R12945060

Problem 1 (Preliminary) (1 pt)

補：$\|x-a\|_2 = \sqrt{(x_1-a_1)^2+(x_2-a_2)^2+\cdots+(x_n-a_n)^2}$
$= \sqrt{(x-a)^T(x-a)}$

(a) (0.2 pts)

(i) (0.1 pts) Given $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Show that

$$\frac{\partial \|\mathbf{x}-\mathbf{a}\|_2}{\partial \mathbf{x}} = \frac{\mathbf{x}-\mathbf{a}}{\|\mathbf{x}-\mathbf{a}\|_2}.$$

$y = \|x-a\|_2 = \sqrt{(x-a)^T(x-a)}$

令 $s = x-a \Rightarrow y = \sqrt{s^Ts}$

$\frac{\partial y}{\partial s} = \frac{1}{2}(s^Ts)^{-1/2} \cdot 2s = \frac{s}{\|s\|_2}$

鏈式：$\frac{\partial s}{\partial x} = \frac{\partial (x-a)}{\partial x} = I$ （I為n×n單位矩陣）

$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial s}\frac{\partial s}{\partial x} = \frac{s}{\|s\|_2} \cdot I = \frac{x-a}{\|x-a\|_2}$

得 $\frac{\partial \|x-a\|_2}{\partial x} = \frac{x-a}{\|x-a\|_2}$

(ii) (0.1 pts) Given $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^n$. Show that

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T.$$

$y = a^TXb \Rightarrow y = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i x_{ij} b_j$

X矩陣中每個元素

$\frac{\partial}{\partial x_{ij}}\left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_i x_{ij} b_j\right) = a_i b_j$

$\frac{\partial a^TXb}{\partial X} = ab^T$

(b) (0.2 pts) Let $\mathbf{X} \in \mathbb{R}^{n\times n}$. Show that

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \det(\mathbf{X})\left(\mathbf{X}^{-1}\right)^{\mathsf{T}}.$$

Hint: Recall the cofactor matrix

$$\mathbf{C} = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix}$$

where $C_{ij} = (-1)^{i+j} M_{ij}$ and $M_{ij} = \det\left((x_{mn})_{m\neq i, n\neq j}\right)$. The adjoint matrix is the transpose of the cofactor matrix

$$\mathrm{adj}(\mathbf{X}) = \mathbf{C}^{\mathsf{T}}.$$

We have an identity

$$\mathbf{X}\mathrm{adj}(\mathbf{X}) = \det(\mathbf{X}).$$

You may check Wikipedia for more details.

$\frac{\partial \det(X)}{\partial X_{ij}} = \text{cofactor}_{ji}(X)$ $\quad X^{-1} = \frac{1}{\det(X)}\,\text{adj}(X)$

$\frac{\partial \det(X)}{\partial X} = \text{adj}(X) = C^T$ $\quad \frac{\partial \det(X)}{\partial X} = \det(X)(X^{-1})^T$

$X \cdot \text{adj}(X) = \det(X) I$

(c) (0.6 pts) Prove that

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial a_{ij}} = \mathbf{e}_j^{\mathsf{T}} \mathbf{A}^{-1} \mathbf{e}_i, \tag{1}$$

where $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \in \mathbb{R}^{m\times m}$ is a (non-singular) matrix, and $\mathbf{e}_j$ is the unit vector along the j-th axis (e.g. $\mathbf{e}_3 = [0, 0, 1, 0, ..., 0]^T$). It is common to write (1) as

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial \mathbf{A}} = \left(\mathbf{A}^{-1}\right)^{\mathsf{T}}$$

Hint: Same as (b).

$\frac{\partial \det(A)}{\partial A} = \det(A)(A^{-1})^T$

$\frac{\partial \log(\det(A))}{\partial A} = \frac{1}{\det(A)} \frac{\partial \det(A)}{\partial A}$

$\frac{\partial \log(\det(A))}{\partial A} = (A^{-1})^T$

$\frac{\partial \log(\det(A))}{\partial a_{ij}} = (A^{-1})_{ji}$

$(A^{-1})_{ji} = e_j^T A^{-1} e_i \Rightarrow \frac{\partial \log(\det(A))}{\partial a_{ij}} = e_j^T A^{-1} e_i$

Problem 2 (Classification with Gaussian Mixture Model) (2.4 pts)

In this question, we tackle the binary classification problem through the generative approach, where we assume the data point X (viewed as a $\mathbb{R}^d$-valued r.v.) and its label Y (viewed as a $\{\mathcal{C}_1, \mathcal{C}_2\}$-valued r.v.) are generated according to the generative model (paramerized by θ) as follows:

$$\mathbb{P}_\theta[X = \mathbf{x}, Y = \mathcal{C}_k] = \pi_k f_{\boldsymbol{\mu}_k, \Sigma_k}(\mathbf{x}) \quad (k \in \{1, 2\}) \tag{2}$$

where $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ for which

$$f_{\boldsymbol{\mu}_k, \Sigma_k}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

Now suppose we observe data points $\mathbf{x}_1, ..., \mathbf{x}_N$ and their corresponding labels $y_1, ..., y_N$.

(a) (1.2 pt)

(i) (0.3 pt) Please write down the likelihood function $L(\theta)$ that describes how likely the generative model would generate the observed data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$.

(ii) (0.3 pt) Find the maximum likelihood estimate $\theta^* = (\pi_1^*, \pi_2^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma_1^*, \Sigma_2^*)$ that maximizes the likelihood function $L(\theta)$.

(iii) (0.3 pt) Write down $\mathbb{P}_\theta[Y = \mathcal{C}_1 | X = \mathbf{x}]$ and $\mathbb{P}_\theta[X = \mathbf{x} | Y = \mathcal{C}_1]$ in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. What are the physical meaning of the aforementioned quantities?

(iv) (0.3 pt) Express $\mathbb{P}_\theta[Y = \mathcal{C}_1 | X = \mathbf{x}]$ in the form of $\sigma(z)$, where $\sigma(\cdot)$ denotes the sigmoid function, and express z in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ and x.

(b) (1.2 pt) Suppose we pose an additional constraint that the covariance matrices of the two Gaussian distributions are identical, namely $\Sigma_1 = \Sigma_2 = \Sigma$, in which the generative model is parameterized by $\vartheta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. Redo questions (a) under such setting.

(a)(i) $L(\theta) = \prod_{i=1}^{N} P_\theta[X = x_i, Y = y_i]$

→ 使用 $P_\theta[X=x, Y=C_k]$ 展開 $L(\theta) = \prod_{i=1}^{N} (\pi_1 f_{\mu_1,\Sigma_1}(x_i))^{1(y_i=1)} (\pi_2 f_{\mu_2,\Sigma_2}(x_i))^{1(y_i=2)}$

(ii) $\log L(\theta) = \sum_{i=1}^{N} (1(y_i=1) \log(\pi_1 f_{\mu_1,\Sigma_1}(x_i)) + 1(y_i=2) \log(\pi_2 f_{\mu_2,\Sigma_2}(x_i)))$

$\Rightarrow \log L(\theta) = \sum_{i=1}^{N} 1(y_i=1)(\log\pi_1 + \log f_{\mu_1,\Sigma_1}(x_i)) + \sum_{i=1}^{N} 1(y_i=2)(\log\pi_2 + \log f_{\mu_2,\Sigma_2}(x_i))$

$\pi_1^* = \frac{1}{N}\sum_{i=1}^{N} 1(y_i=1)$, $\pi_2^* = \frac{1}{N}\sum_{i=1}^{N} 1(y_i=2)$

$\mu_1^* = \frac{\sum_{i=1}^{N} 1(y_i=1) x_i}{\sum_{i=1}^{N} 1(y_i=1)}$, $\mu_2^* = \frac{\sum_{i=1}^{N} 1(y_i=2) x_i}{\sum_{i=1}^{N} 1(y_i=2)}$

$\Sigma_1^* = \frac{\sum_{i=1}^{N} 1(y_i=1)(x_i-\mu_1^*)(x_i-\mu_1^*)^\top}{\sum_{i=1}^{N} 1(y_i=1)}$ $\Sigma_2^* = \frac{\sum_{i=1}^{N} 1(y_i=2)(x_i-\mu_2^*)(x_i-\mu_2^*)^\top}{\sum_{i=1}^{N} 1(y_i=2)}$

(iii) ① $P_\theta[Y=C_1|X=x] = \dfrac{P_\theta[X=x, Y=C_1]}{P_\theta[X=x]}$

$P_\theta[X=x, Y=C_1] = \pi_1 f_{\mu_1,\Sigma_1}(x)$, $P_\theta[X=x] = \pi_1 f_{\mu_1,\Sigma_1}(x) + \pi_2 f_{\mu_2,\Sigma_2}(x)$

$P_\theta[Y=C_1|X=x] = \dfrac{\pi_1 f_{\mu_1,\Sigma_1}(x)}{\pi_1 f_{\mu_1,\Sigma_1}(x) + \pi_2 f_{\mu_2,\Sigma_2}(x)}$

② $P_\theta[X=x|Y=C_1] = f_{\mu_1,\Sigma_1}(x) = \dfrac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)$

在类别 C_1 中，数据点 x 的生成过程

$P_\theta[Y=C_1|X=x]$ ⇒ 在给定数据点 x 的情况下，它属于类别 C_1 的概率

⇒ 后验，在观察到数据 x 后，对其所属类别的信心

$P_\theta[X=x|Y=C_1]$ ⇒ 在数据属于类别 C_1 的条件下，生成数据点 x 的概率

⇒ 生成式概率，在给定类别的情况下数据的分布

(iv) $P_\theta[Y=C_1|X=x] = \sigma(z) = \dfrac{1}{1+e^{-z}} = \dfrac{1}{1+\frac{\pi_2 f_{\mu_2,\Sigma_2}(x)}{\pi_1 f_{\mu_1,\Sigma_1}(x)}}$

取对数：

$P_\theta[Y=C_1|X=x] = \dfrac{1}{1+e^{-z}}$ $\left(z = \log\dfrac{\pi_1 f_{\mu_1,\Sigma_1}(x)}{\pi_2 f_{\mu_2,\Sigma_2}(x)}\right)$

$$z = \log\left(\frac{\pi_1}{\pi_2}\right) + \log\left(\frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}}\right) - \frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$$

$P_\theta[Y=C_1|X=x] = \sigma(z)$

z 为 $\log\left(\frac{\pi_1}{\pi_2}\right) + \log\left(\frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}}\right) - \frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$

(b) (i) $P_\theta[X=x, Y=C_k] = \pi_k f_{\mu_k,\Sigma}(x), \quad k \in \{1,2\}$

$f_{\mu_k,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right)$ ⇒ 統一 Σ

$$\Rightarrow L(\theta) = \prod_{i=1}^{N} \left(\pi_1 f_{\mu_1,\Sigma}(x_i)\right)^{1(y_i=1)} \left(\pi_2 f_{\mu_2,\Sigma}(x_i)\right)^{1(y_i=2)}$$

(ii) $$\log L(\theta) = \sum_{i=1}^{N} 1(y_i=1)\left(\log \pi_1 + \log \mu_{1,\Sigma}(x_i)\right) + \sum_{i=1}^{N} 1(y_i=2)\left(\log \pi_2 + \log f_{\mu_2,\Sigma}(x_i)\right)$$

$$\pi_1^* = \frac{1}{N}\sum_{i=1}^{N} 1(y_i=1), \quad \pi_2^* = \frac{1}{N}\sum_{i=1}^{N} 1(y_i=2), \quad \mu_1^* = \frac{\sum_{i=1}^{N} 1(y_i=1)x_i}{\sum_{i=1}^{N} 1(y_i=1)}, \quad \mu_2^* = \frac{\sum_{i=1}^{N} 1(y_i=2)x_i}{\sum_{i=1}^{N} 1(y_i=2)}$$

$$\Sigma^* = \frac{\sum_{i=1}^{N} 1(y_i=1)(x_i-\mu_1^*)(x_i-\mu_1^*)^T + \sum_{i=1}^{N} 1(y_i=2)(x_i-\mu_2^*)(x_i-\mu_2^*)^T}{N}$$

(iii) $$P_\theta[Y=C_1 | X=x] = \frac{\pi_1 f_{\mu_1,\Sigma}(x)}{\pi_1 f_{\mu_1,\Sigma}(x) + \pi_2 f_{\mu_2,\Sigma}(x)}$$

(iv) $P_\theta[Y=C_1 | X=x] = \sigma(z)$

$$z = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}\left[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]$$

Problem 3 (Closed-Form Linear Regression Solution) (1 pts + Bonus 1.5 pts)

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times d}, \boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \in \mathbb{R}^n$. Denote $\mathbf{X}_i \in \mathbb{R}^{1\times d}$ as the i-th row of $\mathbf{X}$, with the following interpretations:

- If the linear model has the bias term, then write $\boldsymbol{\theta} = [w_1, \cdots, w_m, b]^\mathsf{T}$ and $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}, 1]$, namely $d = m + 1$.
- If the linear model has no bias term, then write $\boldsymbol{\theta} = [w_1, \cdots, w_d]^T$ and $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}]$, namely $d = m$.

(a) Without the bias term, consider the L^2-regularized loss function:

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_i\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2, \ \lambda > 0.$$

Show that the optimal solution that minimizes the loss function is $\boldsymbol{\theta}^* = \left(\boldsymbol{X}^T\boldsymbol{K}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{K}\boldsymbol{y}$, where

$$\boldsymbol{K} = \begin{bmatrix} \kappa_1 & & 0 \\ & \ddots & \\ 0 & & \kappa_n \end{bmatrix}$$

is a diagonal matrix and $\boldsymbol{I}$ is the $d \times d$ identical matrix.

$L(\theta) = (y - X\theta)^T K (y - X\theta) + \lambda\theta^T\theta$

展開 $\Rightarrow L(\theta) = (y - X\theta)^T K (y - X\theta) + \lambda\theta^T\theta = y^T K y - 2y^T K X\theta + \theta^T X^T K X\theta + \lambda\theta^T\theta$

對 θ 求導：$\frac{\partial L(\theta)}{\partial \theta} = -2X^T K y + 2X^T K X\theta + 2\lambda\theta$

最小化 loss function，梯度設為 0：

$-2X^T K y + 2X^T K X\theta + 2\lambda\theta = 0 \Rightarrow X^T K X\theta + \lambda\theta = X^T K y$

$\Rightarrow (X^T K X + \lambda I)\theta = X^T K y \Rightarrow \theta = (X^T K X + \lambda I)^{-1} X^T K y$

(b) (Bonus, 1.5 pts) With the bias term, the L^2-regularized loss function becomes

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_i\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2,\ \lambda > 0.$$

Show that the optimal solution that minimizes the loss function is $\boldsymbol{\theta}^* = [\boldsymbol{w}^{\star T}, b^\star]^T$, where

$$\boldsymbol{w}^\star = \left(\tilde{\boldsymbol{X}}^T\boldsymbol{K}\tilde{\boldsymbol{X}} + \lambda\boldsymbol{I} - \frac{1}{\mathrm{Tr}\left(\boldsymbol{K}\right)}\tilde{\boldsymbol{X}}^T\boldsymbol{K}\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{K}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{K}\left(\boldsymbol{y} - \frac{1}{\mathrm{Tr}\left(\boldsymbol{K}\right)}\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{K}\boldsymbol{y}\right),$$

$$b^\star = \frac{1}{\mathrm{Tr}\left(\boldsymbol{K}\right)}\left(\boldsymbol{e}^T\boldsymbol{K}\boldsymbol{y} - \boldsymbol{e}^T\boldsymbol{K}\tilde{\boldsymbol{X}}\boldsymbol{w}^\star\right)$$

for which $\boldsymbol{e} = [1\ \ldots\ 1]^T$ denotes the all one vector, $\boldsymbol{X} = [\tilde{\boldsymbol{X}}\boldsymbol{e}]$, $\mathrm{Tr}\left(\boldsymbol{K}\right)$ is the trace of the matrix $\boldsymbol{K}$, and that $\mathbf{K}$ and $\mathbf{I}$ are defined as in (a).

$L(\theta) = \sum_i k_i (y_i - \tilde{X}_i w - b)^2 + \lambda w^T w$

$\Rightarrow L(w,b) = (y - \tilde{X}w - be)^T K (y - \tilde{X}w - be) + \lambda w^T w$

對w和b求偏度：

對w求導數：$\frac{\partial L(w,b)}{\partial w} = -2\tilde{X}^T K(y - \tilde{X}w - be) + 2\lambda w$

令等於0：$\tilde{X}^T K \tilde{X} w + \lambda w = \tilde{X}^T K (y - be)$

對b求導數：$\frac{\partial L(w,b)}{\partial b} = -2e^T K (y - \tilde{X}w - be)$

令等於0：$e^T K e b = e^T K (y - \tilde{X}w)$

$w^* = (\tilde{X}^T K \tilde{X} + \lambda I - \frac{1}{Tr(k)} \tilde{X}^T K e e^T K \tilde{X})^{-1} \tilde{X}^T K (y - \frac{1}{Tr(k)} e^T K y)$

$b^* = \frac{1}{Tr(k)} (e^T K y - e^T K \tilde{X} w^*)$

$\theta^* = [w^{*T}, b^*]^T$

Problem 4 (Noise and Regularization) (1 pts)

Consider the linear model $f_{\mathbf{w},b} : \mathbb{R}^k \to \mathbb{R}$, where $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$, defined as

$$f_{\mathbf{w},b}(x) = \mathbf{w}^T\mathbf{x} + b$$

Given dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, if the inputs $\mathbf{x}_i \in \mathbb{R}^k$ are contaminated with input noise $\boldsymbol{\eta}_i \in \mathbb{R}^k$, we may consider the expected sum-of-squares loss in the presence of input noise as

$$\tilde{L}_{ss}(\mathbf{w}, b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i + \boldsymbol{\eta}_i) - y_i)^2\right]$$

where the expectation is taken over the randomness of input noises $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_N$. Additionally, the inputs $(\mathbf{x}_i)$ and the input noise $(\boldsymbol{\eta}_i)$ are independent.

Now assume the input noises $\eta_i = [\eta_{i,1}, \eta_{i,2}, ..., \eta_{i,k}]^T$ are random vectors with zero mean $\mathbb{E}[\eta_{i,j}] = 0$, and the covariance between components is given by

$$\mathbb{E}[\eta_{i,j}\eta_{i',j'}] = \delta_{i,i'}\delta_{j,j'}\sigma^2$$

where $\delta_{i,i'} = \begin{cases} 1 & \text{, if } i = i' \\ 0 & \text{, otherwise.} \end{cases}$ denotes the Kronecker delta.

Please show that

$$\tilde{L}_{ss}(\mathbf{w}, b) = \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \frac{\sigma^2}{2}\|\mathbf{w}\|^2$$

That is, minimizing the expected sum-of-squares loss in the presence of input noise is equivalent to minimizing noise-free sum-of-squares loss with the addition of a L^2-regularization term on the weights. (Hint: $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \mathbf{tr}(\mathbf{x}\mathbf{x}^T)$ and the square of a vector is dot product with itself)

loss function: $\tilde{L}_{ss}(w,b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i+\eta_i)-y_i)^2\right]$

展開 $f_{w,b}(x_i+\eta_i)$:

$f_{w,b}(x_i+\eta_i) = w^T(x_i+\eta_i)+b = w^Tx_i + w^T\eta_i + b$

$f_{w,b}(x_i+\eta_i) - y_i = (w^Tx_i + b - y_i) + w^T\eta_i$

$\tilde{L}_{ss}(w,b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}((w^Tx_i+b-y_i)+w^T\eta_i)^2\right]$

展開平方項:

$\tilde{L}_{ss}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[(w^Tx_i+b-y_i)^2 + 2(w^Tx_i+b-y_i)(w^T\eta_i) + (w^T\eta_i)^2\right]$

① 無 noise 的 loss function:

$\mathbb{E}\left[(w^Tx_i+b-y_i)^2\right] = (w^Tx_i+b-y_i)^2$

② 由於 η_i 均值為 0 $\therefore \mathbb{E}\left[2(w^Tx_i+b-y_i)(w^T\eta_i)\right] = 0$

③ $\mathbb{E}[(w^T\eta_i)^2] = \mathbb{E}[w^T\eta_i\eta_i^Tw] = w^T\mathbb{E}[\eta_i\eta_i^T]w = w^T(\sigma^2 I)w = \sigma^2\|w\|^2$

結合: $\tilde{L}_{ss}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}((w^Tx_i+b-y_i)^2) + \frac{1}{2N}\sum_{i=1}^{N}\sigma^2\|w\|^2$

$\sigma^2\|w\|^2$ 和 i 無關: $\tilde{L}_{ss}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}((w^Tx_i+b-y_i)^2) + \frac{\sigma^2}{2}\|w\|^2$

最小化 input noise 的期望平方損失 = 最小化無 noise 的平方損失 + 加上 L^2-regularization $\frac{\sigma^2}{2}\|w\|^2$

得 $\tilde{L}_{ss}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i)^2 + \frac{\sigma^2}{2}\|w\|^2$

$\frac{\sigma^2}{2}\|w\|^2$ 可減少 noise 的影響, 對 model 更穩定

Problem 5 (Gradient descent for Logistic Regression with Vectorized Feature) (0.6 pts)

This problem is related to the appendix of W2_Logistic_Regression.pdf. Consider the following optimization problem

$$\min_{\mathbf{w}} \ell(\mathbf{w}), \tag{3}$$

where

$$\ell(\mathbf{w}) = \frac{1}{d}\sum_{n=1}^{d} \ell^{(n)}(\mathbf{w}), \quad \ell^{(n)}(\mathbf{w}) = \ln\left(1+\exp\left(-y_n\left(\mathbf{w}^\mathsf{T}\mathbf{x}_n\right)\right)\right).$$

Assume that there are d training data, $\mathbf{x}_n$ is the n-th training data, and the label $y_n = \pm 1$.

(a) (0.2 pts) Prove that $\frac{1}{\ln 2}\ell^{(n)}(\mathbf{w})$ is an upper bound of $\mathbb{1}\{\text{sign}\left(\mathbf{w}^\mathsf{T}\mathbf{x}_n\right) \neq y_n\}$ for any $\mathbf{w}$, where $\mathbb{1}\{\cdot\}$ is the indicator function. Do not use graph calculator for the arguments.

(a) $\ell^{(n)}(w) = \ln(1+\exp(-y_n(w^T x_n)))$

⇒ 當 $y_n(w^T x_n)$ 越小 $\ell^{(n)}(w)$ 越大

由 indicator function 知 $\mathbb{1}\{sign(w^T x_n) \neq y_n\}$ 值是 1 表示錯誤分類，0 為正確分類

⇒ $y_n(w^T x_n) > 0$，分類正確，$y_n(w^T x_n) < 0$，分類錯誤

pf. $\frac{1}{\ln 2}\ell^{(n)}(w) \geq \mathbb{1}\{sign(w^T x_n) \neq y_n\}$

$y_n(w^T x_n) \geq 0 \Rightarrow \ell^{(n)}(w) = \ln(1+\exp(-y_n(w^T x_n))) \leq \ln(1+1) = \ln 2$

$\frac{1}{\ln 2}\ell^{(n)}(w) \leq 1$，分類正確時 loss ≤ 1

$y_n(w^T x_n) < 0 \Rightarrow \ell^{(n)}(w) = \ln(1+\exp(-y_n(w^T x_n))) \geq \ln 2$

$y_n(w^T x_n)$ 越小，loss 越接近 $\ln(1+\exp(0)) = \ln 2 \Rightarrow \frac{1}{\ln 2}\ell^{(n)}(w) \geq 1$

分類錯誤，loss ≥ 1

$\frac{1}{\ln 2}\ell^{(n)}(w) \geq \mathbb{1}\{sign(w^T x_n) \neq y_n\}$

$\frac{1}{\ln 2}$ 是錯誤分類 indicator function 的上界

(b) (0.2 pts) For a given $(\mathbf{x}_n, y_n)$, derive its gradient $\nabla \ell^{(n)}(\mathbf{w})$.

(b) $\ell^{(n)}(w) = \ln(1+\exp(-y_n(w^T x_n)))$

$g(w) = -y_n(w^T x_n)$

$\Rightarrow \ell^{(n)}(w) = \ln(1+\exp(g(w)))$

$\nabla \ell^{(n)}(w) = \frac{d}{dw}\ln(1+\exp(g(w)))$

$\frac{d}{dw}\ln(1+\exp(g(w))) = \frac{1}{1+\exp(g(w))} \cdot \frac{d}{dw}\exp(g(w))$

$(\because g(w) = -y_n(w^T x_n))$

$\frac{d}{dw}\exp(g(w)) = \exp(g(w)) \cdot \frac{d}{dw} g(w)$ 又 $\frac{d}{dw} g(w) = -y_n x_n$

$\frac{d}{dw}\exp(g(w)) = \exp(g(w)) \cdot (-y_n x_n)$

综合 $\nabla \ell^{(n)}(w) = \frac{1}{1+\exp(g(w))} \cdot \exp(g(w)) \cdot (-y_n x_n)$

$\Rightarrow \nabla \ell^{(n)}(w) = \frac{\exp(g(w))}{1+\exp(g(w))} \cdot (-y_n x_n)$

$\frac{\exp(g(w))}{1+\exp(g(w))} = \frac{1}{1+\exp(-g(w))} \Rightarrow \nabla \ell^{(n)}(w) = \frac{1}{1+\exp(y_n(w^T x_n))} \cdot (-y_n x_n)$

(c) (0.2 pts) Prove that the optimization problem 3 is equivalent to minimizing the following objective function

$$\mathcal{L}(\mathbf{w}) = -\sum_{n=1}^{d} \left(\frac{1+y_n}{2} \ln \frac{1+\tanh\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)}{2} + \frac{1-y_n}{2} \ln \frac{1-\tanh\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)}{2} \right).$$

$\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$，$1+\tanh(z) = \frac{2\exp(z)}{\exp(z)+\exp(-z)}$，$1-\tanh(z) = \frac{2\exp(-z)}{\exp(z)+\exp(-z)}$

$\frac{1+y_n}{2} \ln \frac{1+\tanh(\frac{1}{2}w^T x_n)}{2}$：

當 $y_n=1$，分類正確：$\ln(1+\exp(-w^T x_n))$

$\frac{1-y_n}{2} \ln \frac{1-\tanh(\frac{1}{2}w^T x_n)}{2}$

當 $y_n=-1$，分類錯誤：$\ln(1+\exp(w^T x_n))$

得 $\mathcal{L}(w) = \ell(w)$