ML2024 Fall Assignment 5

Wei-Chen Chang, R12227118

Due: Dec, 20 2024

Problem 1

1. Let $x_1, ..., x_4$ be $(1, 2, 3), (3, -2, 2), (-2, -1, -4), (-2, 1, -1)$, hence $n = 4$, $m = 3$. And the covariance matrix is
$$\begin{pmatrix} \frac{18}{4} & -1 & \frac{19}{4} \\ -1 & \frac{10}{4} & \frac{5}{4} \\ \frac{19}{4} & \frac{5}{4} & \frac{30}{4} \end{pmatrix}$$

2. Since Σ is symmetric, we have $\Sigma = U\Lambda U^T$ with $U = \begin{pmatrix} u_1 & u_2 & ... & u_m \end{pmatrix} \in \mathbb{R}^{m\times m}, \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_m) \in \mathbb{R}^{m\times m}$, where u_i, λ_i are the eigenvectors and the corresponding eigenvalues of Σ. And $\lambda_i \geq 0\ \forall i \in \{1, 2, ..., m\}$ since Σ is positive semi-definite.

   Let $Z = U^T\Phi = \begin{pmatrix} z_1 & z_2 & ... & z_k \end{pmatrix} \in \mathbb{R}^{m\times k}$, rewrite the target function as:
$$\begin{aligned} Tr(\Phi^T\Sigma\Phi) &= Tr(\Phi^T U\Lambda U^T\Phi) = Tr(Z^T\Lambda Z) \\ &= Tr\begin{pmatrix} z_1^T\Lambda z_1 & z_1^T\Lambda z_2 & \cdots & z_1^T\Lambda z_k \\ z_2^T\Lambda z_1 & z_2^T\Lambda z_2 & \cdots & z_2^T\Lambda z_k \\ \vdots & \vdots & \ddots & \vdots \\ z_k^T\Lambda z_1 & z_k^T\Lambda z_2 & \cdots & z_k^T\Lambda z_k \end{pmatrix} \\ &= \sum_{i=1}^{k} z_i^T\Lambda z_i = \sum_{i=1}^{k}\lambda_i\|z_i\|_2^2 \\ &= \sum_{i=1}^{k}\lambda_i. \end{aligned}$$

   Note that $\|z_i\|_2^2$ since U is orthogonal, $Z^TZ = \Phi^TUU^T\Phi = I_k$.

   Hence, to attain minimum, one can arrange the eigenvalues λ_i in a non-descending order, $(0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_m)$, and let Φ collect first k eigenvectors as column vectors:
$$\Phi = \begin{pmatrix} u_1 & u_2 & ... & u_k \end{pmatrix}.$$

   Now $Tr(\Phi^T\Sigma\Phi) = \sum_{i=1}^{k}\lambda_i$ is the sum from the first to the k-th smallest eigenvalues of Σ, which is the minimal value.

Problem 2

Let $g_t = \sum_{k=1}^{t} \alpha_k f_k$. In Gradient Boosting, we update g by $g_{t+1} = g_t + \alpha_t f_t$, where:

$$f_t \in \operatorname*{argmin}_{f \in F} \frac{\partial}{\partial \alpha} L(g_t + \alpha f)\Big|_{\alpha=0}, \ \alpha_t \in \operatorname*{argmin}_{\alpha \in \mathbb{R}} L(g_t + \alpha f_t),$$

with the loss function $L(g) = \sum_{i=1}^{N} \log\left(1 + e^{-y_i g(x_i)}\right)$.

For f_t, we aim to minimize $\frac{\partial}{\partial \alpha} L(g_t + \alpha f)\Big|_{\alpha=0}$, which is

$$\begin{aligned}
\frac{\partial}{\partial \alpha} L(g_t + \alpha f)\Big|_{\alpha=0} &= \sum_{i=1}^{N} \frac{1}{1 + \exp\left(y_i(g_t(x_i) + \alpha f(x_i))\right)} \cdot -y_i f(x_i)\Big|_{\alpha=0} \\
&= \sum_{i=1}^{N} \frac{1}{1 + \exp\left(y_i g_t(x_i)\right)} \cdot -y_i f(x_i) \\
&= Z_t \mathbb{E}_{i \sim D_t}[-y_i f(x_i)] \\
&= Z_t \mathbb{E}_{i \sim D_t}[2\mathbf{1}_{y_i \neq f(x_i)} - 1]
\end{aligned}$$

where $Z_t = \sum_{i=1}^{N} \frac{1}{1+\exp(y_i g_t(x_i))}$, D_t is a r.v. with density $D_t(i) = \frac{1}{1+\exp(y_i g_t(x_i))}/Z_t$ for $i = 1, ..., N$. Then

$$f_t \in \operatorname*{argmin}_{f \in F} Z_t \mathbb{E}_{i \sim D_t}[2\mathbf{1}_{y_i \neq f(x_i)} - 1] = \operatorname*{argmin}_{f \in F} \mathbb{P}(y_i \neq f(x_i)),$$

hence the optimal f_t minimize the error rate weighted by D_t.

To optimize over α_t, we set the partial derivative of $L(g_t + \alpha f_t)$ to 0:

$$\frac{\partial}{\partial \alpha} L(g_t + \alpha f_t) = \sum_{i=1}^{N} \frac{1}{1 + \exp\left(y_i(g_t(x_i) + \alpha f_t(x_i))\right)} \cdot -y_i f_t(x_i) = 0$$

To my knowledge, α has no closed form solution, one can solve it by numerical methods and then update it as α_t.

Problem 3

Let z be the latent variable indicates the cluster of the sample, and assume N samples are independent, the log-likelihood can be expressed as:

$$\begin{aligned}
\log p(x_1, ..., x_N; \theta) &= \sum_{i=1}^{N} \sum_{k=1}^{K} p\left(z = k \middle| X = x_i; \theta^{(t)}\right) \log p\left(X = x_i \middle| z = k, \theta\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} p\left(z = k \middle| X = x_i; \theta^{(t)}\right) \left(\log p(X = x_i, z = k; \theta) - \log p(z = k, \theta)\right) \\
&= Q(\theta || \theta^{(t)}) - \sum_{i=1}^{N} \sum_{k=1}^{K} p\left(z = k \middle| X = x_i; \theta^{(t)}\right) \log p(z = k, \theta).
\end{aligned}$$

For the E-step, we derive $Q(\theta||\theta^{(t)})$:

$$\begin{aligned}
Q(\theta||\theta^{(t)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} p\left(z=k\middle|\, X=x_i;\theta^{(t)}\right)\log p(X=x_i,z=k;\theta) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{p(X=x_i,z=k;\theta^{(t)})}{p(X=x_i;\theta^{(t)})}\log p(X=x_i,z=k;\theta) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(t)}(\log(\frac{\pi_k}{\tau_k})-\frac{x_i}{\tau_k}),
\end{aligned}$$

where

$$\delta_{ik}^{(t)} = \frac{\pi_k^{(t)}/\tau_k^{(t)}\cdot e^{(-x_i/\tau_k^{(t)})}}{\sum_{k=1}^{K}\pi_k^{(t)}/\tau_k^{(t)}\cdot e^{(-x_i/\tau_k^{(t)})}}.$$

For M-step, we need to maximize $Q(\theta||\theta^{(t)})$ w.r.t. τ_k,π_k. First we take the partial derivative over τ_k:

$$\frac{\partial}{\partial\tau_k}Q(\theta||\theta^{(t)}) = \sum_{i=1}^{N}\delta_{ik}^{(t)}(\frac{x_i-\tau_k}{\tau_k^2}).$$

Setting it to zero, and rearranging terms, we update τ_k as

$$\tau_k^{(t+1)} = \frac{\sum_{i=1}^{N}\delta_{ik}^{(t)}x_i}{\sum_{i=1}^{N}\delta_{ik}^{(t)}}$$

For π_k, we introduce the Lagrange multiplier λ and set its partial derivative over π_k

$$\frac{\partial}{\partial\pi_k}Q(\theta||\theta^{(t)}) + \lambda\sum_{k=1}^{K}\pi_k = \sum_{i=1}^{N}\delta_{ik}^{(t)}(\frac{1}{\pi_k})-\lambda.$$

Set it to zero, we have:

$$\pi_k = \frac{1}{\lambda}\sum_{i=1}^{N}\delta_{ik}^{(t)}.$$

Note that since $\sum_{k=1}^{K}\pi_k = 1$ and $\sum_{k=1}^{K}\delta_{ik}^{(t)} = 1$, we have $\frac{1}{\lambda}\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}^{(t)} = 1$. This implies $\lambda = N$, hence we update π_k as:

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\delta_{ik}^{(t)}.$$

Problem 4

Problem 5

摘要: 今天演講請來 prof. Morris Chang 介紹在機器學習中如何保護資料或者演算法的隱私的相關議題。一開始先介紹了在資訊時代中隱私問題與相關的事件的歷史介紹，提到如 Cambridge Analytical Scandal 事件。之後引出研究團隊的關注的問題：如何在機器學習裡，保護資料可能包含的個人資訊卻又能保持模型好的表現，並且簡單介紹機器學習領域中幾種可能的隱私遭冒犯的問題。之後也介紹在保護資料的大致想法：是在資料處理過程中，加上雜訊使得

外來攻擊者無法透過模型回推找出原先資料的原始樣貌，並且從中獲取到個人隱私資料。最合則總結講者對於相關領域的願景：大家能有一個平台可以分享資料提供模型訓練，但這些資料又有良好的保護機制，不會被有心人士獲取以冒犯個人隱私。

心得： 本次演講讓我接觸到機器學習應用上可能遭遇到的隱私外洩問題，與一般在只在思考模型架構、訓練和運用在分類/迴歸問題上的技術問題相比，是很不一樣的觀點。而對於隱私的解方，講者提出加上雜訊處理這件簡單的想法，也是很好理解很直觀的想法。最後 QA 有人詢問在 LLM 當道的時代中，模型的隱私問題如何被解決也讓我對相關議題有了新的想法，而講者提到的最終願景也讓我想到開放科學 (open science) 的運動，在提倡資料開放的過程中，今天提到的隱私問題應該也是需要被考量的。

Problem 6

1. Recall that $0 \le r_t \le 1, 0 \le \gamma < 1\ \forall t = 0, 1, 2, ...$, hence $\forall \pi \in \Pi$,

$$0 \le V^\pi(s) \le V^*(s) \le \sum_{t=0}^{\infty} \gamma^t r_t = \frac{1}{1-\gamma},$$

$$0 \le Q^\pi(s) \le Q^*(s) \le \sum_{t=0}^{\infty} \gamma^t r_t = \frac{1}{1-\gamma}.$$

2. For all policy $\pi \in \Pi$, we define a policy $\pi_{s,a,r}(\cdot|\tau) = \pi(\cdot|s, a, r, \tau)$ given the trajectory τ.

   First we claim that $\{\pi_{s,a,r} : \pi \in \Pi\} = \Pi$:

   proof. Since all $\pi_{s,a,r} \in \Pi$, $\{\pi_{s,a,r} : \pi \in \Pi\} \subseteq \Pi$. Next, for every $\pi^* \in \Pi$, we can construct $\pi(\cdot|s, a, r, \tau) = \pi^*(\cdot|\tau) = \pi_{s,a,r}(\cdot|\tau) \in \{\pi_{s,a,r} : \pi \in \Pi\}$, $\Pi \subseteq \{\pi_{s,a,r} : \pi \in \Pi\}$. Hence $\Pi = \{\pi_{s,a,r} : \pi \in \Pi\}$. $\square$

   Next, by Markov's property, we have:

$$E\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, (s_0, a_0, r_0, s_1) = (s, a, r, s')\right] = \gamma E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi_{s,a,r}, s'\right] = \gamma V^{\pi_{s,a,r}}(s').$$

   Hence,

$$\sup_{\pi\in\Pi} E\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, (s_0, a_0, r_0, s_1) = (s, a, r, s')\right] = \gamma \sup_{\pi\in\Pi} V^{\pi_{s,a,r}}(s') = \gamma \sup_{\pi\in\Pi} V^{\pi}(s') = \gamma V^*(s').$$

3. (a) Given a current state s, action $\pi^*(s)$ always selects the option that maximizes the value. In the case of a tie, a deterministic manner can still be applied. This decision does not depend on the past trajectory and is not probabilistic, making it deterministic.

   (b) Since state transition is now deterministic and by (3a) we know the action $\pi^*(s)$ is also deterministic. Let the state transition be $s_{t+1} = s'(s_t, a_t)$, where $s' : (S, A) \to S$, we can rewrite $V^{\pi^*}(s)$:

$$V^{\pi^*}(s_0 = s) = \sum_{t=0}^{\infty} \gamma^t r(s_t, \pi^*(s_t)),$$

   with

$$\pi^*(s) \in \operatorname*{argmax}_{a\in A} r(s, a) + \gamma V^*(s'(s, a)),$$

or,
$$r(s,\pi^*(s)) + \gamma V^*[s'(s,\pi^*(s))] = \sup_{\pi\in\Pi} r(s,\pi(s)) + \gamma V^*[s'(s,\pi(s))].$$

To show $V^* \le V^{\pi^*}$, note that

$$\begin{aligned}
V^*(s) &= \sup_{\pi\in\Pi} E\left[r(s,a_0) + \sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\middle|\, \pi\right] \\
&= \sup_{\pi\in\Pi} E_{a\sim\pi(\cdot|s)}\left[r(s,a) + E\left[\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\middle|\, \pi, (s_0,a_0,r_0,s_1) = (s,a,r,s'(s,a))\right]\middle|\, \pi\right] \\
&\le \sup_{\pi\in\Pi} E_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma V^*(s'(s,a))\middle|\, \pi\right] \\
&= r(s,\pi^*(s)) + \gamma V^*\left(s'(s,\pi^*(s))\right).
\end{aligned}$$

By recursion, we have:

$$\begin{aligned}
V^*(s_0) &\le r(s_0,\pi^*(s_0)) + \gamma V^*(s_1) \\
&\le r(s_0,\pi^*(s_0)) + \gamma r(s_1,\pi^*(s_1)) + \gamma^2 V^*(s_2) \\
&\le r(s_0,\pi^*(s_0)) + \gamma r(s_1,\pi^*(s_1)) + \gamma^2 r(s_2,\pi^*(s_2)) + \ldots
\end{aligned}$$

Denote $X_t = \sum_{k=1}^{t}\gamma^k r(s_k,\pi^*(s_k)) + \gamma^t V^*(s_t)$, and $X_\infty := \lim_{k\to\infty} X_t = \sum_{k=1}^{\infty}\gamma^k r(s_k,\pi^*(s_k))$, then
$$V^*(s_0) \le X_\infty = V^{\pi^*}(s_0) < \infty$$
because $V^\pi(s)$ is finite $\forall s \in S, \pi \in \Pi$. Hence $V^* \le V^{\pi^*}$.
Also since $V^* \ge V^{\pi^*}$, we have $V^* = V^{\pi^*}$.

(c) For all $s \in S, a \in A$, we have:

$$\begin{aligned}
Q^*(s,a) &= \sup_{\pi\in\Pi} E\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, \pi, (s_0,a_0) = (s,a)\right] \\
&= r(s,a) + \sup_{\pi\in\Pi} E\left[\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\,\middle|\, \pi, (s_0,a_0,r_0,s_1) = (s,a,r(s,a),s'(s,a))\right] \\
&= r(s,a) + \gamma V^*(s'(s,a)) \\
&= r(s,a) + \gamma V^{\pi^*}(s'(s,a)) \quad (3b) \\
&= Q^{\pi^*}(s,a). \quad \text{(Bellman's Consistency Equation)}
\end{aligned}$$

Hence $Q^{\pi^*} = Q^*$.