The Australian National University Semester 1, 2022
School of Computing Assignment 1
Group member name(s): Ziyang Chen, Han Zhang
Group member UID(s): u6908560, u7235649

COMP4670/8600: Statistical Machine Learning

Contribution statement:
Ziyang Chen (u6908560) mainly contributed on Section 1 and Section 2.
Han Zhang (u7235649) mainly contributed on Section 1 and Section 3.

Answer to Question 1.1

The class belongs to Program A.

Answer to Question 1.2

Because classes' number and students' number are the same in Program A and Program B, they will not affect the result of the probability.

We set the number of classes is 10 and each class has 100 students. Therefore, we can get:

$$P(A) = \frac{C_{650}^{55} C_{350}^{45}}{C_{1000}^{100}} \quad and \quad P(B) = \frac{C_{450}^{55} C_{550}^{45}}{C_{1000}^{100}} \tag{1}$$

Comparing the ratio of these likelihoods, we can get:

$$\begin{aligned} \frac{P(A)}{P(B)} = \frac{C_{650}^{55} C_{350}^{45}}{C_{450}^{55} C_{550}^{45}} = \frac{\frac{650!}{55! \times 595!} \times \frac{350!}{45! \times 305!}}{\frac{450!}{55! \times 395!} \times \frac{550!}{45! \times 405!}} &= \frac{650! \times 350! \times 395! \times 405!}{595! \times 305! \times 450! \times 550!} \\ &= \frac{650 \times 649 \times ... \times 596 \times 350 \times 349 \times ... \times 306}{450 \times 449 \times ... \times 396 \times 550 \times 549 \times ... \times 406} \end{aligned} \tag{2}$$

It is easy to get $\frac{P(A)}{P(B)} < 1$, thus, $P(A) < P(B)$, which means Program B is more likely.

Answer to Question 1.3

Although 55% (chosen) is equidistant between 45% (Program B) and 65% (Program A), but we can get the variance of Program A and Program B.

Because the situation satisfies the binomial distribution, we can get the following equations when we set the student number is $n(n > 0)$.

$$Var(A) = 0.65 \times 0.35 \times n = 0.2275n \quad and \quad Var(B) = 0.45 \times 0.55 \times n = 0.2475n \tag{3}$$

Because $Var(B) > Var(A)$, Program B is more likely than Program A.

Answer to Question 2.1

According to the given equation (2.1) in spec, we can get:

$$\begin{aligned} q(x|\boldsymbol{\eta}) &= \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta})) \\ &= \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) \exp(-\psi(\boldsymbol{\eta})) \\ &= \frac{1}{\exp(\psi(\boldsymbol{\eta}))} \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) \end{aligned} \tag{4}$$

Therefore, changing the equation 4, we can get:

$$q(x|\boldsymbol{\eta})\exp(\psi(\boldsymbol{\eta})) = \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) \tag{5}$$

Integrating both sides of the equation 5, we can have:

$$\int q(x|\boldsymbol{\eta})\exp(\psi(\boldsymbol{\eta}))dx = \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x))dx \tag{6}$$

Thus,

$$\exp(\psi(\boldsymbol{\eta}))\int q(x|\boldsymbol{\eta})dx = \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x))dx \tag{7}$$

According the given equation (2.2) in spec, we can get:

$$exp(\psi(\boldsymbol{\eta})) = \exp(log\int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x))dx) = \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x))dx \tag{8}$$

Combining equation 7 and equation 8, we can get:

$$\int q(x|\boldsymbol{\eta})dx = 1 \tag{9}$$

Therefore, it can prove $q(x|\boldsymbol{\eta})$ is a valid probability density function.

Answer to Question 2.2

Because $\mathcal{N}$ is 1-dimensional Gaussian distribution with μ and σ, we can get:

$$\begin{aligned}
\mathcal{N}(\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\
&= \exp(\log(2\pi\sigma^2)^{-\frac{1}{2}})\exp(-\frac{1}{2\sigma^2}(x^2-2x\mu+\mu^2)) \\
&= \exp(-\frac{1}{\sigma^2}(x^2-2x\mu) - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)) \\
&= \exp(\begin{pmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{pmatrix}\begin{pmatrix} x \\ x^2 \end{pmatrix} - (\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2))
\end{aligned} \tag{10}$$

comparing with equation (2.1) $q(x|\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}))$ in spec, we can see:

$$\hat{\boldsymbol{u}} = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \quad \hat{\boldsymbol{\eta}} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \tag{11}$$

Answer to Question 2.3

Answer to Question 2.4

The KL-divergence can be expressed into:

$$
\begin{aligned}
D_{KL}[\bar{q}(x) : q(x|\boldsymbol{\eta})] &= \int \bar{q}(x) \log(\frac{\bar{q}(x)}{q(x|\boldsymbol{\eta})}) dx \\
&= \int \bar{q}(x)(\log \bar{q}(x) - \log q(x|\boldsymbol{\eta})) dx \\
&= \int \bar{q}(x) \log \bar{q}(x) dx - \int \bar{q}(x) \log q(x|\boldsymbol{\eta}) dx
\end{aligned} \tag{12}
$$

The first part $\int \bar{q}(x) \log \bar{q}(x) dx$ in equation 12 is not relevant to $\boldsymbol{\eta}$, so it can be ignored. Therefore, the minimisation of $D_{KL}[\bar{q}(x) : q(x|\boldsymbol{\eta})]$ can be seen as minimising the second part $-\int \bar{q}(x) \log q(x|\boldsymbol{\eta}) dx$ in equation 12.

Putting equation (2.6) in spec in the second part, we can get:

$$
-\int \bar{q}(x) \log q(x|\boldsymbol{\eta}) dx = -\frac{1}{N} \sum_{i=1}^{N} \int \delta(x - x_i) \log q(x|\boldsymbol{\eta}) dx \tag{13}
$$

From the hint in the spec, we can know $\int \delta(x - a) \cdot f(x) dx = f(a)$. Thus, we can write the equation 13 into:

$$
-\int \bar{q}(x) \log q(x|\boldsymbol{\eta}) dx = -\frac{1}{N} \sum_{i=1}^{N} \int \delta(x - x_i) \log q(x|\boldsymbol{\eta}) dx = -\frac{1}{N} \sum_{i=1}^{N} \log q(x_i|\boldsymbol{\eta}) = -\ell(\boldsymbol{\eta}) \tag{14}
$$

So, we just need to get the minimisation of the negation of MLE ($-\ell(\boldsymbol{\eta})$). In the same meaning, we just need to get the maximization of MLE ($\ell(\boldsymbol{\eta})$).

Answer to Question 2.5

According to the equation (2.2) and equation (2.7) in spec, we can get:

$$
\begin{aligned}
\lambda = \mathbb{E}[\boldsymbol{u}(x)] = \nabla \psi(\boldsymbol{\eta}) &= \frac{d\psi(\boldsymbol{\eta})}{d\boldsymbol{\eta}} (\log \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) dx) \\
&= \frac{\int \boldsymbol{u}(x) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) dx}{\int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) dx} \\
&= \int \boldsymbol{u}(x) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta})) dx
\end{aligned} \tag{15}
$$

Using the equation (2.5) and equation (2.4) in spec, we can get the KL-divergence of distributions

$\text{EXP}(\boldsymbol{u}, \boldsymbol{\eta}_1)$ and $\text{EXP}(\boldsymbol{u}, \boldsymbol{\eta}_2)$ within the same exponential family.

$$\begin{aligned}
D_{KL}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] &= D_{KL}[q(x|\boldsymbol{\eta}_1) : q(x|\boldsymbol{\eta}_2)] \\
&= \int q(x|\boldsymbol{\eta}_1) \log(\frac{q(x|\boldsymbol{\eta}_1)}{q(x|\boldsymbol{\eta}_2)}) dx \\
&= \int q(x|\boldsymbol{\eta}_1)(\log q(x|\boldsymbol{\eta}_1) - \log q(x|\boldsymbol{\eta}_2)) dx \\
&= \int \exp(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1))(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1) - \boldsymbol{\eta}_2^\top \boldsymbol{u}(x) + \psi(\boldsymbol{\eta}_2)) dx \\
&= \int \exp(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1))(\psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) + (\boldsymbol{\eta}_1^\top - \boldsymbol{\eta}_2^\top)\boldsymbol{u}(x)) dx \\
&= (\psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1)) \int \exp(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1)) dx + ... \\
&\quad (\boldsymbol{\eta}_1^\top - \boldsymbol{\eta}_2^\top) \int \boldsymbol{u}(x) \exp(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1)) dx
\end{aligned} \tag{16}$$

According to the equation 9 and equation 15, we can get $\int \exp(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1)) dx = \int q(x|\boldsymbol{\eta}_1) dx = 1$ and $\int \boldsymbol{u}(x) \exp(\boldsymbol{\eta}_1^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}_1)) dx = \lambda_1$. Thus,

$$D_{KL}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) + (\boldsymbol{\eta}_1^\top - \boldsymbol{\eta}_2^\top)\lambda_1 = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) \tag{17}$$

Answer to Question 2.6

According to the proof in Question 2.5, we have got the solution:

$$D_{KL}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) \tag{18}$$

Therefore, we can get:

$$a^2 = D_{KL}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) \tag{19}$$

$$b^2 = D_{KL}[\boldsymbol{\eta}_2 : \boldsymbol{\eta}_3] = \psi(\boldsymbol{\eta}_3) - \psi(\boldsymbol{\eta}_2) - \lambda_2^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_2) \tag{20}$$

$$c^2 = D_{KL}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_3] = \psi(\boldsymbol{\eta}_3) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_1) \tag{21}$$

When $a^2 + b^2 = c^2$, we can get:

$$\psi(\boldsymbol{\eta}_3) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_1) = \psi(\boldsymbol{\eta}_3) - \psi(\boldsymbol{\eta}_2) + \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \lambda_1^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) - \lambda_2^T(\boldsymbol{\eta}_3 - \boldsymbol{\eta}_2) \tag{22}$$

Thus, simplifying equation 22

$$(\lambda_1 - \lambda_2)^T(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3) = 0 \tag{23}$$

So, $a^2 + b^2 = c^2$ iff the difference in natural parameters $(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3)$ is perpendicular to the difference in expectation parameters $(\lambda_1 - \lambda_2)$.

Answer to Question 3.1

The expectation in EMM EM:

$$
\begin{aligned}
\sum_Z p(Z|X,\vartheta^{old}) \log p(X,Z|\vartheta) &= \sum_Z p(Z|X,\vartheta^{old}) \log[p(X|Z,\vartheta)p(Z|\vartheta)] \\
&= \sum_Z p(Z|X,\vartheta^{old}) \log[\prod_{n=1}^{N}\prod_{k=1}^{K}(q(x_n|\boldsymbol{\eta}_k))^{Z_{nk}}\pi_k^{z_{nk}}] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_Z p(Z|X,\vartheta^{old}) z_{nk}(\log \pi_k + \log q(x_n|\boldsymbol{\eta}_k)) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_Z \frac{p(X|Z,\vartheta^{old})p(Z|\vartheta^{old})}{p(X|\vartheta^{old})} z_{nk}(\log \pi_k + \log q(x_n|\boldsymbol{\eta}_k)) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_Z \frac{z_{nk}\prod_{k=1}^{K}(q(x_n|\boldsymbol{\eta}_k^{old}))^{z_{nk}}(\pi_k^{old})^{z_{nk}}}{\sum_j \pi_j^{old} q(x_n|\boldsymbol{\eta}_j^{old})}(\log \pi_k + \log q(x_n|\boldsymbol{\eta}_k)) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_Z \frac{z_{nk}\prod_{k=1}^{K}[\pi_k^{old} q(x_n|\boldsymbol{\eta}_k^{old})]^{z_{nk}}}{\sum_j \pi_j^{old} q(x_n|\boldsymbol{\eta}_j^{old})}(\log \pi_k + \log q(x_n|\boldsymbol{\eta}_k)) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\pi_k^{old} q(x_n|\boldsymbol{\eta}_k^{old})}{\sum_j \pi_j^{old} q(x_n|\boldsymbol{\eta}_j^{old})}(\log \pi_k + \log q(x_n|\boldsymbol{\eta}_k)) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma^{old}(z_{nk})(\log \pi_k + \log q(x_n|\boldsymbol{\eta}_k))
\end{aligned} \tag{24}
$$

Answer to Question 3.2

According to equation (B.1) in the spec,

$$
\begin{aligned}
\log \ell(\vartheta) &= \log p(X|\vartheta) \\
&= \log(\sum_Z p(X,Z|\vartheta))
\end{aligned} \tag{25}
$$

Apply Lagrange multipliers to it and let it be 0, we have:

$$
\begin{aligned}
0 &= \log(\sum_Z p(X,Z|\vartheta)) + \sigma(\sum_{k=1}^{K} - 1) \\
&= \sum_Z \log(p(X|Z,\vartheta)p(Z|\vartheta)) + \sigma(\sum_{k=1}^{K} - 1)
\end{aligned} \tag{26}
$$

According to Question 3.1,

$$
\begin{aligned}
0 &= \sum_{Z} \log(p(X|Z,\vartheta)p(Z|\vartheta)) + \sigma(\sum_{k=1}^{K} -1) \\
&= \sum_{Z} \log(\prod_{n=1}^{N}\prod_{k=1}^{K}(q(x_n|\boldsymbol{\eta}_k))\pi_k)^{z_{nk}} + \sigma(\sum_{k=1}^{K} -1) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk}) + \sigma\sum_{k=1}^{K} -\sigma \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\pi_k q(x_n|\boldsymbol{\eta}_k)}{\sum_j \pi_j q(x_n|\boldsymbol{\eta}_j)} + \sigma\sum_{k=1}^{K}\pi_k - \sigma \\
&= \sum_{k=1}^{K}(\sum_{n=1}^{N}\frac{\pi_k q(x_n|\boldsymbol{\eta}_k)}{\sum_j \pi_j q(x_n|\boldsymbol{\eta}_j)} + \sigma\pi_k) - \sigma
\end{aligned}
\tag{27}
$$

Set the derivative with respect to π_k to 0, we get:

$$
0 = \sum_{k=1}^{K}(\sum_{n=1}^{N}\frac{q(x_n|\boldsymbol{\eta}_k)}{\sum_j \pi_j q(x_n|\boldsymbol{\eta}_j)} + \sigma)
\tag{28}
$$

Multiply π_k on each side and according to the constraint $\sum_k \pi_k = 1$ we get:

$$
\begin{aligned}
0 &= \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\pi_k q(x_n|\boldsymbol{\eta}_k)}{\sum_j \pi_j q(x_n|\boldsymbol{\eta}_j)} + \sigma \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk}) + \sigma \\
&= \sum_{k=1}^{K} N_k + \sigma \\
&= N + \sigma
\end{aligned}
\tag{29}
$$

So $\sigma = -N$. Thus

$$
\begin{aligned}
0 &= \sum_{k=1}^{K} N_k + \sigma \\
&= \sum_{k=1}^{K} \pi_k N_k + \sigma\pi_k \\
&= N_k - N\pi_k,
\end{aligned}
\tag{30}
$$

$$
\pi_k = \frac{N_k}{N}
\tag{31}
$$

$\eta_k = \nabla\varphi(\lambda_k)$, according to equation (2.7) in the spec,

$$
\begin{aligned}
\lambda \overset{def}{=} \underset{x_n \sim \textsc{Exp}(\boldsymbol{u},\eta)}{\mathbb{E}}[u(x_n)] = \nabla\psi(\boldsymbol{\eta}) \\
&= \frac{d\psi(\boldsymbol{\eta})}{d\boldsymbol{\eta}}(\log \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x_n))dx_n) \\
&= \int \boldsymbol{u}(x_n)\exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x_n) - \psi(\boldsymbol{\eta}))dx_n \\
&= \int \boldsymbol{u}(x_n) q(x_n|\boldsymbol{\eta})dx_n
\end{aligned} \tag{32}
$$

According to equation (2.1) in the spec and Question 2.5,

$$
\begin{aligned}
\lambda &= \frac{1}{\sum_{n=1}^{N}\gamma^{old}(z_{nk})}\sum_{n=1}^{N}(\boldsymbol{u}(x_n)\gamma^{old}(z_{nk})) \\
&= \frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{u}(x_n)\gamma^{old}(z_{nk}))
\end{aligned} \tag{33}
$$

Answer to Question 3.3

The *weighted_probs(), e_step_EMM(), m_step_EMM() in emm_question.py have been implemented and submitted. The visualisation result have been shown in implementation_viewer.ipynb and it has been submitted as well.*

Answer to Question 3.4

Answer to Question 3.5

Answer to Question 3.6

Answer to Question 3.7

The *single_EM_iter_blur() in blr_question.py have been implemented and submitted. The visualisation result have been shown in implementation_viewer.ipynb and it has been submitted as well.*