function $g(\lambda)$ we have,

$$
g(\lambda) = \inf_{\{\pi_c\}_{c=1}^{C}} \Big[\sum_{c=1}^{C} \alpha_c \log \pi_c + \lambda(1 - \sum_{c=1}^{C} \pi_c) \Big]. \qquad (24)
$$

By taking derivatives of $\sum_{c=1}^{C} \alpha_c \log \pi_c + \lambda(1 - \sum_{c=1}^{C} \pi_c)$ w.r.t. π_c, we get

$$\begin{aligned} &\frac{\partial \sum_{c=1}^{C} \alpha_c \log \pi_c + \lambda(1 - \sum_{c=1}^{C} \pi_c)}{\partial \pi_c} = 0, \\ \Longrightarrow &\frac{\alpha_c}{\pi_c^*} - \lambda = 0, \\ \Longrightarrow &\pi_c^* = \frac{\alpha_c}{\lambda}. \end{aligned} \tag{25}$$

Further, π_c^* should satisfy the condition of the optimization problem, that is,

$$\sum_{c=1}^{C} \pi_c^* = 1, \quad \Longrightarrow \quad \sum_{c=1}^{C} \frac{\alpha_c}{\lambda} = 1, \quad \Longrightarrow \quad \lambda = \sum_{c=1}^{C} \alpha_c, \quad \Longrightarrow \quad \lambda = N. \tag{26}$$

Therefore, we get $\pi_c^* = \frac{\alpha_c}{\lambda} = \frac{\alpha_c}{N} = \frac{\sum_{n:y_n=c} 1}{N}$.

- **Finding μ_{cd}^*:** Note that the first term of (22) does not depend on μ_{cd}. Hence,

$$\max_{\boldsymbol{\mu}} \log P(\mathcal{D}) \quad \Leftrightarrow \quad \min_{\boldsymbol{\mu}} \sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} (x_{nd} - \mu_{cd})^2. \tag{27}$$

By taking the derivatives w.r.t. μ_{cd}, and set it to zero, we get

$$\sum_{n:y_n=c} -2(x_{nd} - \mu_{cd}^*) = 0, \quad \Longrightarrow \quad \mu_{cd}^* = \frac{\sum_{n:y_n=c} x_{nd}}{\sum_{n:y_n=c} 1}. \tag{28}$$

- **Finding σ_{cd}^*:** Note that the first term of (22) does not depend on σ_{cd}^2. Hence,

$$\max_{\boldsymbol{\sigma}} \log P(\mathcal{D}) \quad \Leftrightarrow \quad \min_{\boldsymbol{\sigma}} \sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} [+\frac{1}{2} \log 2\pi\sigma_{cd}^2 + \frac{1}{2\sigma_{cd}^2}(x_{nd} - \mu_{cd})^2]. \tag{29}$$

Since $\sum_{c=1}^{C} \sum_{n:y_n=c} \sum_{d=1}^{D} [+\frac{1}{2} \log 2\pi\sigma_{cd}^2 + \frac{1}{2\sigma_{cd}^2}(x_{nd} - \mu_{cd})^2]$ is a function of only σ_{cd}^2 (and not σ_{cd} directly), we can take the derivatives w.r.t. σ_{cd}^2, and set it to zero to get

$$\sum_{n:y_n=c} [\frac{1}{2\sigma_{cd}^{*2}} - \frac{(x_{nd} - \mu_{cd}^*)^2}{2\sigma_{cd}^{*4}}] = 0, \quad \Longrightarrow \quad \sigma_{cd}^{*2} = \frac{\sum_{n:y_n=c} (x_{nd} - \mu_{cd}^*)^2}{\sum_{n:y_n=c} 1}. \tag{30}$$