ML2023 Fall Homework Assignment 1
Handwritten

Lecturor: Pei-Yuan Wu
TAs: Po-Yang Hsieh, Le-Rong Hsu

Sep 2023

Problem 1 (Preliminary) (1 pt)

In this problem, you need to find the derivative of 2-norm or a scalar with respect to a vector or a matrix. For (b) and (c), you may start by considering

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{n1}} & \cdots & \frac{\partial y}{\partial x_{nn}} \end{bmatrix}.$$

(Hint: Find a partial derivative with respect to the (i, j)-th component and sort out the vector or matrix form.)

(a) (0.2 pts)

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

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

(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

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

(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

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

You may check Wikipedia for more details.

(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).

Problem 1 ans

(a) (i) Recall that $\|\mathbf{x}-\mathbf{a}\|_2 = \sqrt{(x_1-a_1)^2+\cdots+(x_n-a_n)^2}$. Observe that

$$\frac{\partial \|\mathbf{x}-\mathbf{a}\|_2}{\partial x_1} = \frac{1}{2}\frac{2(x_1-a_1)}{\sqrt{(x_1-a_1)^2+\cdots+(x_n-a_n)^2}} = \frac{x_1-a_1}{\sqrt{(x_1-a_1)^2+\cdots+(x_n-a_n)^2}}.$$

Actually,

$$\frac{\partial \|\mathbf{x}-\mathbf{a}\|_2}{\partial x_j} = \frac{1}{2}\frac{2(x_j-a_j)}{\sqrt{(x_1-a_1)^2+\cdots+(x_n-a_n)^2}} = \frac{x_j-a_j}{\sqrt{(x_1-a_1)^2+\cdots+(x_n-a_n)^2}}.$$

Thus,

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

(ii) Express $\mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{b}$ as

$$[a_1,\ldots,a_m]\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

Calculate the partial derivative with respect to x_{ij}:

$$\frac{\mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{b}}{\partial x_{ij}} = a_i b_j.$$

Write the Jacobian into a matrix form $\frac{\mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^\mathsf{T}$.

(b) Recall that

$$\det(\mathbf{X}) = \sum_{i=1}^{m}(-1)^{i+j}x_{ij}M_{ij} = \sum_{i=1}^{m} x_{ij}C_{ij}$$

where $M_{ij} = \det\left((x_{mn})_{m\neq i, n\neq j}\right)$ and $C_{ij} = (-1)^{i+j}M_{ij}$.

$$\frac{\partial \det(\mathbf{X})}{\partial x_{ij}} = C_{ij}.$$

Express the derivative as the matrix form

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{C} = \operatorname{adj}(\mathbf{X})^\mathsf{T}$$

where

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

is the cofactor matrix. Thus,

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

(c) Suppose that $\mathbf{A}$ is invertible. Observe that

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

Recall that for any square matrix $\mathbf{A}$ and fixed $1 \leq j \leq m$,

$$\det(\mathbf{A}) = \sum_{i=1}^{m} (-1)^{i+j} a_{ij} M_{ij} = \sum_{i=1}^{m} a_{ij} C_{ij}$$

where $M_{ij} = \det(a_{mn})_{m \neq i, n \neq j}$ and $C_{ij} = (-1)^{i+j} M_{ij}$. Remember the relation

$$\mathbf{A}\operatorname{adj}(\mathbf{A}) = \det(\mathbf{A}).$$

Fix $1 \leq i, j \leq m$ and expand $\det(\mathbf{A}) = \sum_{k=1}^{m} a_{kj} C_{kj}$. Then

$$\begin{aligned}
&= \frac{1}{\det(\mathbf{A})} \frac{\partial \det(\mathbf{A})}{\partial a_{ij}} \\
&= \frac{1}{\det(\mathbf{A})} \frac{1}{\partial a_{ij}} \sum_{k=1} a_{kj} C_{kj} \\
&= \frac{1}{\det(\mathbf{A})} C_{ij} = \frac{1}{\det(\mathbf{A})} \operatorname{adj}(\mathbf{A})_{ji}.
\end{aligned}$$

Since $\operatorname{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{A}^{-1}$, we have

$$\frac{1}{\det(\mathbf{A})} \operatorname{adj}(\mathbf{A})_{ji} = (\mathbf{A}^{-1})_{ji}.$$

Thus,

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

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

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

Now suppose we observe data points $\mathbf{x}_1, ..., \mathbf{x}_N$ and their corresponding labels $y_1, ..., y_N$, and $\pi_1 + \pi_2 = 1$.

(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.

Problem 2 ans

(a) (i) The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{N} (\mathbb{1}(y_i = C_1)\pi_1 f_{\boldsymbol{\mu}_1,\Sigma_1}(\mathbf{x}_i) + \mathbb{1}(y_i = C_2)\pi_2 f_{\boldsymbol{\mu}_2,\Sigma_2}(\mathbf{x}_i))$$

Since the indicator function is not differentiable, you may write it in a another format. W.L.O.G we may assume that there are N_1 numbers of $y_i \in C_1$, N_2 numbers of $y_i \in C_2$ and $N_1 + N_2 = N$The likelihood function is given by

$$\begin{aligned} L(\theta) =& \frac{1}{(2\pi)^{dN/2}} \pi_1^{N_1} \pi_2^{N_2} \frac{1}{|\Sigma_1|^{N_1/2}} \frac{1}{|\Sigma_2|^{N_2/2}} \times \\ & \prod_{i,y_i=C_1}^{N_1} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_1)^\mathsf{T} \Sigma_1^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)\right) \prod_{j,y_j=C_2}^{N_2} \exp\left(-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_2)^\mathsf{T} \Sigma_2^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_2)\right) \end{aligned}$$

(ii) To make the calculation easier later, we change the above answer into log likelihood function

$$\begin{aligned} L - log(\theta) =& log(\frac{1}{(2\pi)^{dN/2}}) + N_1 log(\pi_1) + N_2 log(\pi_2) + \frac{-N_1}{2} log(|\Sigma_1|) + \frac{-N_2}{2} log(|\Sigma_2|) + \\ & \sum_{i,y_i=C_1}^{N_1} \left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_1)^\mathsf{T} \Sigma_1^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)\right) + \sum_{j,y_j=C_2}^{N_2} \left(-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_2)^\mathsf{T} \Sigma_2^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_2)\right) \end{aligned}$$

Now we calculate the optimal π_1^*, π_2^*. Note that $\pi_1 + \pi_2 = 1$

$$\frac{\partial L - log(\theta)}{\partial \pi_1} = \frac{N_1}{\pi_1} + \frac{N_2}{1 - \pi_1} = 0$$

$$(1 - \pi_1)N_1 + \pi_1 N_2 = 0 \Rightarrow \pi_1 * = \frac{N_1}{N}$$

same for π_2 we have

$$\pi_2^* = \frac{N_2}{N}$$

for μ_1^*, μ_2^*

$$\frac{\partial L - log(\theta)}{\partial \mu_1} = \sum_{i,y_i=C_1}^{N_1} \left(\Sigma_1^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\right) = \Sigma_1^{-1} \sum_{i,y_i=C_1}^{N_1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) = 0$$

$$\sum_{i,y_i=C_1}^{N_1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) = 0 \Rightarrow \mu_1 * = \frac{\sum_{i,y_i=C_1}^{N_1} x_i}{N_1}$$

same for μ_2 we have

$$\mu_2^* = \frac{\sum_{i,y_i=C_2}^{N_2} x_i}{N_2}$$

for Σ_1^*, Σ_2^*, note that

$$\frac{\partial L-log(\theta)}{\partial \Sigma_1} = \frac{\partial L-log(\theta)}{\partial \Sigma_1^{-1}}\frac{\partial \Sigma_1^{-1}}{\partial \Sigma_1} = 0$$

Since the later one is not 0. We need the former one to be 0.

$$\begin{aligned}
\frac{\partial L-log(\theta)}{\partial \Sigma_1^{-1}} &= \frac{1}{2}\frac{\partial - N_1 log(|\Sigma_1|)}{\partial \Sigma_1^{-1}} + \frac{\partial \sum_{i,y_i=C_1}^{N_1}\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}\Sigma_1^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_1)\right)}{\partial \Sigma_1^{-1}} \\
&= \frac{1}{2}\frac{\partial N_1 log(|\Sigma_1^{-}1|)}{\partial \Sigma_1^{-1}} + -\frac{1}{2}\frac{\partial \sum_{i,y_i=C_1}^{N_1} tr\left((\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}\Sigma_1^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_1)\right)}{\partial \Sigma_1^{-1}} \\
&= \frac{1}{2}\left(N_1\Sigma_1^T - \frac{\partial \sum_{i,y_i=C_1}^{N_1} tr\left(\Sigma_1^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}\right)}{\partial \Sigma_1^{-1}}\right) \\
&= \frac{1}{2}\left(N_1\Sigma_1^T - \sum_{i,y_i=C_1}^{N_1}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}\right) = 0
\end{aligned}$$

$$\left(N_1\Sigma_1^T - \sum_{i,y_i=C_1}^{N_1}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}\right) = 0 \Rightarrow \Sigma_1^* = \frac{\sum_{i,y_i=C_1}^{N_1}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}}{N_1}$$

same for Σ_2 we have

$$\Sigma_2^* = \frac{\sum_{i,y_i=C_2}^{N_2}(\mathbf{x}_i-\boldsymbol{\mu}_2)(\mathbf{x}_i-\boldsymbol{\mu}_2)^\mathsf{T}}{N_2}$$

(iii) $\mathbb{P}_\theta[X=\mathbf{x}|Y=\mathcal{C}_1] = \frac{\mathbb{P}(X=x,Y=C_1)}{\mathbb{P}(Y=C_1)} = f_{\boldsymbol{\mu}_1,\Sigma_1}(\mathbf{x})$ which means the probability of$x$ given the class
$\mathbb{P}_\theta[Y=\mathcal{C}_1|X=\mathbf{x}] = \frac{\mathbb{P}(X=x,Y=C_1)}{\mathbb{P}(X=x)} = \frac{\pi_1 f_{\boldsymbol{\mu}_1,\Sigma_1}(\mathbf{x})}{\pi_1 f_{\boldsymbol{\mu}_1,\Sigma_1}(\mathbf{x})+\pi_2 f_{\boldsymbol{\mu}_2,\Sigma_2}(\mathbf{x})}$ which means when we sample a new x how likely is it belongs to C_1

(iv) Same as the induction in class we have

$$z = ln\frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} - \frac{1}{2}x^t(\Sigma_1)^{-1}x + \mu_1^T(\Sigma_1)^{-1}x - \frac{1}{2}\mu_1^T(\Sigma_1)^{-1}\mu^1 + \frac{1}{2}x^t(\Sigma_2)^{-1}x - \mu_2^T(\Sigma_2)^{-1}x + \frac{1}{2}\mu_2^T(\Sigma_2)^{-1}\mu_2 + ln\frac{N_1}{N_2}$$

(b) we only show those are modified.

(i)

$$\begin{aligned}
L(\theta) = &\frac{1}{(2\pi)^{dN/2}}\pi_1^{N_1}\pi_2^{N_2}\frac{1}{|\Sigma|^{N/2}} \\
&\prod_{i,y_i=C_1}^{N_1}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T}\Sigma^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\right)\prod_{j,y_j=C_2}^{N_2}\exp\left(-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu}_2)^\mathsf{T}\Sigma^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_2)\right)
\end{aligned}$$

(ii)

$$\Sigma^* = \frac{\sum_{i,y_i=C_1}^{N_1}(\mathbf{x}_i-\boldsymbol{\mu}_1)(\mathbf{x}_i-\boldsymbol{\mu}_1)^\mathsf{T} + \sum_{i,y_i=C_2}^{N_2}(\mathbf{x}_i-\boldsymbol{\mu}_2)(\mathbf{x}_i-\boldsymbol{\mu}_2)^\mathsf{T}}{N}$$

(iii) the same

(iv)

$$z = (\mu_1-\mu_2)^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T(\Sigma_1)^{-1}\mu_1 + \frac{1}{2}\mu_2^T(\Sigma_1)^{-1}\mu_2 + ln\frac{N_1}{N_2}$$

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:

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_{\mathrm{i}}\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2, \quad \lambda > 0,\ \kappa_i > 0 \text{ for all } i.$$

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.

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

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_{\mathrm{i}}\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2, \quad \lambda > 0,\ \kappa_i > 0 \text{ for all } i.$$

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

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

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

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

Problem 3 ans

(a) First, represent the loss function as

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T\boldsymbol{K}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$

Next, take gradient of $\boldsymbol{\theta}$ and set it to 0, you will get the optimal solution $\boldsymbol{\theta}^* = \left(\boldsymbol{X}^T\boldsymbol{K}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{K}\boldsymbol{y}$

(b) First, represent the loss function as

$$(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w} - b\boldsymbol{e})^T\boldsymbol{K}(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w} - b\boldsymbol{e}) + \lambda\boldsymbol{w}^T\boldsymbol{w}$$

Next, take gradient of both $\boldsymbol{w}$ and b and set them to 0 respectively, you will get two equations. By solving the system of equations carefully, you will get the optimal solution

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

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

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

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

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

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

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

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: $\|\mathbf{x}\|^{\mathbf{2}} = \mathbf{x^T x} = \mathbf{tr}(\mathbf{x x^T})$ and the square of a vector is dot product with itself)

Problem 4 ans

By definition,

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

Note that $\mathbb{E}(\eta_i) = 0$ Now, calculate $\mathbb{E}\left[(\mathbf{w}^T\eta_i)^2\right]$

$$\begin{aligned}
\sum_{i=1}^{N}\mathbb{E}(\mathbf{w}^T\eta_i)^2 &= \sum_{i=1}^{N}\mathbb{E}(\sum_{j=1}^{k} w_j\eta_{i,j}) \\
&= \sum_{i=1}^{N}\mathbb{E}(\sum_{j=1}^{k}\sum_{l=1}^{k} w_j w_l \eta_{i,j}\eta_{i,l}) \\
&= \sum_{j=1}^{k}\sum_{l=1}^{k} w_j w_l \sum_{i=1}^{N}\mathbb{E}(\eta_{i,j}\eta_{i,l}) \\
&= N\sigma^2\sum_{j=1}^{k}\sum_{l=1}^{k} w_j w_l = N\sigma^2\|w\|^2
\end{aligned}$$

Hence,

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

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.

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

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

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

Problem 5 ans

(a) First, Consider the case $y_n = 1$. We want to show $\frac{1}{\ln 2}\ln\left(1+\exp\left(-\mathbf{w}^{\mathsf{T}}\mathbf{x}_n\right)\right)$ is an upper bound of $\mathbb{1}\{\text{sign}\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n\right) \neq 1\}$. For $\mathbf{w}^{\mathsf{T}}\mathbf{x}_n > 0$, we can easily obtain that

$$\mathbb{1}\{\text{sign}\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n\right) \neq 1\} = 0 \leq \frac{1}{\ln 2}\ln\left(1+\exp\left(-\mathbf{w}^{\mathsf{T}}\mathbf{x}_n\right)\right) > 0.$$

For $\mathbf{w}^{\mathsf{T}}\mathbf{x}_n \leq 0$,

$$\mathbb{1}\{\text{sign}\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n\right) \neq 1\} = 1$$

and

$$\frac{1}{\ln 2}\ln\left(1+\exp\left(-\mathbf{w}^\mathsf{T}\mathbf{x}_n\right)\right) \geq \frac{1}{\ln 2}\ln\left(1+\exp\left(0\right)\right) = 1.$$

For the case $y_n = -1$, same results can be obtained and then Q.E.D.

(b) Consider the definition of gradient

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(x)}{\partial x^1} \\ \frac{\partial f(x)}{\partial x^2} \\ \vdots \\ \frac{\partial f(x)}{\partial x^n} \end{bmatrix}.$$

Then

$$\frac{\partial \ell^{(n)}}{\partial w^1} = \frac{1}{1+\exp(-y_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n))} \cdot \exp(-y_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n)) \cdot (-y_n x_n^1).$$

Hence,

$$\nabla \ell^{(n)}(\mathbf{w}) = \frac{-y_n \exp(-y_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n))}{1+\exp(-y_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n))} \cdot \mathbf{x}_n.$$

(c) The first term in Σ is considered when $y_n = 1$

$$\ln\frac{1+\tanh\left(\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{x}_n\right)}{2} = \ln\frac{1+\frac{e^{2\cdot\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{x}_n}-1}{e^{2\cdot\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{x}_n}+1}}{2} = \ln\frac{1+\frac{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_n}-1}{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_n}+1}}{2} = \ln\frac{\frac{2e^{\mathbf{w}^\mathsf{T}\mathbf{x}_n}}{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_n}+1}}{2} = \ln\frac{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_n}}{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_n}+1} = \ln\frac{1}{1+e^{-\mathbf{w}^\mathsf{T}\mathbf{x}_n}}$$

With the minus in front of Σ, the term becomes

$$\ln(1+e^{-\mathbf{w}^\mathsf{T}\mathbf{x}_n})$$

which is exactly $\ell^{(n)}(\mathbf{w})$. This can also be obtained for $y_n = -1$ and then Q.E.D.

Problem 6 (Mathematical Background) (0 pt)

Please click the following link `https://www.cs.cmu.edu/~mgormley/courses/10601/homework/hw1.zip` to download the Homework 1 from CMU 2023 Machine Learning Website. You are encouraged to practice Section 3 to Section 6 of this homework to brush up some of the mathematical background that will be useful for this course. **This problem will not be graded**. However, you are encouraged to consult TA by joining TA hour if you find any questions.

Some Tools You Need to Know

1. Orthogonal Matrix
2. Positive Definite, Semipositive Definite
3. Eigenvalue Decomposition, Singular value decomposition
4. Lagrange Multiplier
5. Trace

You can find the definition and the usage by yourself. It is also welcome to discuss with TA in TA hour.