Problem 1 (Trace Optimization)(1%)

1. Let $\Sigma \in R^{m\times m}$ be a symmetric positive semi-definite matrix, $\mu \in R^m$. Please construct a set of points $x_1, ..., x_n \in R^m$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = \Sigma, \quad \frac{1}{n}\sum_{i=1}^{n}x_i = \mu$$

Hint: n is not given by the problem. WLOG, you may assume $\mu = 0 \in \mathbb{R}^d$.

2. Let $1 \le k \le m$, solve the following optimization problem and justify with proof:

minimize $\quad Trace(\Phi^T\Sigma\Phi)$
subject to $\quad \Phi^T\Phi = I_k$
variables $\quad \Phi \in R^{m\times k}$

In other words, you need to find Φ and verify that your Φ minimize the trace.

1. 當 $\mu=0$ 時，$x_1, x_2, \dots x_n \in \mathbb{R}^m$，使得

$\frac{1}{n}\sum_{i=1}^{n} x_i = 0$ ， $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \Sigma$

$\Sigma = U\Lambda U^T$ $\begin{cases} U \in R^{m\times m} \ (U^TU = I_m) \\ \Lambda = diag(\lambda_1, \lambda_2, \dots \lambda_m) \text{ 是包含非負特徵值的對角矩陣} \end{cases}$

希望產生一組 $x_1, \dots x_n$ 使 sample 共變異數為 Σ

① 多變量常態分佈 $N(u,\Sigma)$，從中取無限多組獨立樣本平均必接近 μ 且 Σ

② 獨立高斯分佈 $z_i \sim N(0, I_m)$ 產生 n 組樣本 $z_1, z_2, \dots, z_n$

$x_i = U\sqrt{\Lambda}z_i$ ，$\because z_i$ 的共變異數為 I_m，則 x_i 的共變異數為 $U\Lambda U^T = \Sigma$

令 $\mu=0$ ， $x_i = U\sqrt{\Lambda}z_i$ 的樣本平均將隨著 n 越來越大而接近 0.

⇒當樣本數 n 夠大，樣本均值和樣本共變異數會近似期望值

2. 令 $\Phi' = U^T\Phi$ ，由於 U 是正交矩陣，$\Phi^T\Phi = I_k$ 等價於 $\Phi'^T\Phi' = I_k$

$Trace(\Phi^T\Sigma\Phi) = Trace(\Phi'^T\Lambda\Phi') \Rightarrow Trace(\Phi'^T\Lambda\Phi') = \sum_{j=1}^{k}\phi_j'^T\Lambda\phi_j' = \sum_{j=1}^{k}\sum_{i=1}^{m}\lambda_i(\phi_j')_i^2$

λ_i 是權重，目標是將 Φ' 的列向量 ϕ_j' 對應到 λ_i 最小的方向上

選擇 Φ 的列向量為對應於 Σ 最小的 k 個特徵值 $\lambda_1, \lambda_2, \dots, \lambda_k$ 的特徵向量

將 $\Phi = [u_1, u_2, \dots, u_k]$ $\Rightarrow Trace(\Phi^T\Sigma\Phi) = \sum_{i=1}^{k}\lambda_i$ 為最小值

Problem 2 (Gradient Boosting)(1%) 二元分類

Consider the binary classification problem, where we are given training data set $\{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$. Let $F = \{f \mid f : \mathbb{R}^d \to \{1, -1\}\}$ be the collection of classifiers. Given number of epochs $T \in \mathbb{N}$. Suppose that we want to find the function

$$g(x) = \sum_{t=1}^{N} \alpha_t f_t(x)$$

加權 弱分類器

where $f_t \in F$ and $\alpha_t \in \mathbb{R}$ for all $t = 1, \ldots, T$, by which the aggregated classifier is given by

$$h(x) = \begin{cases} 1, & \text{if } g(x) > 0 \\ -1, & \text{if } g(x) \leq 0. \end{cases}$$

Please apply gradient boosting to show how the functions f_t and the coefficients α_t are computed with an aim to minimize the following loss function

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

$g_0(x) = \arg\min_c \sum_{i=1}^{N} \log(1+e^{-y_i c})$

L(g)是對整體模型g(x)(包含所有的學習的加權組合)的表現進行評估

Iterative Steps:

For $t = 1, 2, \ldots, T$:

Compute residuals: (計算損失函數對模型預測的負梯度，作為模型的改進方向)

$$r_i^{(t)} = -\frac{\partial L(g)}{\partial g(x_i)} = \frac{y_i}{1 + e^{y_i g_{t-1}(x_i)}}$$

Fit a weak classifier $f_t(x)$:

最小平方損失：$\sum_{i=1}^{N} (r_i^{(t)} - f_t(x_i))^2$

$f_t(x) = \arg\min_{f \in F} \sum_{i=1}^{N} (r_i^{(t)} - f(x_i))^2$

weight α_t:

當前模型 新的學習器 $f(x)$

$$\alpha_t = \arg\min_{\alpha} \sum_{i=1}^{N} \log\left(1 + e^{-y_i (g_{t-1}(x_i) + \alpha f_t(x_i))}\right)$$

找一個最佳的α_t，使得在當前模型$g^{(t-1)}(x)$的基礎上，加上新的學習器$f_t(x)$的修正後，損失函數L(g)最小

Update the model

$$g_t(x) = g_{t-1}(x) + \alpha_t f_t(x)$$

迭代T次，或 another convergence criterion is satisfied.

$$g_T(x) = \sum_{t=1}^{T} \alpha_t f_t(x)$$

Final Aggregated Classifier:

$$h(x) = \begin{cases} 1 & \text{if } g_T(x) > 0, \\ -1 & \text{if } g_T(x) \leq 0 \end{cases}$$

Problem 3 (EM algorithm for mixture of exponential model)(1%)

Given N samples $x_1, \ldots, x_N \in [0, \infty)$, we would like to cluster them into K clusters. Assume the samples are generated according to Exponential mixture models

$$X \sim \sum_{j=1}^{K} \pi_j Exp(\tau_j)$$

where $\pi_1 + \ldots + \pi_K = 1$, and $Exp(\tau)$ denotes the exponential distribution with probability density function

$$f_\tau(x) = \begin{cases} (1/\tau)e^{-x/\tau} & , x \geq 0 \\ 0 & , x < 0. \end{cases}$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \tau_k)\}_{k=1}^K$.

(a) Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{(\pi_k^{(t)}, \tau_k^{(t)})\right\}_{k=1}^K$ to $\theta^{(t+1)} = \left\{(\pi_k^{(t+1)}, \tau_k^{(t+1)})\right\}_{k=1}^K$ in the following form:

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

(b) What is the closed form expression of $\delta_{ik}^{(t)}$?

(a) E-step

$$\delta_{ik}^{(t)} = P\left(Z_i = k \,\middle|\, x_i ; \theta^{(t)}\right)$$

Bayes' rule:

$$\Rightarrow \delta_{ik}^{(t)} = \frac{P(Z_i = k, x_i \mid \theta^{(t)})}{P(x_i \mid \theta^{(t)})}$$

$\because Z_i = k$ 表示 x_i 由第 k 個成分生成 $\Rightarrow P(Z_i = k, x_i \mid \theta^{(t)}) = \pi_k^{(t)} f_{\tau_k^{(t)}}(x_i)$

$$P(x_i \mid \theta^{(t)}) = \sum_{j=1}^{K} \pi_j^{(t)} f_{\tau_j^{(t)}}(x_i)$$

$f_{\tau_j}(x) = \frac{1}{\tau_j} e^{-x/\tau_j}$ 帶入:

$$\delta_{ik}^{(t)} = \frac{\pi_k^{(t)} \frac{1}{\tau_k^{(t)}} e^{-x_i/\tau_k^{(t)}}}{\sum_{j=1}^{K} \pi_j^{(t)} \frac{1}{\tau_j^{(t)}} e^{-x_i/\tau_j^{(t)}}}$$

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z \mid X, \theta^{(t)}}\left[\log p(X, Z \mid \theta)\right] = \sum_{k=1}^{K}\left[\sum_{i=1}^{N} \delta_{ik}^{(t)} \log \pi_k - \sum_{i=1}^{N} \delta_{ik}^{(t)} \log \tau_k - \sum_{i=1}^{N} \delta_{ik}^{(t)} \frac{x_i}{\tau_k}\right]$$

M-step: 最大化 $Q(\theta \mid \theta^{(t)})$

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$$

對 π_k 和 τ_k 求偏導數, 並設為 0, 另時考慮 $\sum_{k=1}^{K} \pi_k = 1$

1. updating π_k:

令 $N_k^{(t)} = \sum_{i=1}^{N} \delta_{ik}^{(t)}$, 在 $Q(\theta \mid \theta^{(t)})$ 中和 π_k 有關的式子

$$\sum_{i=1}^{N} \delta_{ik}^{(t)} \log \pi_k = N_k^{(t)} \log \pi_k$$

最大化須考慮 $\sum_{k=1}^{K} \pi_k = 1$:

$$L(\pi_1, \ldots, \pi_K, \lambda) = \sum_{k=1}^{K} N_k^{(t)} \log \pi_k + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$$

$$\frac{\partial L}{\partial \pi_k} = \frac{N_k^{(t)}}{\pi_k} - \lambda = 0 \Rightarrow \pi_k = \frac{N_k^{(t)}}{\lambda}$$

$$\sum_{k=1}^{K} \frac{N_k^{(t)}}{\lambda} = 1 \Rightarrow \lambda = \sum_{k=1}^{K} N_k^{(t)} = N,$$

$\because \sum_{k=1}^{K} N_k^{(t)} = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} = N$

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

2. updating τ_k:

含有關 τ_k 的部份 $-\sum_{i=1}^{N} \delta_{ik}^{(t)} \log \tau_k - \sum_{i=1}^{N} \delta_{ik}^{(t)} \frac{x_i}{\tau_k} = Q_k(\tau_k)$

$$\frac{dQ_k}{d\tau_k} = \frac{-\sum_{i=1}^{N} \delta_{ik}^{(t)}}{\tau_k} + \sum_{i=1}^{N} \delta_{ik}^{(t)} \frac{x_i}{\tau_k^2} = 0 \Rightarrow \frac{1}{\tau_k^2}\left(-\tau_k \sum_{i=1}^{N} \delta_{ik}^{(t)} + \sum_{i=1}^{N} \delta_{ik}^{(t)} x_i\right)$$

$$-\tau_k \sum_{i=1}^{N} \delta_{ik}^{(t)} + \sum_{i=1}^{N} \delta_{ik}^{(t)} x_i = 0$$

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

(b) from E-step: $\delta_{ik}^{(t)} = \frac{\pi_k^{(t)} f_{\tau_k^{(t)}}(x_i)}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\tau_j^{(t)}}(x_i)}$, $f_{\tau_k}(x) = \frac{1}{\tau_k} e^{-x/\tau_k}$, $x \geq 0$ 帶入

$$\Rightarrow \delta_{ik}^{(t)} = \frac{\pi_k^{(t)} \frac{1}{\tau_k^{(t)}} e^{-x_i/\tau_k^{(t)}}}{\sum_{j=1}^{K} \pi_j^{(t)} \frac{1}{\tau_j^{(t)}} e^{-x_i/\tau_j^{(t)}}}$$

Problem 4 (Sparse SVM)(2%)

Given training data of N input-output pairs $\mathscr{D} = ((x_i, y_i))_{i=1}^N$, where $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the p-norm of the vector $\alpha = (\alpha_1, ..., \alpha_N)$ that defines the weight vector $\mathbf{w}$, for some $p \geq 1$. In this question we consider the case $p = 2$, which leads to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & f(\alpha, b, \boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N C_i \xi_i \\ \text{subject to} & y_i \left(\sum_{j=1}^N \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b\right) \geq 1 - \xi_i, \ i \in \{1, ..., N\} \\ \text{variables} & b \in \mathbb{R}, \alpha_i \geq 0, \xi_i \geq 0, \ i \in \{1, ..., N\} \end{array}$$

which can be rewritten in the following primal problem:

$$\left.\begin{array}{ll} \text{minimize} & f(\alpha, b, \boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N C_i \xi_i \\ & g_{1,i}(\alpha, b, \boldsymbol{\xi}) = 1 - \xi_i - y_i \left(\sum_{j=1}^N \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b\right) \leq 0 \\ \text{subject to} & g_{2,i}(\alpha, b, \boldsymbol{\xi}) = -\alpha_i \leq 0 \\ & g_{3,i}(\alpha, b, \boldsymbol{\xi}) = -\xi_i \leq 0 \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array}\right\} \ i \in \{1, ..., N\} \quad (1)$$

as well as its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\boldsymbol{\omega}, \beta, \boldsymbol{\gamma}) = \inf_{\alpha \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} L(\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma}) \\ \text{subject to} & \omega_i \geq 0, \beta_i \geq 0, \gamma_i \geq 0, \ i \in [\![1, N]\!] \\ \text{variables} & \boldsymbol{\omega} = (\omega_1, ..., \omega_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_N) \in \mathbb{R}^N \end{array} \quad (2)$$

1. Write down the Lagrangian function $L(\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma})$ in explicit form of $\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma}$.
2. Show that the duality gap between (1) and (2) is zero.
3. Derive $\theta(\boldsymbol{\omega}, \beta, \boldsymbol{\gamma})$ in explicit form of dual variables $\boldsymbol{\omega}, \beta, \boldsymbol{\gamma}$.

1. $L(\alpha, b, \xi, w, \beta, r) = f(\alpha, b, \xi) + \sum_{i=1}^N w_i g_{1,i}(\alpha, b, \xi) + \sum_{i=1}^N \beta_i g_{2,i}(\alpha, b, \xi) + \sum_{i=1}^N r_i g_{3,i}(\alpha, b, \xi)$

$L(\alpha, b, \xi, w, \beta, r) = \underbrace{\frac{1}{2}\sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N C_i \xi_i}_{\text{原目標函式時一部分}} + \sum_{i=1}^N w_i \left[1 - \xi_i - y_i \left(\sum_{j=1}^N \alpha_j y_j x_j \cdot x_j + b\right)\right] - \sum_{i=1}^N \beta_i \alpha_i - \sum_{i=1}^N r_i \xi_i$

2. $f(\alpha, b, \xi) = \frac{1}{2}\sum_{i=1}^N \alpha_i^2 + \sum_{i=1}^N C_i \xi_i$

⇒ 是 α - 的二次可微分凸函式加上 ξ - 的線性函式；所有限制式

$1 - \xi_i - y_i \left(\sum_j \alpha_j y_j x_j \cdot x_i + b\right) \leq 0$，$-\alpha_i \leq 0$，$-\xi_i \leq 0$ ⇒ 為線性不等式，convex optimization problem

Slater：

對線性約束時凸問題，只要存在某個 (α, b, ξ) 使所有不等式嚴格成立，即保證 strong Duality

例：$\alpha_i = 0$，$b = 0$，$\xi_i = k$ ⇒ $1 - k - y_i(0) = 1 - k < 0$，$\alpha_i = 0 \Rightarrow -\alpha_i = 0$，$\xi_i = k > 0 \Rightarrow -\xi_i < 0$

$k > 1$ 即可嚴格滿足全部原始限制式

∵ 滿足 slater，此凸問題時對偶問必然不存在鴻溝 ⇒ Duality Gap = 0

3. $\theta(w, \beta, r) = \inf_{\alpha, b, \xi} L(\alpha, b, \xi, w, \beta, r)$

① 對 b 做極小化

Lagrangian b 部分：$-\sum_{i=1}^N w_i y_i b$　$\frac{\partial L}{\partial b} = 0$　$-\sum_{i=1}^N w_i y_i = 0 \Rightarrow \sum_{i=1}^N w_i y_i = 0$

② 對 ξ 做極小化

Lagrangian ξ 部分：$\sum_{i=1}^N C_i \xi_i - \sum_{i=1}^N w_i \xi_i - \sum_{i=1}^N r_i \xi_i$　$\frac{\partial L}{\partial \xi_i} = 0$　$C_i - w_i - r_i = 0 \Rightarrow C_i = w_i + r_i$

③ 對 α 做極小化

Lagrangian α 部分：$\frac{1}{2}\sum_{i=1}^N \alpha_i^2 - \sum_{i=1}^N \beta_i \alpha_i - \sum_{i=1}^N w_i y_i \left(\sum_j \alpha_j y_j x_j \cdot x_i\right)$　$\frac{\partial L}{\partial \alpha_r} = 0$　$\alpha_r - \sum_{i=1}^N w_i y_i y_r (x_i \cdot x_r) - \beta_r = 0 \Rightarrow \alpha_r = \sum_{i=1}^N w_i y_i y_r (x_i \cdot x_r) + \beta_r$

④ 得 $\theta(w, \beta, r)$ 閉式解

$\theta(w, \beta, r) = \sum_{i=1}^N w_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N w_i w_j y_i y_j (x_i \cdot x_j)$ 且同時滿足 $\sum_{i=1}^N w_i y_i = 0$，$0 \leq w_i \leq C_i$，$\beta_i \geq 0$，$r_i = C_i - w_i \geq 0$

⇒ $\theta(w, \beta, r) = \sum_{i=1}^N w_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N w_i w_j y_i y_j (x_i \cdot x_j)$ subject to $\sum_{i=1}^N w_i y_i = 0$，$0 = w_i \leq C_i$

4. Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \omega_i - \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \omega_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i \right)_+^2 \\ \text{subject to} & \sum_{i=1}^{N} \omega_i y_i = 0 \\ \text{variables} & 0 \le \omega_i \le C_i, \ i = 1, \ldots, N \end{array} \tag{3}$$

5. Suppose $(\bar{\alpha}, \bar{b}, \bar{\boldsymbol{\xi}})$ and $(\bar{\omega}, \bar{\beta}, \bar{\gamma})$ are the optimal solutions to problems (1) and (2) respectively. Denote $\bar{\mathbf{w}} = \sum_{j=1}^{N} \bar{\alpha}_j y_j \mathbf{x}_j$.

(a) Prove that

$$\bar{\alpha}_i = \max \left(\sum_{j=1}^{N} \bar{\omega}_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i, 0 \right) \ \forall i = 1, \ldots, N \tag{4}$$

(b) Prove that

$$\bar{b} = \arg\min_{b \in \mathbb{R}} \sum_{i=1}^{N} C_i \max \left(1 - y_i \left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + b \right), 0 \right), \tag{5}$$

(c) Prove that $\bar{\xi}_i = \max \left(1 - y_i \left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b} \right), 0 \right)$ for all $i = 1, \ldots, N$.

(d) Prove that

$$\left.\begin{array}{ll} \bar{\omega}_i = C_i, & \text{if } y_i \left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b} \right) < 1 \\ \bar{\omega}_i = 0, & \text{if } y_i \left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b} \right) > 1 \\ 0 \le \bar{\omega}_i \le C_i, & \text{if } y_i \left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b} \right) = 1 \end{array}\right\} \ \forall i = 1, \ldots, N$$

4. $L(\alpha, b, \xi, \omega, \beta, \gamma) = \frac{1}{2}\sum_{i=1}^{N}\alpha_i^2 + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N}\omega_i\left(1-\xi_i - y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_i\cdot x_j + b\right)\right) - \sum_{i=1}^{N}\beta_i\alpha_i - \sum_{i=1}^{N}\gamma_i\xi_i$

Dual Function $\theta(\omega) = \inf_{\alpha,b,\xi} L(\alpha, b, \xi, \omega, \beta, \gamma)$ ⇒ minimize $L(\alpha, b, \xi, \omega, \beta, \gamma)$ with respect to α, b, and ξ

① 和 ξ 相关: $\sum_{i=1}^{N} C_i\xi_i - \sum_{i=1}^{N}(\omega_i+\gamma_i)\xi_i \xrightarrow{\text{极小化}} \frac{\partial L}{\partial \xi_i} = C_i - \omega_i - \gamma_i = 0 \quad \therefore \omega_i + \gamma_i \le C_i \Rightarrow \xi_i = 0 \Rightarrow \gamma_i = C_i - \omega_i$ 且 $0 \le \omega_i \le C_i$

② 和 b 相关: $-\sum_{i=1}^{N}\omega_i y_i b \xrightarrow{\text{极小化}} \frac{\partial L}{\partial b} = -\sum_{i=1}^{N}\omega_i y_i = 0$ 约束条件 $\sum_{i=1}^{N}\omega_i y_i = 0$

③ 和 α 相关: $\frac{1}{2}\sum_{i=1}^{N}\alpha_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\omega_i\omega_j y_i y_j x_i\cdot x_j \xrightarrow{\text{极小化}} \frac{\partial L}{\partial \alpha_i} = \alpha_i - \sum_{j=1}^{N}\omega_j y_j y_i x_j = 0 \Rightarrow \alpha_i = \sum_{j=1}^{N}\omega_j y_j y_i x_j$

$\theta(\omega) = \sum_{i=1}^{N}\omega_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\omega_i\omega_j y_i y_j x_i\cdot x_j \Rightarrow \max_{\omega} \sum_{i=1}^{N}\omega_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\omega_i\omega_j y_i y_j x_i\cdot x_j$ 约束条件 $\begin{cases}\sum_{i=1}^{N}\omega_i y_i = 0 \\ 0 \le \omega_i \le C_i, \ \forall i\end{cases}$

5 (a) KKT:

$\alpha_i = \sum_{j=1}^{N}\omega_j y_j x_j\cdot x_i + \beta_i$ where $\beta_i \ge 0$ ensures $\alpha_i \ge 0$

$\Rightarrow \alpha_i = \max\left(\sum_{j=1}^{N}\omega_j y_j x_j\cdot x_i, 0\right)$

(b) 找 b 使得 $\sum_{i=1}^{N} C_i \max(1 - y_i(\bar{\omega}\cdot x_i + b), 0)$ 最小化

$\bar{b}$ 控制决策边界 $\bar{\omega}\cdot x + b = 0$，最佳 b 是最支持向量（满足 $1 - y_i(\bar{\omega}\cdot x_i + b) = 0$）正好位于分界线上的偏置

$\bar{b} = \arg\min_{b\in\mathbb{R}} \sum_{i=1}^{N} C_i \max(1 - y_i(\bar{\omega}\cdot x_i + b), 0)$

(c) 原始问题中有约束 $1 - \xi_i - y_i(\bar{\omega}\cdot x_i + b) \le 0 \Rightarrow \xi_i \ge 1 - y_i(\bar{\omega}\cdot x_i + \bar{b})$ 用时 $\xi_i \ge 0$

$\bar{\xi}_i = \max(1 - y_i(\bar{\omega}\cdot x_i + \bar{b}), 0)$

(d) KKT:

$\bar{\omega}_i = C_i$: if sample i 位于间隔内部 ($y_i(\bar{\omega}\cdot x_i + \bar{b}) < 1$) → 需满足 $\bar{\xi}_i > 0 \Rightarrow \bar{\omega}_i = C_i$

$\bar{\omega}_i = 0$: if sample i 位于分界线外部 ($y_i(\bar{\omega}\cdot x_i + \bar{b}) > 1$) 则 $\bar{\xi}_i = 0$ 且 $\bar{\omega}_i = 0$

$0 \le \bar{\omega}_i \le C_i$: if sample i 位于分界线上 ($y_i(\bar{\omega}\cdot x_i + \bar{b}) = 1$) 则 $\bar{\omega}_i$ 取值于区间 $[0, C_i]$

Problem 5 (Invited Talk)(1%)

Please share your thoughts about this talk.

本次演講探討隱私保護已成為大數據與AI領域不可忽視的議題。隨著機器學習應用的普及，如何平衡技術進步與隱私風險將是未來研究的重要方向。其風險可能為從模型推測出訓練數據，判斷特定數據是否參與模型訓練。因此，若能通過加入雜訊來隱藏每一數據樣本的貢獻，或將模型訓練過程分佈在本地端設備上，減少數據集中化的風險，也可更進一步做加密技術，保護數據處理過程中的機密性。在未來，我想可以結合隱私保護技術與模型壓縮技術，在降低隱私風險的同時提升計算效率，這對於隱私保護機器學習的應用有重大突破。

Problem 6 (Bellman Optimality Equations)(2%)

In this problem, we aim to help students review and understand the proof of the Bellman optimality equations.

Theorem. *Let* Π *be the set of nonstationary and randomized policies. Define*

$$V^*(s) = \sup_{\pi \in \Pi} V^\pi(s)$$
$$Q^*(s,a) = \sup_{\pi \in \Pi} Q^\pi(s,a).$$

Then there exists a stationary and deterministic policy π such that for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$V^\pi(s) = V^*(s)$$
$$Q^\pi(s,a) = Q^*(s,a).$$

Remark. The notations are consistent with those in the lecture notes.

1. Verify that V^* and Q^* is bounded between 0 and $\frac{1}{1-\gamma}$. Hence, V^* and Q^* must be finite.
2. Show that given $(s_0, a_0, r_0, s_1) = (s, a, r, s')$, the optimal discounted value γV^*, from $t = 1$ onwards, does not depend on the initial conditions s, a, and r:

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

3. Let π^* be a policy such that

$$\forall\, s \in \mathcal{S}, \quad \pi^*(s) \in \arg\max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')].$$

(a) Explain that π^* is deterministic.

(b) Now suppose that the transition of states and actions is deterministic. In order to show that π^* is an optimal policy, i.e. $V^*(s) = V^{\pi^*}(s)$, we have to show two inequalities: $V^* \geq V^{\pi^*}$ and $V^* \leq V^{\pi^*}$. The first one is trivial since $\pi^* \in \Pi$. Now, please show the other inequality $V^* \leq V^{\pi^*} < \infty$.

(c) Similarly, show that $Q^{\pi^*} = Q^*$ under the assumption that the transition is deterministic.

(a) 依據 Bellman: π^* 是通過對於每個狀態 $s \in S$ 選擇使得 $Q^*(s,a)$ 最大化的行為 a 定義的:

$$\pi^*(s) \in \arg\max_{a \in A} r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V^*(s')]$$

對於每個狀態 s，π^* 都選擇能使上式最大化的行動 a

∵ 最大化操作 argmax 是針對一個確定的行動值 $a \in A$，而每個 a 的期望值 $r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V^*(s')]$ 是確定的數值

∴ 每個 s，最多只有一個 a 能達到該最大值

⇒ π^* 是不隨機的，每個狀態對應一個唯一的行動選擇

(b) 證明 $V^*(s) \leq V^{\pi^*}(s)$

令對任何狀態 s 和動作 a，下一個狀態是個函數 $s' = f(s,a)$

Bellman optimality equation for V^*:

$V^*(s) = \max_{a \in A} [r(s,a) + \gamma V^*(f(s,a))]$ —①

∵ $\pi^*(s)$ achieves that maximum: $V^*(s) = r(s, \pi^*(s)) + \gamma V^*(f(s, \pi^*(s)))$

$V^{\pi^*}(s) = r(s, \pi^*(s)) + \gamma V^{\pi^*}(f(s, \pi^*(s)))$ —②

Comparing ① and ②, the difference is just $V^*(f(s, \pi^*(s)))$ & $V^{\pi^*}(f(s, \pi^*(s)))$

by forward iteration that for the next state $s_1 = f(s, \pi^*(s))$, we also have $V^*(s_1) \leq V^{\pi^*}(s_1)$

Repeating this forward along the entire state sequence generated by π^* yields $V^*(s) \leq V^{\pi^*}(s)$ for all $s \in S$

證明 $V^{\pi^*}(s) < \infty$

Assume rewards are bounded: $0 \leq r(s,a) \leq R_{max} < \infty$

$V^{\pi^*}(s) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \leq \sum_{t=0}^{\infty} \gamma^t R_{max} = \frac{R_{max}}{1-\gamma} < \infty$

$V^*(s) \leq V^{\pi^*}(s) < \infty$ (Bellman argument + bounded rewards)

∴ $V^{\pi^*}(s) = V^*(s)$, π^* is optimal in the deterministic transition MDP

(c) $Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V^*(s')]$

In the deterministic case, if next state is $s' = f(s,a)$ ⇒ $Q^*(s,a) = r(s,a) + \gamma V^*(f(s,a))$

For π^*, $Q^{\pi^*}(s,a) = r(s,a) + \gamma V^{\pi^*}(f(s,a))$

∵ $V^{\pi^*}(s') = V^*(s')$ for all states s', it follows: $Q^{\pi^*}(s,a) = Q^*(s,a)$ for all s,a

∴ $Q^{\pi^*} = Q^*$, completing the argument for action-value optimality as well.