Artificial Intelligence: HW 2

Jeong Min Lee

November 6, 2023

1 Linear Regression

1.a

Let f be the target function. I'll use the superscript with parenthesis to describe the n-th sample vector.

$$f(\boldsymbol{\omega}, \overline{\boldsymbol{x}^{(n)}}) = \sum_n \frac{1}{2} \left(t^{(n)} - \boldsymbol{\omega}^T \overline{\boldsymbol{x}^{(n)}} \right)^2 \tag{1}$$

To minimize f, differentiate it by $\boldsymbol{\omega}$ and find the $\boldsymbol{\omega}_0$ which makes the derivative zero. To make expression simple, I used the Einstein notation.

$$\begin{aligned}
\frac{\partial f}{\partial \omega_j} &= -(t^{(n)} - \boldsymbol{\omega}^T \overline{\boldsymbol{x}^{(n)}}) \cdot \frac{\partial}{\partial \omega_j} \boldsymbol{\omega}^T \bar{\boldsymbol{x}}^{(n)} \\
&= -(t^{(n)} - \boldsymbol{\omega}^T \overline{\boldsymbol{x}^{(n)}}) x_j^{(n)} \\
&= 0
\end{aligned}$$

By enumerating the $\frac{\partial f}{\partial \omega_j}$ horizontally, one can get $\frac{\partial f}{\partial \boldsymbol{\omega}}$.

$$\sum_n t^{(n)} \begin{pmatrix} x_1^{(n)} \\ \vdots \\ x_M^{(n)} \end{pmatrix}^T = \sum_n \begin{pmatrix} \boldsymbol{\omega}^T \overline{\boldsymbol{x}^{(n)}} x_1^{(n)} \\ \vdots \\ \boldsymbol{\omega}^T \overline{\boldsymbol{x}^{(n)}} x_M^{(n)} \end{pmatrix}^T \tag{2}$$

The left hand side is simply $\sum_n t^{(n)} \overline{\boldsymbol{x}^{(n)}}^T$. From the linearity of vector summation rule, the right hand side is $\left(\left(\sum_n \overline{\boldsymbol{x}^{(n)}} \cdot \overline{\boldsymbol{x}^{(n)}}^T \right) \boldsymbol{\omega} \right)^T$. By taking transpose to both sides, one can get the following equation.

$$\left[\sum_n \overline{\boldsymbol{x}^{(n)}} \cdot \overline{\boldsymbol{x}^{(n)}}^T \right] \boldsymbol{\omega} = \sum_n t^{(n)} \overline{\boldsymbol{x}^{(n)}} \tag{3}$$

Therefore, $\boldsymbol{A} = \sum_n \overline{\boldsymbol{x}^{(n)}} \cdot \overline{\boldsymbol{x}^{(n)}}^T$ and $\boldsymbol{b} = \sum_n t^{(n)} \overline{\boldsymbol{x}^{(n)}}$

1.b

$\overline{\boldsymbol{x}^{(1)}} = (1, 0)^T, t^{(1)} = 1$. $\overline{\boldsymbol{x}^{(2)}} = (1, \epsilon)^T, t^{(2)} = 1$. $\boldsymbol{A} = \overline{\boldsymbol{x}^{(1)}} \cdot \overline{\boldsymbol{x}^{(1)}}^T + \overline{\boldsymbol{x}^{(1)}} \cdot \overline{\boldsymbol{x}^{(2)}}^T = \begin{pmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{pmatrix}$ $\boldsymbol{b} = \overline{\boldsymbol{x}^{(1)}} + \overline{\boldsymbol{x}^{(2)}} = (2, \epsilon)^T$. Since $\boldsymbol{A}$ is invertible(determinant is nonzero.),

$$\begin{aligned}
\boldsymbol{\omega} &= \boldsymbol{A}^{-1} \boldsymbol{b} && (4) \\
&= \frac{1}{\epsilon^2} \begin{pmatrix} \epsilon^2 & -\epsilon \\ -\epsilon & 2 \end{pmatrix} \begin{pmatrix} 2 \\ \epsilon \end{pmatrix} && (5) \\
&= \begin{pmatrix} 1 \\ 0 \end{pmatrix} && (6)
\end{aligned}$$

1.c

$\boldsymbol{A}$ is same to the above one. $\boldsymbol{b} = (1+\epsilon)\cdot\overline{\boldsymbol{x}^{(1)}} + \overline{\boldsymbol{x}^{(2)}} = (2+\epsilon, \epsilon)^T$

$$\begin{aligned}\boldsymbol{\omega} &= \boldsymbol{A}^{-1}\boldsymbol{b} &(7)\\ &= \frac{1}{\epsilon^2}\begin{pmatrix}\epsilon^2 & -\epsilon\\ -\epsilon & 2\end{pmatrix}\begin{pmatrix}2+\epsilon\\ \epsilon\end{pmatrix} &(8)\\ &= \begin{pmatrix}1+\epsilon\\ -1\end{pmatrix} &(9)\end{aligned}$$

1.d

$\boldsymbol{\omega}_b = (1,0)^T, \boldsymbol{\omega}_c = (1.1,-1)^T$. The difference of $\Delta\boldsymbol{\omega} = \boldsymbol{\omega}_c - \boldsymbol{\omega}_b = (\epsilon,-1)^T = (0.1,-1)^T$

2 Linear Regression with Regularization

2.a

Claim 1 : $\boldsymbol{A}$ is positive semi-definite.
proof

$\boldsymbol{A}$ is trivially symmetry matrix. $\forall \boldsymbol{v}\in\mathbb{R}^n, \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} = \sum_n \boldsymbol{v}^T\overline{\boldsymbol{x}^{(n)}}\cdot\overline{\boldsymbol{x}^{(n)}}^T\boldsymbol{v} = \sum_n \|\boldsymbol{v}^T\overline{\boldsymbol{x}^{(n)}}\|^2 \geq 0.\blacksquare$

Claim 2 : $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \iff \boldsymbol{A}^{-1}\boldsymbol{x} = \lambda^{-1}\boldsymbol{x}$ where $\lambda \neq 0$ and $\boldsymbol{A}$ is invertible.
proof

$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \iff \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{A}^{-1}\boldsymbol{x} \iff \lambda^{-1}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{x}.\blacksquare$

Let S(A) be the set of all eigenvalues of A.

$$S(\boldsymbol{A}) \equiv \{\lambda_i | \text{for some } \boldsymbol{x}\in\mathbb{R},\ \boldsymbol{A}\boldsymbol{x} = \lambda_i\boldsymbol{x}\} \tag{10}$$

$\forall\tilde{\lambda}\in S(\boldsymbol{A}+\lambda\boldsymbol{I})$ s.t. $(\boldsymbol{A}+\lambda I)\boldsymbol{x} = \tilde{\lambda}\boldsymbol{x}$.
By multiplying $\boldsymbol{x}^T, \boldsymbol{x}^T(\boldsymbol{A}+\lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \lambda = \tilde{\lambda} \geq \lambda.(\because \boldsymbol{A}$ is positive semi-definite.)
This implies that $\min(S(\boldsymbol{A}+\lambda\boldsymbol{I})) \geq \lambda$. Equivalently,due to the **Claim 2**, this also means that $\max(S((\boldsymbol{A}+\lambda I)^{-1})) \leq \lambda^{-1}$. By noticing that $\max(S((\boldsymbol{A}+\lambda I)^{-1})) = \rho((\boldsymbol{A}+\lambda I)^{-1})$, the proof is done. Note that for the equality, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ must have nontrivial solution.

2.b

For both problems, the $\boldsymbol{A}+\lambda\boldsymbol{I}$ is following.

$$\boldsymbol{A}+\lambda\boldsymbol{I} = \begin{pmatrix}2+\lambda & \epsilon\\ \epsilon & \epsilon^2+\lambda\end{pmatrix} \tag{11}$$

Since (11) is invertible, one can get $\boldsymbol{\omega_b}, \boldsymbol{\omega_c}$.

$$\begin{aligned}\boldsymbol{\omega_b} &= \frac{1}{(1+\lambda)\epsilon^2 + \lambda(\lambda+2)}\cdot\begin{pmatrix}\epsilon^2+\lambda & -\epsilon\\ -\epsilon & 2+\lambda\end{pmatrix}\cdot\begin{pmatrix}2\\ \epsilon\end{pmatrix}\\ &= \frac{1}{(1+\lambda)\epsilon^2 + \lambda(\lambda+2)}\cdot\begin{pmatrix}\epsilon^2+2\lambda\\ \epsilon\lambda\end{pmatrix}\\ &= \begin{pmatrix}0.973\\ 0.044\end{pmatrix}\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\omega_c} &= \frac{1}{(1+\lambda)\epsilon^2+\lambda(\lambda+2)} \cdot \begin{pmatrix} \epsilon^2+\lambda & -\epsilon \\ -\epsilon & 2+\lambda \end{pmatrix} \cdot \begin{pmatrix} 2+\epsilon \\ \epsilon \end{pmatrix} \\
&= \frac{1}{(1+\lambda)\epsilon^2+\lambda(\lambda+2)} \cdot \begin{pmatrix} \epsilon^3+\epsilon^2+\lambda\epsilon+2\lambda \\ -\epsilon^2+\lambda\epsilon \end{pmatrix} \\
&= \begin{pmatrix} 1.026 \\ -0.044 \end{pmatrix}
\end{aligned}
$$

Furthemore, $\Delta\boldsymbol{\omega} = \boldsymbol{\omega_c} - \boldsymbol{\omega_b}$ can be obtained.

$$
\Delta\boldsymbol{\omega} = \frac{1}{(1+\lambda)\epsilon^2+\lambda(\lambda+2)} \begin{pmatrix} \epsilon^3+\lambda\epsilon \\ -\epsilon^2 \end{pmatrix} = \begin{pmatrix} 0.0531 \\ -0.088 \end{pmatrix} \tag{12}
$$

2.c

One can notice that $\Delta\boldsymbol{\omega}$ with regularization is much smaller than $\Delta\boldsymbol{\omega}$ without regularization. This implies that regularization makes the parameters less variable with respect to small noise in input data. This can be verified by the following figure.

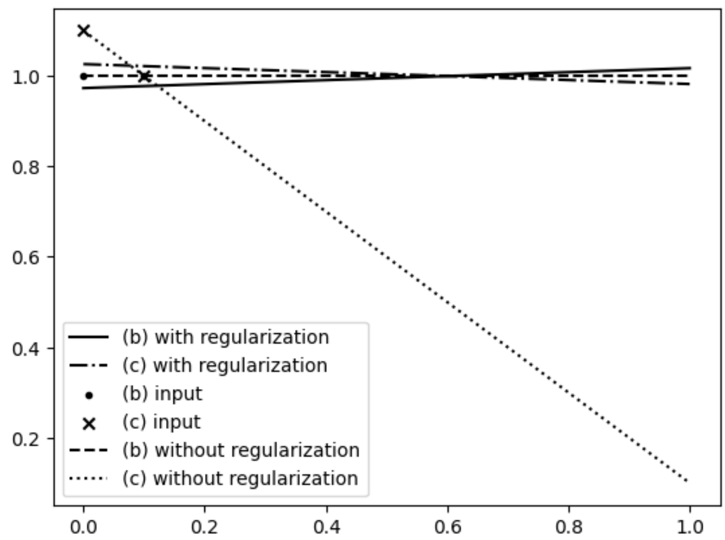

Figure 1: The linear regression result of (b) and (c)

3 LR with Regularization: A Probabilistic Perspective

4 Logistic Regression