Wensong Hu
EECS
University of Michigan

January 30, 2024

HW1 — Linear Regression

1 Derivation and Proof

1.1 Close form solution

The sum square error loss function is:

$$L(w) = \frac{1}{2}\sum_{n=1}^{N}(w\phi(x^{(n)}) - y^{(n)})^2 \tag{1-1}$$

We know that $\phi(x) = [1, x]^T$, so the following equation is derived:

$$\begin{aligned}
L(w) &= \frac{1}{2}\sum_{n=1}^{N}(w_0 + w_1x^{(n)} - y^{(n)})^2 && \text{(1-2)}\\
&= \frac{1}{2}\sum_{n=1}^{N}((w_0 + w_1x^{(n)})^2 + (y^{(n)})^2 - 2(w_0 + w_1x^{(n)})y^{(n)}) && \text{(1-3)}\\
&= \frac{N}{2}w_0^2 + \sum_{n=1}^{N}(w_0w_1x^{(n)} + \frac{1}{2}w_1^2(x^{(n)})^2 + \frac{1}{2}(y^{(n)})^2 - w_0y^{(n)} - w_1x^{(n)}y^{(n)}) && \text{(1-4)}
\end{aligned}$$

Take derivative with respect to w_0 and w_1 of the loss function:

$$\nabla_{w_0}L(w) = Nw_0 + \sum_{n=1}^{N}w_1(x^{(n)})^2 - \sum_{n=1}^{N}y^{(n)} \tag{1-5}$$

$$\nabla_{w_1}L(w) = \sum_{n=1}^{N}w_0x^{(n)} + \sum_{n=1}^{N}w_1(x^{(n)})^2 - \sum_{n=1}^{N}x^{(n)}y^{(n)} \tag{1-6}$$

In order to find w_0, let the derivative with respect to w_0 equals 0:

$$\begin{aligned}
\nabla_{w_0}L(w) = Nw_0 + \sum_{n=1}^{N}w_1(x^{(n)})^2 - \sum_{n=1}^{N}y^{(n)} &= 0 && \text{(1-7)}\\
\implies w_0 &= \frac{1}{N}\sum_{n=1}^{N}y^{(n)} - \frac{1}{N}\sum_{n=1}^{N}w_1x^{(n)} && \text{(1-8)}\\
&= \bar{Y} - w_1\bar{X} && \text{(1-9)}
\end{aligned}$$

Where $\bar{X}$ is the mean of x data, and $\bar{Y}$ is the mean of y data.
Toke derivative w.r.t. w_1 and let it equals to 0, and plug w_0 in:

$$\nabla_{w_1} L(w) = \sum_{n=1}^{N} w_0 x^{(n)} + \sum_{n=1}^{N} w_1 (x^{(n)})^2 - \sum_{n=1}^{N} x^{(n)} y^{(n)} = 0 \tag{1-10}$$

$$\implies w_1 = \frac{\sum_{n=1}^{N} x^{(n)} y^{(n)} - \sum_{n=1}^{N} w_0 x^{(n)}}{\sum_{n=1}^{N} (x^{(n)})^2} \tag{1-11}$$

$$= \frac{\frac{1}{N}\sum_{n=1}^{N} x^{(n)} y^{(n)} - \bar{Y}\bar{X}}{\frac{1}{n}\sum_{n=1}^{N} (x^{(n)})^2 - \bar{X}^2} \tag{1-12}$$

1.2 Positive definite

1.2.1 Prove A is PD iff $\lambda_i > 0$ for each i

First, prove if $\lambda_i > 0$, a symmetric matrix A is PD:

$$z^T A z = z^T U \Lambda U^T z \tag{1-13}$$

$$\tag{1-14}$$

Where, U is an orthogonal matrix, so $UU^T = I$, Λ is diagonal matrix with all value is eigenvalues.

$$z^T A z = \sum_i z^T u_i \lambda_i u_i^T z \tag{1-15}$$

$$= \sum_i z^T u_i u_i^T z \lambda_i \tag{1-16}$$

$$= \sum_i z^T z \lambda_i \tag{1-17}$$

$$= \sum_i ||z||_2^2 \lambda_i \tag{1-18}$$

Since all $\lambda_i > 0$, the conclusion can be drawn $z^T A z > 0$, thus, A is positive definite. Second, prove if a symmetric matrix A is PD, the all the eigenvalues $\lambda_i > 0$:

$$z^T A z > 0 \tag{1-19}$$

$$\implies z^T U \Lambda U^T z > 0 \tag{1-20}$$

$$\implies \sum_i z^T u_i \lambda_i u_i^T z > 0 \tag{1-21}$$

$$\implies \sum_i z^T u_i u_i^T z \lambda_i > 0 \tag{1-22}$$

$$\implies \sum_i ||z||_2^2 \lambda_i > 0 \tag{1-23}$$

Therefore all the eigenvalues should be bigger than 0.
Thus: A is PD $\iff$ $\lambda_i > 0$

1.2.2 Derive eigenvalue and eigenvectors for regularized close form solution

$$\Phi^T\Phi + \beta I = U\Lambda U^T + \beta I \tag{1-24}$$

For each eigenvalue and eigenvector:

$$u_i u_i^T \lambda_i + \beta = \lambda_i + \beta \tag{1-25}$$

Thus, eigenvalue of $\Phi^T\Phi + \beta I$ is $\lambda_i + \beta$, and the eigenvectors are also u_i
To prove matrix $\Phi^T\Phi + \beta I$ is PD when $\beta > 0$, first construct the quadratic form:

$$\begin{aligned} z^T(\Phi^T\Phi + \beta)z &= z^T(U\lambda + \beta I U^T)z && \text{(1-26)}\\ &= \sum_i (z^T u_i u_i^T z)(\lambda_i + \beta) && \text{(1-27)}\\ &= \sum) i ||z||_2^2(\lambda_i + \beta) && \text{(1-28)} \end{aligned}$$

Since $\lambda_i > 0$ for all i, only $\beta > 0$, $z^T(\Phi^T\Phi + \beta)z$ would be positive. Thus matrix $\Phi^T\Phi + \beta I$ is PD if $\beta > 0$.

1.3 Prove relation between MLE and logistic regression

Since the label is -1 and 1, the log-likelihood could be expressed differently:

$$\sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)}) = \sum_{n=1}^{N} \log \sigma(y^{(n)} \cdot w^T\phi(x^{(n)})) \tag{1-29}$$

Where:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1-30}$$

In this scenario, the class labels are $y \in \{-1, +1\}$. The equation needs to account for the fact that the output of the logistic regression model should reflect these two possibilities. The term $y \cdot w^T x$ does exactly that by flipping the sign of $w^T x$ when $y = -1$, which effectively models the correct probability for each class.
Plug sigmoid function in, so we can get the transformed log-likelihood:

$$\sum_{n=1}^{N} \log \frac{1}{1 + e^{-y^{(n)} \cdot w^T\phi(x^{(n)})}} = \sum_{n=1}^{N} -\log(1 + e^{-y^{(n)} \cdot w^T\phi(x^{(n)})}) \tag{1-31}$$

Statement proved.

2 Linear regression on a polynomial

2.1 GD and SGD

2.1.1 Code implementation

Submitted to Autograder

2.1.2 Plot

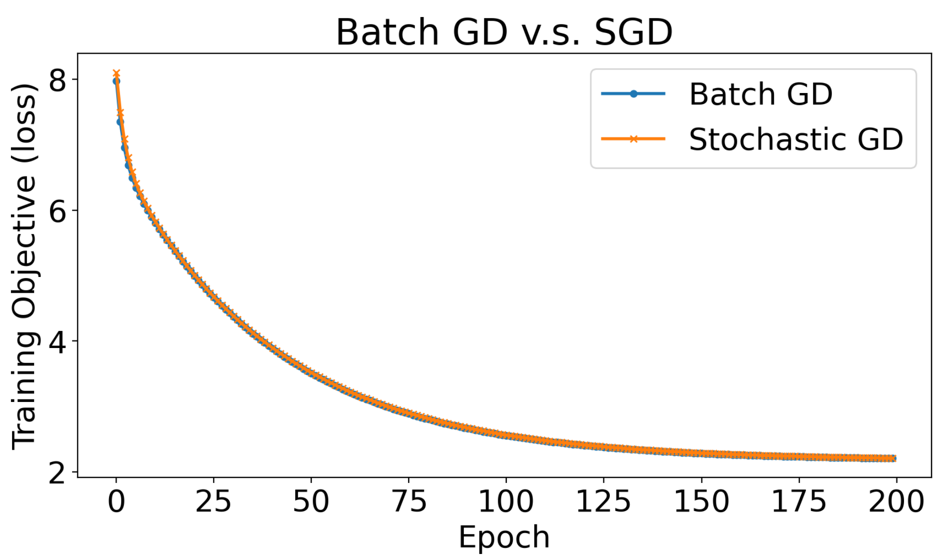
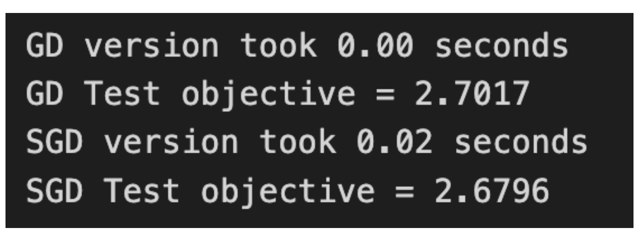


Figure 1: Loss for batch GD and SGD



```
GD version took 0.00 seconds
GD Test objective = 2.7017
SGD version took 0.02 seconds
SGD Test objective = 2.6796
```

Figure 2: Terminal output for training on GD and SGD

GD takes less time, but SGD shows lower test objective.

2.2 Over-fitting study

2.2.1 Code implementation

Submitted to Autograder

2.2.2 Plot

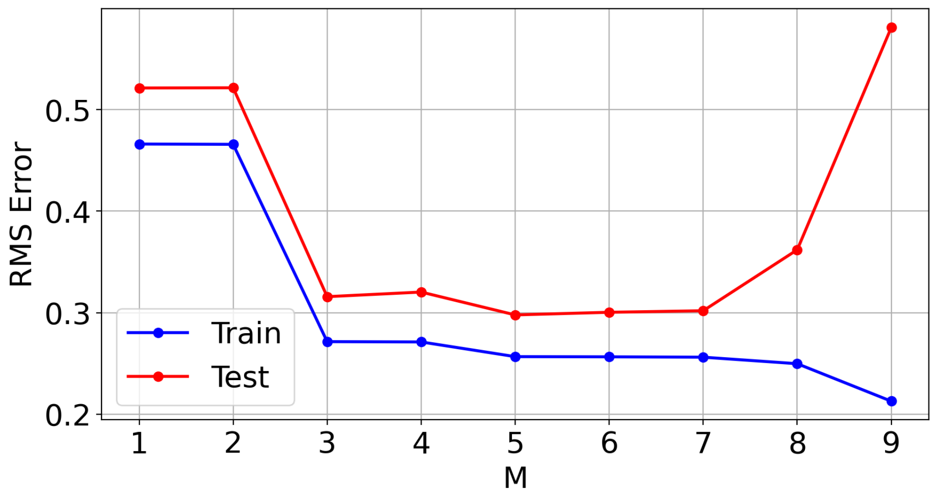


Figure 3: RMSE with different polynomial order

2.2.3 Discussion

The 5 degree polynomial fits the data best, since both the training and testing RMSE are lowest.
The 0 degree polynomial underfits the data, since the training and testing RMSE are high.
The 8 degree and 9 degree polynomial overfits the data, since the training RMSE is low, but teating RMSE are high.

2.3 Ridge Regression

2.3.1 Code implementation

Submitted to Autograder

2.3.2 Plot

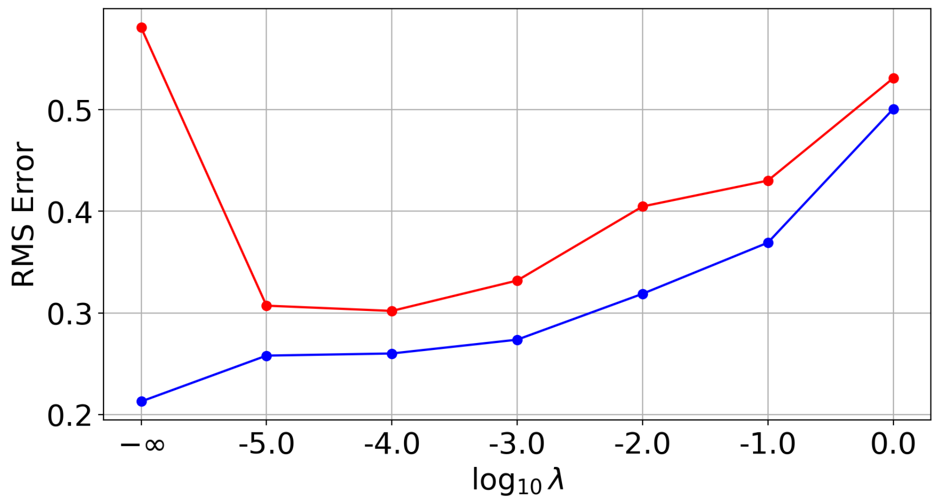


Figure 4: RMSE with different regularization factor

2.3.3 Discussion

When $\log\lambda = -4$,the 9degree polynomial has the best performance, as both training and testing error are kept low, and increases after that.

3 Locally weighted linear regression

3.1 Matrix form loss function

$$
\begin{aligned}
E_D(w) &= \frac{1}{2}\sum_{i=1}^{N} r^{(i)}(w^T x^{(i)} - y^{(i)})^2 && \text{(1-32)}\\
&= \sum_{n=1}^{N}(w^T x^{(i)} - y^{(i)})\frac{r^{(i)}}{2}(w^T x^{(i)} - y^{(i)}) && \text{(1-33)}\\
&= (w^T X - Y^T)\begin{bmatrix} \frac{r^{(1)}}{2} & & \\ & \ddots & \\ & & \frac{r^{(N)}}{2} \end{bmatrix}(w^T X - Y^T)^T && \text{(1-34)}\\
&= (w^T X - Y^T)R(w^T X - Y^T)^T && \text{(1-35)}
\end{aligned}
$$

Where matrix $R = \begin{bmatrix} \frac{r^{(1)}}{2} & & \\ & \ddots & \\ & & \frac{r^{(N)}}{2} \end{bmatrix}$.

3.2 Close form solution

$$\nabla E_D(w) = \nabla(w^T X - Y^T) R (w^T X - Y^T)^T \tag{1-36}$$
$$= \nabla(w^T X R - Y^T R)(w^T X - Y^T)^T \tag{1-37}$$
$$= \nabla(w^T X T X^T w - w^T X R Y - Y^T R X^T w - Y^T R Y) \tag{1-38}$$
$$= \nabla(w^T X T X^T w - 2 w^T X R Y - Y^T R Y) \tag{1-39}$$
$$= 2 X R^T X^T w - 2 X R Y \tag{1-40}$$

$$\nabla E_D(w) = 0 \tag{1-41}$$
$$\Longrightarrow X R^T X^T w = X R Y \tag{1-42}$$
$$\Longrightarrow w = (X R X^T)^{-1} X R Y \tag{1-43}$$

3.3 Proof relation between MLE and weighted linear regression

The probability density funciton is:

$$p(y^{(i)}|x^{(i)}; w) = \frac{1}{\sqrt{2\pi\sigma^{(i)}}} \exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right) \tag{1-44}$$

The likelihood is:

$$L(w) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}; w) \tag{1-45}$$

The log-likelihood is:

$$\ell(w) = \sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}; w) \tag{1-46}$$
$$= \sum_{i=1}^{N} \left(-\log(\sqrt{2\pi\sigma^{(i)}}) - \frac{(y^{(i)} - w^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right) \tag{1-47}$$
$$\tag{1-48}$$

Remove the constants that do not depend on w as they do not affect the maximization, and define $r^{(i)} = \frac{1}{(\sigma^{(i)})^2}$:

$$\ell(w) = -\frac{1}{2} \sum_{i=1}^{N} r^{(i)} (y^{(i)} - w^T x^{(i)})^2 \tag{1-49}$$
$$\tag{1-50}$$

Recall that the weighted linear regression:

$$E_D(w) = \frac{1}{2} \sum_{i=1}^{N} r^{(i)} (y^{(i)} - w^T x^{(i)})^2 \tag{1-51}$$

$$\max_w \ell(w) \Leftrightarrow \min_w -\ell(w) \Leftrightarrow \min_w E_D(w) \tag{1-52}$$

3.4 Code implementation

3.4.1 Code implementation

Submitted to Autograder

3.4.2 Plot

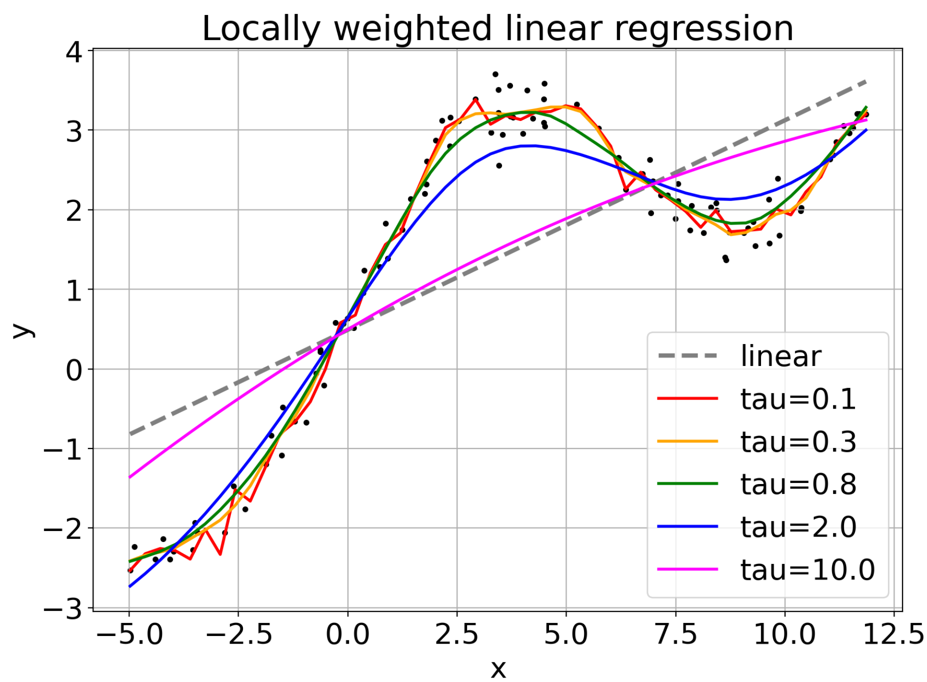


Figure 5: Locally weighted linear regression with different weight parameters

3.4.3 Discussion

When τ is small, the model returns overfitted model which have lots of sharp turns, although the general trend is correct, but it is not robust to data variance.
When τ is large, the model look at too much data once so the result is very close the the non-local-weighted linear regression.

Submitted by Wensong Hu on January 30, 2024.