Machine Learning Course 2024 Spring: Homework 1

March 1, 2024

1 Problem 1

1. **Solution:**

   The original optimization problem can be rewritten as

   $$\min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||_2^2,$$

   where

   $$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix} \in \mathbb{R}^{m\times(d+1)}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m, \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} \in \mathbb{R}^{d+1}.$$

   □

2. The optimal solution $\boldsymbol{\beta}^* = (\boldsymbol{w}^*, b^*)$ for ridge regression is unique.

   Proof:

   $$J(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y})^{\mathrm{T}}(X\boldsymbol{\beta} - \boldsymbol{y}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}.$$

   The first-order derivative of $J(\boldsymbol{\beta})$ w.r.t $\boldsymbol{\beta}$ is

   $$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}) + \lambda\boldsymbol{\beta}.$$

   The second-order derivative of $J(\boldsymbol{\beta})$ w.r.t $\boldsymbol{\beta}$ (Hessian matrix) is

   $$\frac{\partial^2 J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I},$$

where $\mathbf{I} \in \mathbb{R}^{(d+1)\times(d+1)}$ is an identity matrix. Since the Hessian matrix is positive definite ($\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I} \succ 0$), the objective function $J(\boldsymbol{\beta})$ is strictly convex w.r.t. $\boldsymbol{\beta}$. Thus, the optimal parameter $\boldsymbol{\beta}^*$ is unique. $\square$

3. **Solution:**

$$\mathbf{X} = \begin{pmatrix} 2 & 1 & 3 & 1 \\ 3 & 5 & 2 & 1 \\ 5 & 3 & 6 & 1 \\ 1 & 7 & 2 & 1 \\ 4 & 2 & 5 & 1 \\ 6 & 1 & 4 & 1 \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} 0 \\ -3 \\ 0 \\ -3 \\ 0 \\ 3 \end{pmatrix}.$$

Let $\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}_{d+1}$:

$$\begin{aligned} \mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}) + \lambda\boldsymbol{\beta} &= \mathbf{0}_{d+1}, \\ (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} &= \mathbf{X}^{\mathrm{T}}\boldsymbol{y}. \end{aligned}$$

Since $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}$ is positive definite and invertible for $\lambda = 0.1$,

$$\begin{aligned} \boldsymbol{\beta}^* &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} \\ &= \begin{pmatrix} 0.539 \\ -0.599 \\ -0.101 \\ -0.114 \end{pmatrix}. \end{aligned}$$

$\square$

4. **Proof:**

y_i follows a Gaussian distribution with the mean $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_i$ and standard deviation σ:

$$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_i)^2}{2\sigma^2}\right] \tag{1}$$

Then we have:

$$
\begin{aligned}
\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^m (y_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_i)^2}{2\sigma^2} - \frac{\sum_{j=1}^d \theta_i^2}{2\tau^2} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^m (y_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_i)^2}{2\sigma^2} + \frac{\sum_{j=1}^d \theta_i^2}{2\tau^2} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_{i=1}^m (y_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_i)^2 + \frac{\sigma^2}{2\tau^2}||\boldsymbol{\theta}||_2^2
\end{aligned} \tag{2}
$$

Notice that $\lambda = \frac{\sigma^2}{\tau^2}$. □

2 Problem 2

1. For the sigmoid function,

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

the first-order derivative of σ w.r.t z is

$$\nabla\sigma = \frac{\partial\sigma}{\partial z} = \sigma(z)(1-\sigma(z)),$$

the second-order derivative of σ w.r.t z (Hessian Matrix) is

$$\nabla^2\sigma = \frac{\partial^2\sigma(z)}{\partial z^2} = \sigma(z)(1-\sigma(z))(1-2\sigma(z)),$$

Since $0 \le \sigma(z) \le 1$, it follows that $0 \le 1-\sigma(z) \le 1$ and $-1 \le 1-2\sigma(z) \le 1$. Therefore, $1-2\sigma(z)$ can take values between -1 and 1, but it is not necessarily non-negative. Therefore the sigmoid function is non-convex.

For the equation below

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^m \left(-y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i + \ln\left(1+e^{\beta^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right),$$

the first-order derivative of σ w.r.t z is

$$\nabla\ell = \frac{\partial\ell}{\partial\boldsymbol{\beta}} = \sum_{i=1}^m \left(-y_i\hat{\boldsymbol{x}}_i + \frac{e^{\boldsymbol{\beta}^T\hat{\boldsymbol{w}}_i}\hat{\boldsymbol{x}}_i}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}}\hat{\boldsymbol{w}}_i}\right),$$

the second-order derivative of ℓ w.r.t $\boldsymbol{\beta}$ (Hessian Matrix) is

$$\nabla^2 \ell = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{m} \frac{e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{w}}_i} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}}}{\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}\right)^2},$$

For $\forall \boldsymbol{v} \in \mathbb{R}^{d+1} \neq \mathbf{0}_{d+1}$, we have

$$\boldsymbol{v}^{\mathrm{T}} \sum_{i=1}^{m} \frac{e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{w}}_i} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}}}{\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}\right)^2} \boldsymbol{v} = \sum_{i=1}^{m} \frac{e^{\beta^{\mathrm{T}} \hat{\omega}_i}}{\left(1 + e^{\beta^{\mathrm{T}} \hat{\boldsymbol{w}}_j}\right)^2} \left(\hat{\boldsymbol{x}}_i^{\mathrm{T}} \boldsymbol{v}\right)^2, \ \geq 0$$

such that $\nabla^2 \ell \succeq 0$, ℓ is convex.

2. Construct K-1 log odds (logit) for K-class classification:

$$\begin{aligned}
\ln \tfrac{p(y=1|\boldsymbol{x})}{p(y=K|\boldsymbol{x})} &= \boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{x} + b_1, \\
\ln \tfrac{p(y=2|\boldsymbol{x})}{p(y=K|\boldsymbol{x})} &= \boldsymbol{w}_2^{\mathrm{T}} \boldsymbol{x} + b_2, \\
&\cdots \\
\ln \tfrac{p(y=K-1|\boldsymbol{x})}{p(y=K|\boldsymbol{x})} &= \boldsymbol{w}_{K-1}^{\mathrm{T}} \boldsymbol{x} + b_{K-1},
\end{aligned}$$

such that

$$\begin{aligned}
p(y=1 \mid \boldsymbol{x}) &= p(y=K \mid \boldsymbol{x}) e^{\boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{x} + \dot{b}_1}, \\
p(y=2 \mid \boldsymbol{x}) &= p(y=K \mid \boldsymbol{x}) e^{\boldsymbol{w}_2^{\mathrm{T}} \boldsymbol{x} + \boldsymbol{b}_2}, \\
&\cdots \\
p(y=K-1 \mid \boldsymbol{x}) &= p(y=K \mid \boldsymbol{x}) e^{\boldsymbol{w}_{K-1}^{\mathrm{T}} \boldsymbol{x} + b_{K-1}}.
\end{aligned}$$

To guarantee the sum of the probability is 1 , we have

$$p(y=K \mid \boldsymbol{x}) = 1 - \sum_{i=1}^{K-1} p(y=K \mid \boldsymbol{x}) e^{\boldsymbol{w}_j^T \boldsymbol{w} + b_j},$$

such that

$$\begin{aligned}
p(y=K \mid \boldsymbol{x}) &= \tfrac{1}{1+\sum_{i=1}^{K-1} e^{\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x} + b_i}}, \\
p(y=j \mid \boldsymbol{x}) &= \tfrac{e^{\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x} + b_j}}{1+\sum_{i=1}^{K-1} e^{\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x} + b_i}} (j \neq K).
\end{aligned}$$

Let $\boldsymbol{\beta}_j = (\boldsymbol{w}_j; b_j)\,(j \neq K), \boldsymbol{\beta}_K = (\mathbf{0}_d; 0), \hat{\boldsymbol{x}} = (\boldsymbol{x}; 1)$, we have

$$p(y=j \mid \boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}_j^{\mathrm{T}} \hat{\boldsymbol{x}}}}{\sum_{i=1}^{K} e^{\boldsymbol{\beta}_i^{\mathrm{T}} \hat{\boldsymbol{x}}}} (1 \leq j \leq K),$$

the log-likelihood function is

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{m} \ln p\left(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}\right) \\
&= \sum_{i=1}^{m} \ln \left(\frac{e^{\boldsymbol{\beta}_{y_i}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}}{\sum_{j=1}^{K} e^{\boldsymbol{\beta}_j^{\mathrm{T}} \hat{\boldsymbol{x}}}} \right) \\
&= \sum_{i=1}^{m} \left(\boldsymbol{\beta}_{y_i}^{\mathrm{T}} \hat{\boldsymbol{x}} - \ln \left(\sum_{j=1}^{K} e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}} \right) \right)
\end{aligned}
$$

3 Problem 3

1. For classifier C_1, ranking the predicted value y_{C_1} first:

y	0	0	1	1	1	1	0	1	0	0
y_{C_1}	0.88	0.75	0.67	0.54	0.43	0.38	0.38	0.29	0.28	0.11
rank	10	9	8	7	6	5	4	3	2	1

so

$$
\mathrm{AUC}_{C_1} = \frac{3+3+3+2.5+2}{5 \times 5} = \frac{13.5}{25} = \frac{27}{50}
$$

For classifier C_2, ranking the predicted value y_{C_2} first:

y	0	1	0	0	1	1	1	0	0	1
y_{C_1}	0.95	0.89	0.89	0.66	0.66	0.49	0.47	0.23	0.19	0.15
rank	10	9	8	7	6	5	4	3	2	1

so

$$
\mathrm{AUC}_{C_2} = \frac{3.5+2.5+2+2+0}{5 \times 5} = \frac{10}{25} = \frac{2}{5}
$$

2. For y_{C_1}

		Prediction	
		Positive	Negative
Ground Truth	Positive	4	1
	Negative	3	2

For y_{C_2}

		Prediction	
		Positive	Negative
Ground Truth	Positive	2	3
	Negative	3	2

3. $F1$-Score

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \\ F1 = \frac{2\times P\times R}{P+R} \end{cases}$$

For classifier C_1 ,

$$TP_{C_1} = 4, FP_{C_1} = 3, FN_{C_1} = 1, TN_{C_1} = 2,$$
$$P_{C_1} = \frac{4}{4+3} = \frac{4}{7}, R_{C_1} = \frac{4}{4+1} = \frac{4}{5},$$
$$F1_{C_1} = \frac{2\times\frac{4}{7}\times\frac{4}{5}}{\frac{4}{7}+\frac{4}{5}} = \frac{2}{3}.$$

For classifier C_2 ,

$$TP_{C_1} = 2, FP_{C_1} = 3, FN_{C_1} = 3, TN_{C_1} = 2,$$
$$P_{C_1} = \frac{2}{2+3} = \frac{2}{5}, R_{C_1} = \frac{2}{2+3} = \frac{2}{5},$$
$$F1_{C_1} = \frac{2\times\frac{2}{5}\times\frac{2}{5}}{\frac{2}{5}+\frac{2}{5}} = \frac{2}{5}.$$