

Artificial Intelligence: HW 2

Jeong Min Lee

November 8, 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.

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

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.

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

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

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

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

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

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

1.b

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

$$\boldsymbol{\omega} = \mathbf{A}^{-1} \mathbf{b} \quad (4)$$

$$= \frac{1}{\epsilon^2} \begin{pmatrix} \epsilon^2 & -\epsilon \\ -\epsilon & 2 \end{pmatrix} \begin{pmatrix} 2 \\ \epsilon \end{pmatrix} \quad (5)$$

$$= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (6)$$

1.c

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

$$\boldsymbol{\omega} = \mathbf{A}^{-1}\mathbf{b} \quad (7)$$

$$= \frac{1}{\epsilon^2} \begin{pmatrix} \epsilon^2 & -\epsilon \\ -\epsilon & 2 \end{pmatrix} \begin{pmatrix} 2 + \epsilon \\ \epsilon \end{pmatrix} \quad (8)$$

$$= \begin{pmatrix} 1 + \epsilon \\ -1 \end{pmatrix} \quad (9)$$

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 : \mathbf{A} is positive semi-definite.

proof

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

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

proof

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

Let $S(\mathbf{A})$ be the set of all eigenvalues of \mathbf{A} .

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

$\forall \tilde{\lambda} \in S(\mathbf{A} + \lambda\mathbf{I})$ s.t. $(\mathbf{A} + \lambda\mathbf{I})\mathbf{x} = \tilde{\lambda}\mathbf{x}$.

By multiplying $\mathbf{x}^T, \mathbf{x}^T(\mathbf{A} + \lambda\mathbf{I})\mathbf{x} = \mathbf{x}^T \mathbf{A} \mathbf{x} + \lambda = \tilde{\lambda} \geq \lambda$. ($\because \mathbf{A}$ is positive semi-definite.)

This implies that $\min(S(\mathbf{A} + \lambda\mathbf{I})) \geq \lambda$. Equivalently, due to the **Claim 2**, this also means that $\max(S((\mathbf{A} + \lambda\mathbf{I})^{-1})) \leq \lambda^{-1}$. By noticing that $\max(S((\mathbf{A} + \lambda\mathbf{I})^{-1})) = \rho((\mathbf{A} + \lambda\mathbf{I})^{-1})$, the proof is done. Note that for the equality, $\mathbf{A}\mathbf{x} = \mathbf{0}$ must have nontrivial solution.

2.b

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

$$\mathbf{A} + \lambda\mathbf{I} = \begin{pmatrix} 2 + \lambda & \epsilon \\ \epsilon & \epsilon^2 + \lambda \end{pmatrix} \quad (11)$$

Since (11) is invertible, one can get $\boldsymbol{\omega}_b, \boldsymbol{\omega}_c$.

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

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

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

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

2.c

One can notice that $\Delta\omega$ with regularization is much smaller than $\Delta\omega$ 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.

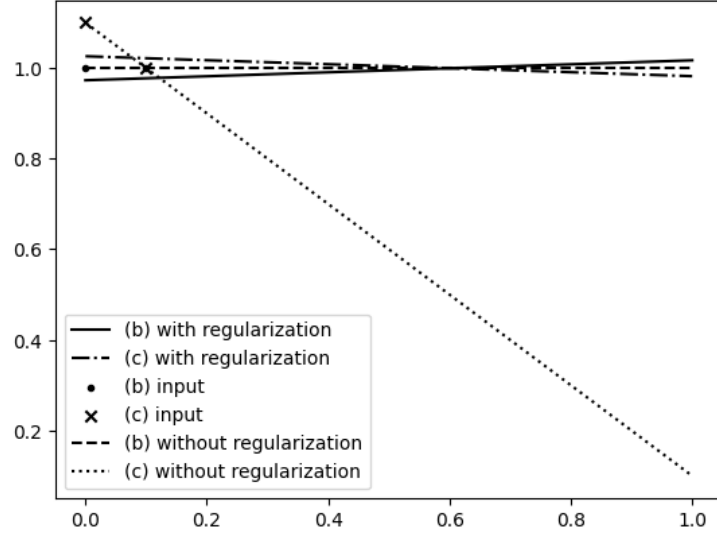


Figure 1: The linear regression result of (b) and (c)

3 LR with Regularization: A Probabilistic Perspective

$$\begin{aligned}
\Pr(\omega) &= \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} \mathbf{I}) \\
&= \frac{\lambda^{N/2}}{2\pi^{N/2}} \exp\left(-\frac{1}{2} \omega^T \lambda \omega\right)
\end{aligned}$$

$$\begin{aligned}
t^{(i)} &= \omega^T \mathbf{x}^{(i)} + \epsilon^{(i)} \quad \& \quad \Pr(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\epsilon^{(i)2}/2\sigma^2\right) \\
\Rightarrow \Pr(t^{(i)} \mid \omega, \mathbf{x}^{(i)}) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(t^{(i)} - \omega^T \mathbf{x}^{(i)})^2/2\sigma^2\right)
\end{aligned}$$

$$\begin{aligned}\Pr(\mathbf{t} \mid \boldsymbol{\omega}^T \mathbf{x}^{(i)}) &= \prod_{i=1}^N \Pr(t^{(i)} \mid \boldsymbol{\omega}^T \mathbf{x}^{(i)}) \\ &= \frac{1}{2\pi^{N/2}\sigma^N} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (t^{(i)} - \boldsymbol{\omega}^T \mathbf{x}^{(i)})^2\right)\end{aligned}$$

From the discussion above, the posterior probability is following.

$$\Pr(\boldsymbol{\omega} \mid \mathbf{x}, \mathbf{t}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (t^{(i)} - \boldsymbol{\omega}^T \mathbf{x}^{(i)})^2 - \frac{\lambda}{2} \boldsymbol{\omega}^T \boldsymbol{\omega}\right) \quad (13)$$

Minimizing (13) is equivalent to minimizing its logarithm.

$$\log(\Pr(\boldsymbol{\omega} \mid \mathbf{x}, \mathbf{t})) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^N (t^{(i)} - \boldsymbol{\omega}^T \mathbf{x}^{(i)})^2 - \frac{\lambda}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} \quad (14)$$

One can see from (14) that maximizing logarithm of posterior is equivalent to minimizing the target function of linear regression with L2 regularization.

4 Logistic Regression

$$\log L(\boldsymbol{\omega}) = -\sum_i \left[t^{(i)} \log\left(\frac{1}{1 + e^{-\boldsymbol{\omega}^T \mathbf{x}^{(i)}}}\right) + (1 - t^{(i)}) \log\left(\frac{e^{-\boldsymbol{\omega}^T \mathbf{x}^{(i)}}}{1 + e^{-\boldsymbol{\omega}^T \mathbf{x}^{(i)}}}\right) \right] \quad (15)$$

To minimize (15), one can differentiate it about $\boldsymbol{\omega}$ to find the $\boldsymbol{\omega}_0$ that makes the derivative zero. By denoting $P^{(n)} = (1 + e^{\boldsymbol{\omega}^T \mathbf{x}^{(n)}})^{-1} = \sigma(\boldsymbol{\omega}^T \mathbf{x}^{(n)})$

$$\begin{aligned}\frac{\partial}{\partial \omega_i} \log L(\boldsymbol{\omega}) &= -\sum_n \frac{\partial}{\partial \omega_i} [t^{(n)} \log P^{(n)} + (1 - t^{(n)}) \log (1 - P^{(n)})] \\ &= -\sum_n \left[\frac{t^{(n)}}{P^{(n)}} - \frac{1 - t^{(n)}}{1 - P^{(n)}} \right] \cdot \frac{\partial P^{(n)}}{\partial \omega_i} \\ &= -\sum_n [t^{(n)}/P^{(n)} - (1 - t^{(n)})/P^{(n)}] \cdot P^{(n)}(1 - P^{(n)}) \cdot \frac{\partial}{\partial \omega_i} \boldsymbol{\omega}^T \mathbf{x}^{(n)} \\ &= -\sum_n [t^{(n)}(1 - P^{(n)}) - (1 - t^{(n)})P^{(n)}] x_i^{(n)} \\ &= -\sum_n (t^{(n)} - P^{(n)}) x_i^{(n)} = 0\end{aligned}$$

Note that $d\sigma(x)/dx = \sigma(x)(1 - \sigma(x))$. The result of the last equation above can be written in vector form.

$$\sum_n (t^{(n)} - \sigma(\boldsymbol{\omega}^T \mathbf{x})) \mathbf{x}^{(n)} = 0 \quad (16)$$

The equation (16) can be written as following equation.

$$\sum_n t^{(n)} \mathbf{x}^{(n)} = \sum_n \sigma(\boldsymbol{\omega}^T \mathbf{x}) \mathbf{x}^{(n)} \quad (17)$$

Unfortunately, since sigmoid function σ is non-linear function, the equation (17) is non-linear equation that doesn't have the closed form of solution in general.