EECS 545: Homework #1

Mingliang Duanmu
duanmuml@umich.edu

January 25, 2022

1 Linear regression on a polynomial

a

i

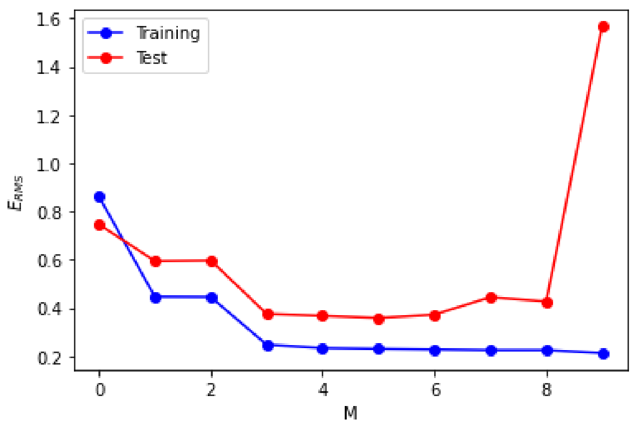
For both methods, we have iterations = 1000, learning rate = 0.01, and degree = 2.
For batch gradient descent, we have coefficients = array([-2.82412688, 1.94687097]).
For stochastic gradient descent, we have coefficients = array([-2.82979555, 1.942896]).

ii

We use the same parameters as before, iterations = 1000, learning rate = 0.01, and degree = 2.
Since the training data is relatively small, we do not notice a distinct difference in speed of convergence, but by printing out the E_{MS} for the first 100 epochs, we find batch gradient descent converges a little faster than stochastic gradient descent.

b

i



ii

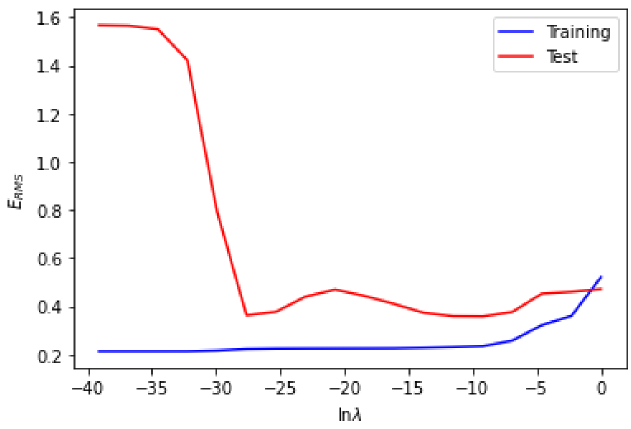
From the plot we find the polynomial with $M = 5$ best fits the data since the RMS error of the test data is minimized. If the degree is high ($M = 9$), we can see the testing error is much greater than training error, which is an over-fitting. When the degree is too small ($M < 3$), we can see both training and testing error are great, which is an under-fitting.

c

i

The closed form solution for ridge regression is

$$\mathbf{w} = (\mathbf{\Phi}^T \mathbf{\Phi} + \lambda \mathbf{I})^{-1} \mathbf{\Phi}^T \mathbf{Y}$$



ii

From the plot above we observe that when $\lambda = 10^{-4}$ the test error is minimized.

2 Locally weighted linear regression

a

$$E_D(\mathbf{w}) = (X\mathbf{w} - \mathbf{y})^T R(X\mathbf{w} - \mathbf{y}) = \mathbf{z}^T R\mathbf{z} = \sum_{i=1}^{N} R^{(i)} z_i^2 = \frac{1}{2}\sum_{i=1}^{N} r^{(i)}(z_i)^2$$

where $z_i = \mathbf{w}^T x^{(i) - y^{(i)}}$. So we have

$$R = \frac{1}{2} diag(r_1, r_2, \cdots, r_N)$$

b

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} r^{(i)}(\mathbf{w}^T x^{(i)})^2 - \sum_{i=1}^{N} r^{(i)} y^{(i)} \mathbf{w}^T (x^{(i)})^2 + \frac{1}{2}\sum_{i=1}^{N} r^{(i)}(y^{(i)})^2$$
$$= \frac{1}{2}\mathbf{w}^T X^T \mathbf{R} X\mathbf{w} - \mathbf{w}^T X^T \mathbf{R} Y + \frac{1}{2} Y^T \mathbf{R} Y$$

By calculating the gradient of expectation and set to zero,

$$\nabla E_D(\mathbf{w}) = X^T \mathbf{R} X\mathbf{w} - X^T \mathbf{R} Y = 0$$

So we have the close form

$$\mathbf{w} = (X^T \mathbf{R} X)^{-1} X^T \mathbf{R} X$$

c

$$\log P(Y|X, \mathbf{w}) = \log P(y^{(1)}, y^{(2)}, \cdots, y^{(N)}|X, \mathbf{w})$$
$$= \sum_{i=1}^{N} \log[\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\|y^{(i)} - W^T x^{(i)}\|^2}{2(\sigma^{(i)})^2})]$$
$$= \sum_{i=1}^{N} (-\frac{1}{2}\log(2\pi) - \log(\sigma^{(i)}) - \frac{\|y^{(i)} - W^T x^{(i)}\|^2}{2(\sigma^{(i)})^2})$$
$$= -\frac{N}{2}\log(2\pi) - \sum_{i=1}^{N} \log(\sigma^{(i)} - \sum_{i=1}^{N} \frac{\|y^{(i)} - W^T x^{(i)}\|^2}{2(\sigma^{(i)})^2})$$

Let $\beta^{(i)} = \frac{1}{\sigma^{(i)}}$ and calculate the gradient and set to zero,

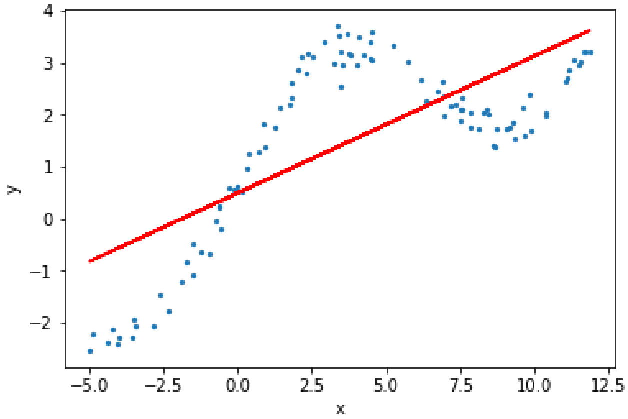
$$\nabla \log P(Y|X, \mathbf{w}) = \sum_{i=1}^{N} X(y^{(i)} - \mathbf{w}^T) x^{(i)} x^{(n)} \beta^{(i)}$$

$$= \sum_{i=1}^{N} y^{(i)} x^{(i)} \beta^{(i)} - x^{(i)} (x^{(i)})^T \mathbf{w} \beta^{(i)} = X^T \beta Y - X^T \beta X\mathbf{w} = 0$$

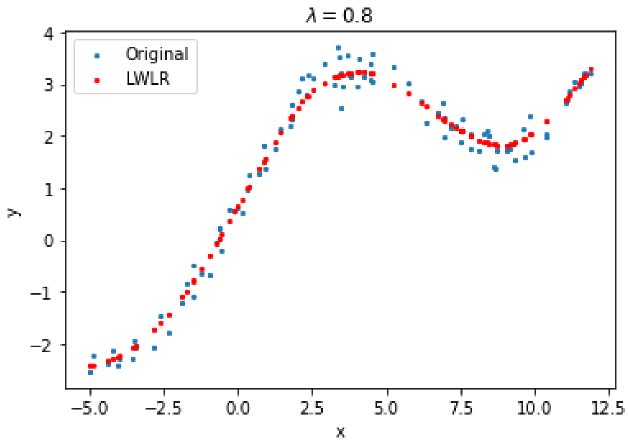
The maximum likelihood estimate of $\mathbf{w}$ can be written as $(X^T \beta X)^{-1} X^T \beta Y$, where $\beta = diag(\frac{1}{\sigma^{(1)}}, \frac{1}{\sigma^{(1)}}, \cdots, \frac{1}{\sigma^{(N)}})$ is a diagonal matrix like $\mathbf{R}$ in the previous question. Therefore, finding the maximum likelihood estimate of $\mathbf{w}$ reduces to solving a weighted linear regression problem.

d

i

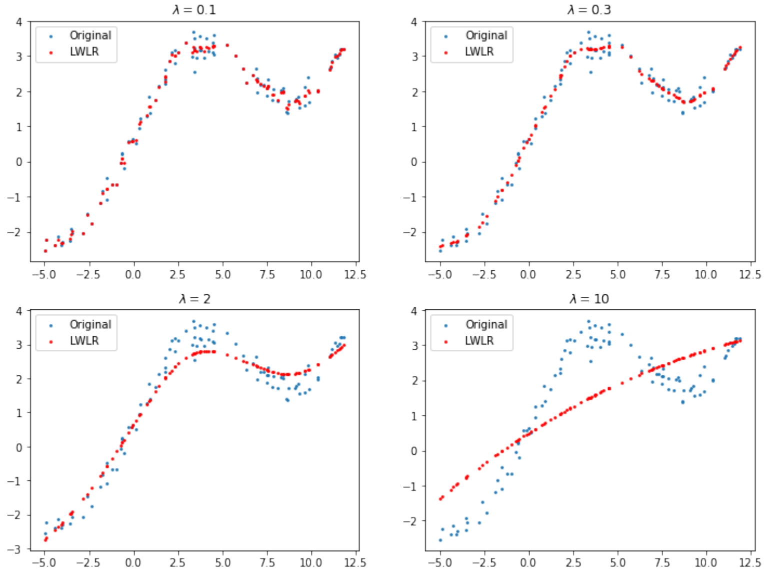


ii



iii

When τ is too small, an over-fit occurs. When τ is too large, an under-fit occurs.



3 Derivation and Proof

a

Suppose we have $\widehat{Y}_i = \omega_0 + \omega_1 X_i$. To minimize the error

$$E = \sum_{i=1}^{N}(Y - \widehat{Y}_i)^2 = \sum_{i=1}^{N}(Y - \omega_0 - \omega_1 X_i)^2$$

we do partial differentiation for ω_0 and ω_1 respectively:

$$\frac{\partial E}{\partial \omega_0} = -2\sum_{i=1}^{N}(Y_i - \omega_0 - \omega_1 X_i) = 0$$

$$\frac{\partial E}{\partial \omega_1} = -2\sum_{i=1}^{N}(Y_i - \omega_0 - \omega_1 X_i)X_i = 0$$

Solving the first equation we can get

$$\sum_{i=1}^{N} Y_i - N\omega_0 - \omega_1 \sum_{i=1}^{N}(X_i)$$

Thus,

$$\omega_0 = \frac{1}{N}(\sum_{i=1}^{N} Y_i - \omega_1 \sum_{i=1}^{N} X_i) = \bar{Y} - \omega_1 \bar{X}$$

Similarly, solving the second equation we can get

$$\sum_{i=1}^{N} X_i Y_i - \omega_0 \sum_{i=1}^{N} X_i - \omega_1 \sum_{i=1}^{N} X_i^2 = 0$$

Thus,

$$\omega_1 = \frac{\sum_{i=1}^{N} X_i Y_i - \bar{Y}\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} X_i^2 - \bar{X}\sum_{i=1}^{N} X_i}$$

b

i

Necessity:
According to the question, $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_N)$, so we have for any $\mathbf{z} \neq 0$

$$\mathbf{z}^T\mathbf{A}\mathbf{z} = \mathbf{z}^T\mathbf{U}\Lambda\mathbf{U}^T\mathbf{z} = \mathbf{z}'\mathbf{\Lambda}\mathbf{z}'^T = \sum_{i=1}^{N}\lambda_i(z_i')^2$$

Since $\lambda_i > 0$, we can prove $\mathbf{A}$ is PD.
Sufficiency:
As $\mathbf{A}$ is symmetric and PD, according to the concept of eigenvalue, we have $\mathbf{A}x = \lambda x$. We multiply x^T on both sides and get

$$x^T\mathbf{A}\mathbf{X} = \lambda x^T x = \lambda\sum_{i=0}^{N}x_i^2 > 0$$

Therefore we prove for any $\lambda_i > 0$, $\mathbf{A}$ is PD.

ii

Since $\mathbf{\Phi}^T\mathbf{\Phi}$ is symmetric, we can express it as $\mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_N)$. So we have

$$\mathbf{\Phi}^T\mathbf{\Phi} + \beta\mathbf{I} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T + \beta\mathbf{U}\mathbf{I}\mathbf{U}^T = \mathbf{U}(\mathbf{\Lambda} + \beta\mathbf{I})\mathbf{U}^T$$

Therefore we prove the ridge regression has an effect of shifting all singular values by a constant β.
Let $\mathbf{\Lambda}' = \mathbf{\Lambda} + \beta\mathbf{I}$, since $\beta > 0$, $\mathbf{\Lambda}' = diag(\lambda_1', \lambda_2', \cdots, \lambda_N')$ where $\lambda_i' > 0$. So we have for any $\mathbf{z} \neq 0$

$$\mathbf{z}^T(\mathbf{\Phi}^T\mathbf{\Phi} + \beta\mathbf{I})\mathbf{z} = \mathbf{z}^T\mathbf{U}\mathbf{\Lambda}'\mathbf{U}^T\mathbf{z} = \lambda_i'\sum_{i=1}^{N}z_i^2$$

where $\mathbf{z}' = \mathbf{z}^T\mathbf{U}$.
Therefore we prove $\mathbf{\Phi}^T\mathbf{\Phi} + \beta\mathbf{I}$ is PD.