Regularized Linear Regression

Tuan Nguyen

Ngày 18 tháng 5 năm 2022

Overview

Non-linear data

Dataset split

Overfitting

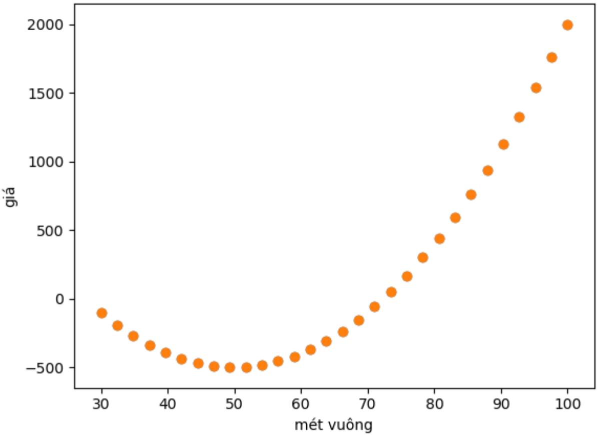
Posterior

Ridge regression

Lasso regression

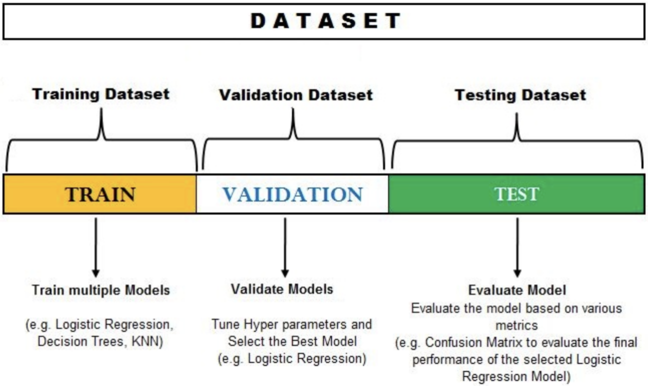
ElasticNet regression

Non-linear data



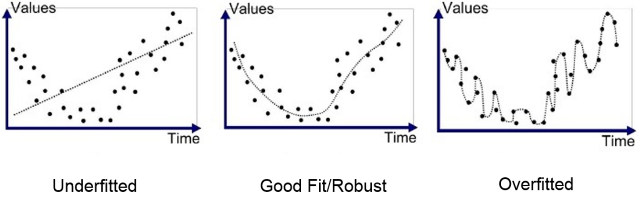
Hình 1: Non-linear data

Dataset split



Hình 2: Train test split

Overfitting



Hình 3: Overfitting and underfitting

Training set error	1%	15%	0.5%
Validation set error	11%	16%	1%

How to solve?

- Underfitting: increase complexity of model
- Overfitting:
  - Add more data
  - Regularization: L1, L2, Dropout,...
  - Early stopping
  - ...

Posterior

Bayes theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$\Leftrightarrow \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$\Rightarrow p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{x}, \mathbf{t}, \alpha, \beta)}$$

$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$ is a posterior. While likelihood is given the parameter how the parameter fit the data, posterior is given the data, what is the probability of parameter. In the posterior, we also includef our belief.

We expect to maximinze the posterior to find **w**.

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Because $p(\mathbf{x}, \mathbf{t}, \alpha, \beta)$ is dependent of **w**

Suppose $p(\mathbf{w}|\alpha)$ is a normal distribution. We have

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}I) = (\frac{\alpha}{2\pi})^{(M+1)/2} \exp\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\}$$

So

$$\begin{aligned}
&p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \\
&\quad \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \\
&\quad \propto \exp\{-\frac{\beta}{2}\sum_{n=1}^{n}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\}
\end{aligned}$$

we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{n}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Posterior (cont.)

or we minimize

$$Q = \|X\mathbf{w} - \mathbf{t}\|_2^2 + \lambda \mathbf{w}^T \mathbf{w}$$

Q is MSE loss with L2 regularization.

By minimizing Q, we can find $\mathbf{w} = (X^T X + \lambda I)^{-1} X^T t$

Gaussian prior is called conjugate prior because the posterior is also Gaussian distribution. So conjugate prior is the distribution that makes the likelihood and posterior have the same distribution.

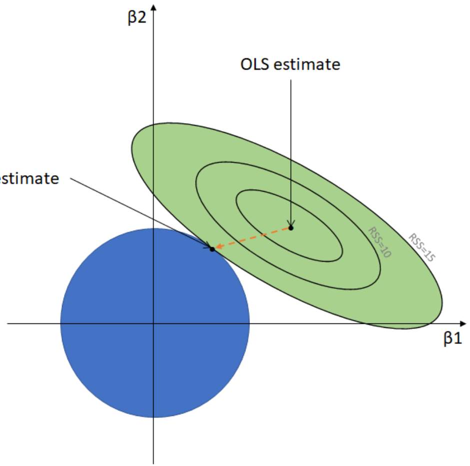
Ridge regression

$$L = \frac{1}{2N}\sum_{i=1}^{N}(w_0 + w_1 x_i - y_i)^2 + \lambda w_1^2$$

Remark:

- Loss function is added with the penalty equivalent to square of the magnitude of the all parameters.
- Ridge regression shrinks the parameters and it helps to reduce the model complexity => avoid overfitting.

Ridge regression (cont.)



Hình 4: Ridge regression

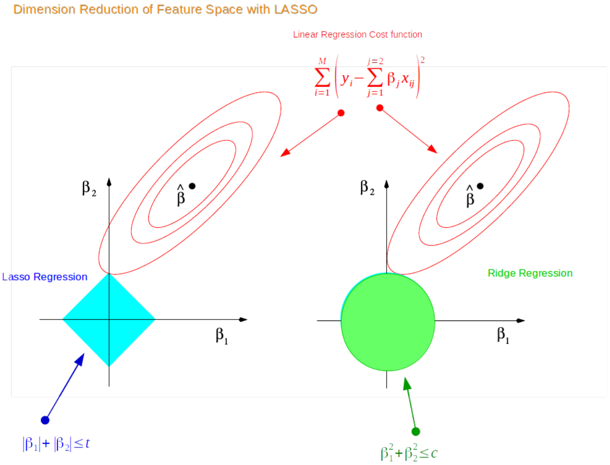
Lasso regression

$$L = \frac{1}{2N}\sum_{i=1}^{N}(w_0 + w_1 x_i - y_i)^2 + \lambda|w_1|$$

Remark:

- Loss function is added with the penalty equivalent to absolute value of the magnitude of the all parameters.
- Lasso regression not only shrinks the parameters and it helps to reduce the model complexity => avoid overfitting but also selects the important feature.

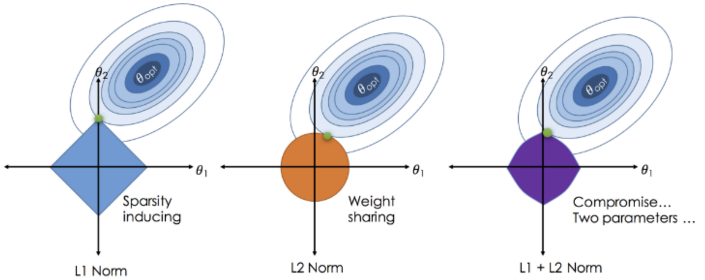
Lasso regression (cont.)



Hình 5: Lasso regression

ElasticNet regression

$$L = \frac{1}{2N}\sum_{i=1}^{N}(w_0 + w_1 x_i - y_i)^2 + \lambda(\frac{1-\alpha}{2}w_1^2 + \alpha|w_1|)$$



Hình 6: ElasticNet regression