Homework 3

2023-11398 Yun sukmin

November 15, 2024

1 Linear Regression

1.1 Problem - (a)

$$\min_w \frac{1}{2}\sum_{n=1}^{N}(y_n - w^T\bar{x}_n)^2 = \min_w \frac{1}{2}\sum_{n=1}^{N}(w^T\bar{x}_n - y_n)^2$$

$$\begin{aligned}
\nabla_w E(w) &= \sum_{n=1}^{N}(w^T\bar{x}_n - y_n)\bar{x}_n \\
&= \sum_{n=1}^{N}(w^T\bar{x}_n)\bar{x}_n - \sum_{n=1}^{N}(y_n)\bar{x}_n \\
&= \sum_{n=1}^{N}\bar{x}_n\bar{x}_n^T w - \sum_{n=1}^{N}y_n\bar{x}_n \\
&= \left(\sum_{n=1}^{N}\bar{x}_n\bar{x}_n^T\right) w - \left(\sum_{n=1}^{N}y_n\bar{x}_n\right) \\
&= Aw - b = 0 \\
A &= \sum_{n=1}^{N}\bar{x}_n\bar{x}_n^T,\ b = \sum_{n=1}^{N}t_n\bar{x}_n
\end{aligned}$$

Thus:

$$A = X^TX, b = X^Ty$$

where X is the matrix with rows $\bar{x}_n^T$, and T is the vector of target values t_n. This linear system can now be solved to obtain w as:

$$w = A^{-1}b = (X^TX)^{-1}X^T$$

1.2 Problem - (b)

Let's define the training data matrix X and vector T as follows:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} \quad T = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Using the expression derived in (a), we have:

$$A = X^T X = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} = \begin{bmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix} \quad b = X^T T = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$

Now, the equation Aw = b becomes:

$$\begin{bmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$

two equations:

$$2w_0 + \epsilon w_1 = 2$$

$$\epsilon w_0 + \epsilon^2 w_1 = \epsilon$$

Dividing the second equation by $\epsilon(\epsilon \neq 0)$, we get:

$$w_0 + \epsilon w_1 = 1$$

Subtracting the above equation from the first equation, we get:

$$w_0 = 1$$

we can substitute w_0 into the first equation:

$$2 + \epsilon w_1 = 2$$

$$w_1 = 0$$

Thus, the solution to the linear regression problem is:

$$w = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

1.3 Problem - (c)

Let's define the training data matrix X and vector T as follows:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} \quad T = \begin{bmatrix} 1 + \epsilon \\ 1 \end{bmatrix}$$

Using the expression derived in (a), we have:

$$A = X^T X = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} = \begin{bmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix} \quad b = X^T T = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 + \epsilon \\ 1 \end{bmatrix} = \begin{bmatrix} 2 + \epsilon \\ \epsilon \end{bmatrix}$$

Now, the equation Aw = b becomes:

$$\begin{bmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2 + \epsilon \\ \epsilon \end{bmatrix}$$

two equations:

$$2w_0 + \epsilon w_1 = 2 + \epsilon$$
$$\epsilon w_0 + \epsilon^2 w_1 = \epsilon$$

Dividing the second equation by $\epsilon (\epsilon \neq 0)$, we get:

$$w_0 + \epsilon w_1 = 1$$

Subtracting the above equation from the first equation, we get:

$$w_0 = 1 + \epsilon$$

we can substitute w_0 into the first equation:

$$2 + 2\epsilon + \epsilon w_1 = 2 + \epsilon$$

$$w_1 = -1$$

Thus, the solution to the linear regression problem is, and substitute $\epsilon = 0.1$ to the solution is:

$$w = \begin{bmatrix} 1+\epsilon \\ -1 \end{bmatrix} = \begin{bmatrix} 1.1 \\ -1 \end{bmatrix}$$

1.4 Problem - (d)

$$w_b = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad w_c = \begin{bmatrix} 1.1 \\ -1 \end{bmatrix}$$

So difference between w_b and w_c is:

$$Difference = w_c - w_b = \begin{bmatrix} 1.1 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ -1 \end{bmatrix}$$

2 Linear Regression with Regularization

2.1 Problem - (a)

Let's show that A is positive semidefinite matrix.
Defining in Problem 1, Let $A = X^T X$ where X is N by D matrix.
we need to show that for any vector $v \in \mathbb{R}^D$, $v^T A v \geq 0$.

$$\begin{aligned} v^T A v &= v^T X^T X v \\ &= (Xv)^T X v \\ &= ||Xv||^2 \geq 0 \end{aligned}$$

so A is positive semidefinite matrix.
Since that all of eigenvalues of matrix A are non-negative. When add λI to A, all of eigenvalues of matrix A are at least λ.

$$Eigenvalues\ of\ (A + \lambda I) \geq \lambda$$

Eigenvalues of $(A+\lambda I)^{-1}$ are the reciprocals of the eigenvalues of $A+\lambda I$. eigenvalues of $A+\lambda I$ are at least λ, so eigenvalues of $(A+\lambda I)^{-1}$ are at most $\frac{1}{\lambda}$ so $\sigma((A+\lambda I)^{-1}) \leq \frac{1}{\lambda}$ is right.

2.2 Problem - (b)

in Problem 1-(b), training datas are:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} \quad T = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Constructing A and b as in Problem 1-(b), we have:

$$A = X^T X = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} = \begin{bmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix} \quad b = X^T T = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$

Now, the equation $(A+\lambda I)w = b$ becomes:

$$\begin{bmatrix} 2+\lambda & \epsilon \\ \epsilon & \epsilon^2+\lambda \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$

two equations:

$$(2+\lambda)w_0 + \epsilon w_1 = 2$$
$$\epsilon w_0 + (\epsilon^2+\lambda)w_1 = \epsilon$$

Substituting $\epsilon = 0.1, \lambda = 0.05$ into two equations:

$$2.05w_0 + 0.1w_1 = 2$$

$$0.1w_0 + 0.06w_1 = 0.1$$

Simplify two equations:

$$123w_0 + 6w_1 = 120$$
$$10w_0 + 6w_1 = 10$$

Subtracting the second equation from the first equation:

$$113w_0 = 110 \quad w_0 = \frac{110}{113}$$

And substituting w_0 in upper equaions, we gets:

$$w_1 = \frac{5}{113}$$

So, vector w is:

$$w = \frac{5}{113} \begin{bmatrix} 22 \\ 1 \end{bmatrix}$$

in Problem 1-(c), training datas are:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} \quad T = \begin{bmatrix} 1+\epsilon \\ 1 \end{bmatrix}$$

Constructing A and b as in Problem 1-(b), we have:

$$A = X^T X = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & \epsilon \end{bmatrix} = \begin{bmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix} \quad b = X^T T = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1+\epsilon \\ 1 \end{bmatrix} = \begin{bmatrix} 2+\epsilon \\ \epsilon \end{bmatrix}$$

Now, the equation $(A + \lambda I)w = b$ becomes:

$$\begin{bmatrix} 2+\lambda & \epsilon \\ \epsilon & \epsilon^2 + \lambda \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2+\epsilon \\ \epsilon \end{bmatrix}$$

two equations:

$$(2+\lambda)w_0 + \epsilon w_1 = 2 + \epsilon$$

$$\epsilon w_0 + (\epsilon^2 + \lambda)w_1 = \epsilon$$

Substituting $\epsilon = 0.1, \lambda = 0.05$ into two equations:

$$2.05 w_0 + 0.1 w_1 = 2.1$$

$$0.1 w_0 + 0.06 w_1 = 0.1$$

Simplify two equations:

$$123 w_0 + 6 w_1 = 126$$

$$10 w_0 + 6 w_1 = 10$$

Subtracting the second equation from the first equation:

$$113 w_0 = 116 \quad w_0 = \frac{116}{113}$$

And substituting w_0 in upper equaions, we gets:

$$w_1 = \frac{-5}{113}$$

So, vector w is:

$$w = \frac{1}{113} \begin{bmatrix} 116 \\ -5 \end{bmatrix}$$

Difference in between upper two setups is:

$$w_c - w_b = \frac{1}{113} \begin{bmatrix} 116 \\ -5 \end{bmatrix} - \frac{5}{113} \begin{bmatrix} 22 \\ 1 \end{bmatrix} = \frac{1}{113} \begin{bmatrix} 116 - 110 \\ -5 - 5 \end{bmatrix} = \frac{1}{113} \begin{bmatrix} 6 \\ -10 \end{bmatrix}$$

2.3 Problem - (c)

Difference in Problem 1-(d) is $\begin{bmatrix} 0.1 \\ -1 \end{bmatrix}$, and difference in Problem 2-(c) is $\frac{1}{113}\begin{bmatrix} 6 \\ -10 \end{bmatrix}$.
The difference is due to the addition of epsilon to the data. Therefore, the difference indicates how sensitive the model is to the addition of a small noise value, epsilon.
The difference in Problem 1-(d) is larger than the difference in Problem 2-(c). This is because the model in Problem 1-(d) is more sensitive to the addition of a small noise value, epsilon, than the model in Problem 2-(c).
This means that the regularization term λI helps to prevent the model from becoming overly sensitive to small variations or noise in the target values.
So, the regularization term allows the the model to maintain stability and robustness.

3 LR with Regularization: A Probabilistic Perspective

By using Bayes' rule, the posterior probability is proportional to the product of the prior probability and the likelihood:

$$Pr(w|X,y) \propto Pr(y|X,w)Pr(w)$$

Given Pr(w) follows a Gaussian distribution with mean 0 and covariance matrix $\frac{1}{\lambda}I$, the prior probability is:

$$Pr(w) = N(0, \frac{1}{\lambda}I) = \frac{1}{(\frac{2\pi}{\lambda})^{d/2}} exp(-\frac{\lambda}{2}w^T w)$$

where d is dimension of w. Given the likelihood $Pr(y|X,w)$ is the probability of the ovsered data given the weights w:

$$\begin{aligned} Pr(y|X,w) &= N(y|Xw, \sigma^2 I) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} exp(-\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)) \\ &= \prod_{n=1}^{N} N(y_n|w^T \bar{x}_n, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} exp(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^T \bar{x}_n)^2) \end{aligned}$$

For maximizing the posterior probability, we need to maximize the log of the posterior probability:

$$\begin{aligned}
log(Pr(w|X,y)) &\propto log(Pr(y|X,w)Pr(w)) \\
&= log(Pr(y|X,w)) + log(Pr(w)) \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^T\bar{x}_n)^2 - \frac{\lambda}{2}w^Tw + C \\
&= -(\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^T\bar{x}_n)^2 + \frac{\lambda}{2}w^Tw)
\end{aligned}$$

For maximizing the posterior probability, we minimize the objective function of linvear regression with L2 regularization.
So finding the best w by maximizing the posterior probability is equivalent to linear regression with L2 regularization.

4 Logistic Regression

$$L(w) = -\sum_{n=1}^{N} t_n log(\sigma(w^T\bar{x}_n)) + (1-t_n)log(1-\sigma(w^T\bar{x}_n))$$

Differentiating the loss function with respect to w, we get:

$$\begin{aligned}
\frac{\partial}{\partial w}L(w) &= -\sum_{n=1}^{N}\left(t_n\frac{1}{\sigma(w^T\bar{x}_n)}\frac{\partial}{\partial w}\sigma(w^T\bar{x}_n) - (1-t_n)\frac{1}{1-\sigma(w^T\bar{x}_n)}\frac{\partial}{\partial w}\sigma(w^T\bar{x}_n)\right) \quad (chain-rule) \\
&= -\sum_{n=1}^{N}\left(t_n\frac{1}{\sigma(w^T\bar{x}_n)} - (1-t_n)\frac{1}{1-\sigma(w^T\bar{x}_n)}\right)\cdot\frac{\partial}{\partial w}\sigma(w^T\bar{x}_n) \\
&= -\sum_{n=1}^{N}\left(t_n\frac{1}{\sigma(w^T\bar{x}_n)} - (1-t_n)\frac{1}{1-\sigma(w^T\bar{x}_n)}\right)\sigma(w^T\bar{x}_n)(1-\sigma(w^T\bar{x}_n))\frac{\partial}{\partial w}w^T\bar{x}_n \quad (chain-rule) \\
&= -\sum_{n=1}^{N}\left(t_n\frac{1}{\sigma(w^T\bar{x}_n)} - (1-t_n)\frac{1}{1-\sigma(w^T\bar{x}_n)}\right)\sigma(w^T\bar{x}_n)(1-\sigma(w^T\bar{x}_n))\bar{x}_n \\
&= -\sum_{n=1}^{N}\big(t_n(1-\sigma(w^T\bar{x}_n)) - (1-t_n)\sigma(w^T\bar{x}_n)\big)\,\bar{x}_n \\
&= -\sum_{n=1}^{N}\big(t_n - t_n\sigma(w^T\bar{x}_n) - \sigma(w^T\bar{x}_n) + t_n\sigma(w^T\bar{x}_n)\big)\,\bar{x}_n \\
&= -\sum_{n=1}^{N}\big(t_n - \sigma(w^T\bar{x}_n)\big)\,\bar{x}_n
\end{aligned}$$

Thus, the gradient of the loss function is:

$$\frac{\partial}{\partial w} L(w) = -\sum_{n=1}^{N} \left(t_n - \sigma(w^T \bar{x}_n) \right) \bar{x}_n$$

setting the gradient to zero, we can solve for w.
Here, $\sigma(w^T \bar{x}_n) = \frac{1}{1+exp(-w^T \bar{x}_n)}$, sigmoid function is nonlinear functions on w, this means that we can't isolate w in a closed-form expression, because the sigmoid introduces a dependency on w that is nonlinear and not easily inverted.