Introduction to Intelligent Systems / Lecture 2

Regularized regression

Jae Yun JUN KIM*

July 4, 2019

Recall the unregularized regression case with $h_\theta(x) = \theta_0 + \theta_1 x_1$.

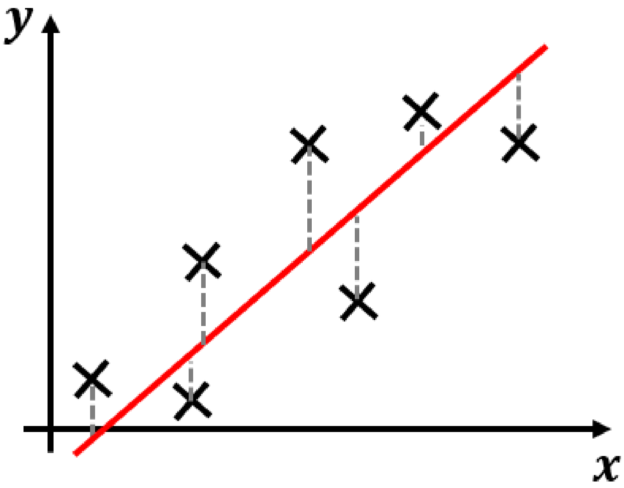


Figure 1: Unregularized regression

An unregularized regression problem can be defined as

$$\min_\theta \quad \frac{1}{2}\sum_{i=1}^{I}\left(h_\theta(x^{(i)} - y^{(i)}\right)^2 \tag{1}$$

For this problem, there exists the closed-form solution

$$\theta = \left(X^T X\right)^{-1} X^T y \tag{2}$$

By the way, the method used to find the optimal parameters θ is known as the **ordinary least squares (OLS)**.
This result gives the optimal parameter values for

- the chosen model ($h_\theta(x)$, a linear model),
- the considered training examples.

But, is this result still good for other samples that are not considered in the training set?

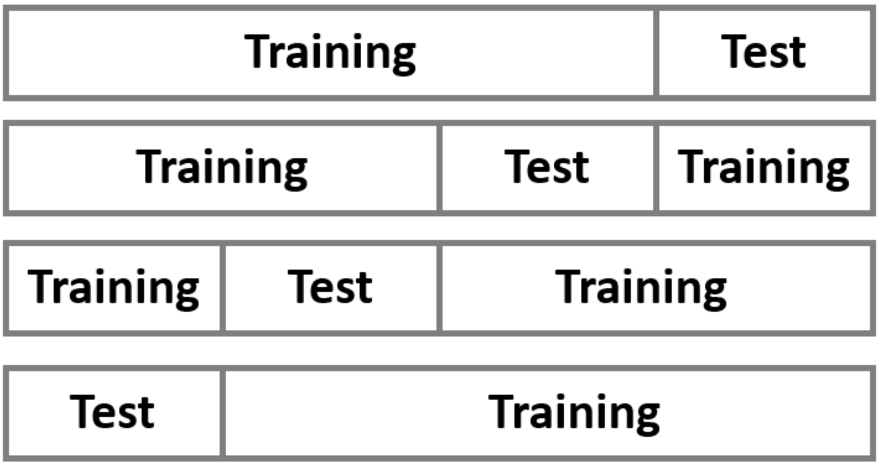


Figure 2: Test and training

To find out this, for a given data set only a portion of data is used for training and the rest for testing the goodness of the found parameters.
So far, we have only considered linear models to fit some given data, but we could also consider other models and choose the one that performs the best for the given data set. But, how to choose such a model? The following graph shows the Mean square errors vs. the complexity of models.

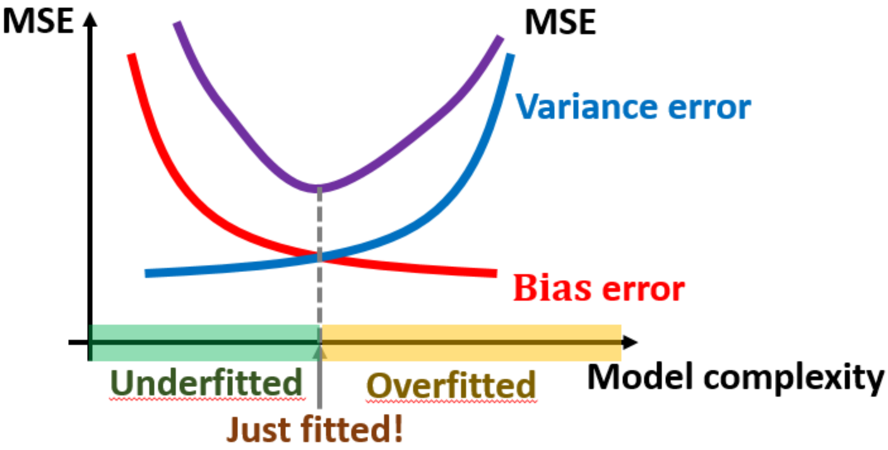


Figure 3: MSE vs. model complexity

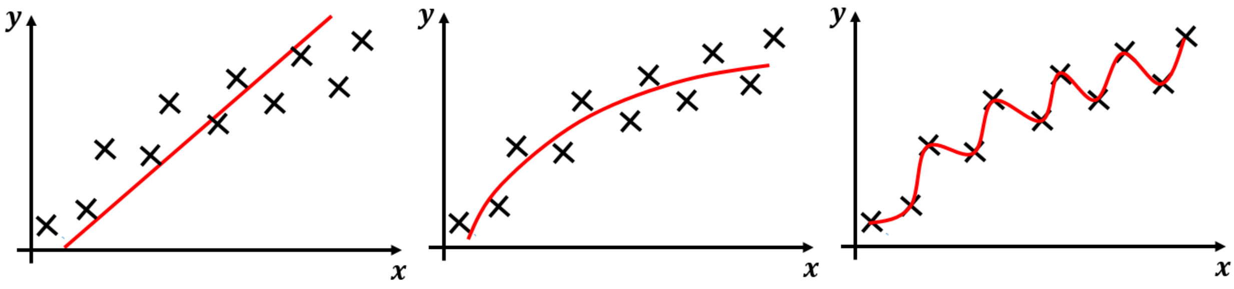
The **mean square error (MSE)** is defined as

$$\frac{1}{I}\sum_{i=1}^{I}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2 \tag{3}$$

The **Bias error** is the error computed with respect to the training data set (i.e., the data used to estimate the model parameters). While the **variance error** is the error computed with respect to the test data set (i.e., the data set used to estimate the model parameters).

Example

[*] ECE Paris Graduate School of Engineering, 37 quai de Grenelle 75015 Paris, France; jae-yun.jun-kim@ece.fr



(a) Linear model (b) Parabolic model (c) High order model

Figure 4: Model comparison

1 Underfitting or high bias

"Biased" means that the chosen algorithm has a strong preconception (or bias) and despite the evidence shown by the data, the algorithm is dominated by this preconception (bias) and therefore the fitting result is not good.

2 Overfitting problem

High variance means that the complexity is high enough that

- high order polynomials can fit the given data well.
- the space of the hypothesis is too large and too variable and we do not have enough data to constrain it to get a good hypothesis function.

If we have too many features, the learned hypothesis may fit the training set very well. That is,

$$E = \frac{1}{2} \sum_{i=1}^{I} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2 \approx 0, \tag{4}$$

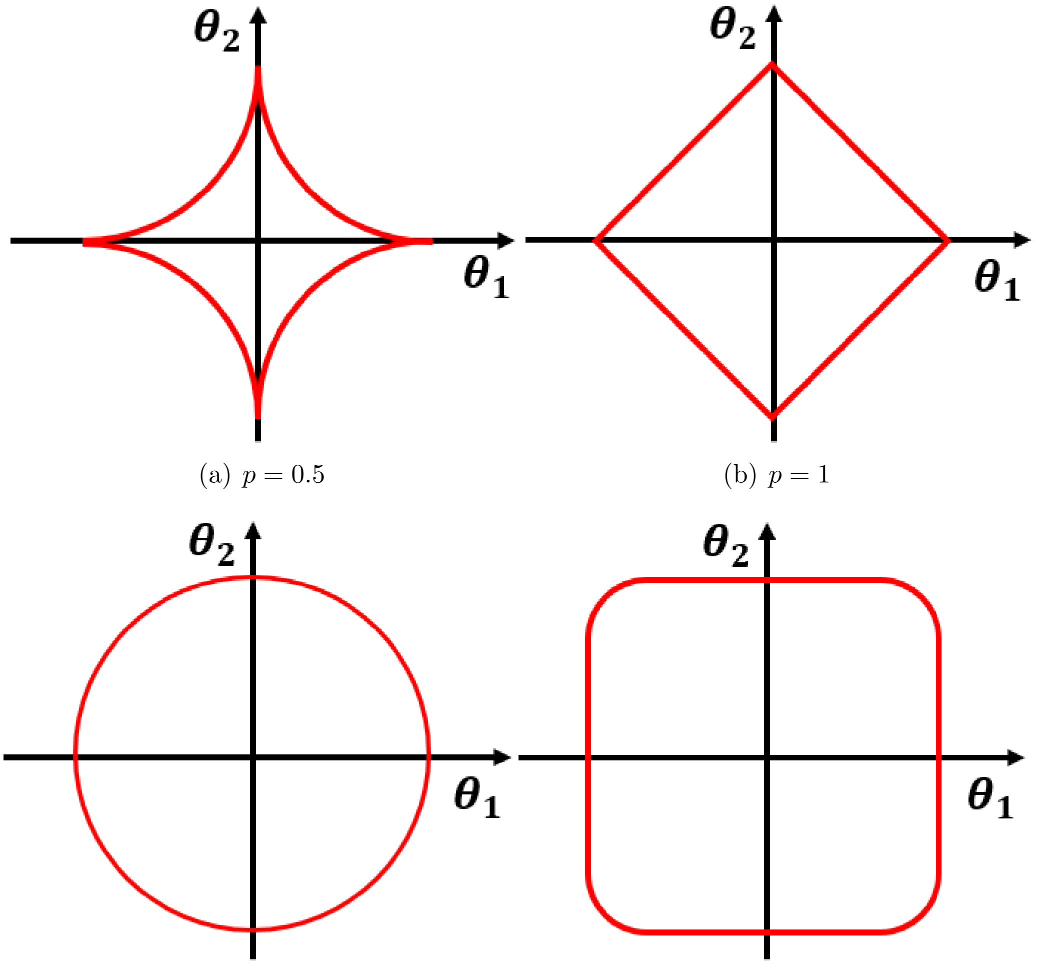
but it fails to generalize to new examples.
The overfitting problem is specially significant when the number of features is large and not enough data is available.
There are two ways to overcome the overfitting problem. The first one is that by using the model selection algorithm one can manually select the features to keep and reduce the number of features. One drawback of this approach is that by throwing away some features, we lose the available information. The second option is by regularizing the cost function. This method allows one to keep all the features, but reduce the magnitude (values) of parameters θ. This can work well even when we have a lot of features, where each of these features contributes a bit to predicting y.

3 Regularization

Some characteristics of regularization are following:

- Small values for parameters
- Simpler hypothesis
- Less prone to overfitting

3.1 Various regularization functions



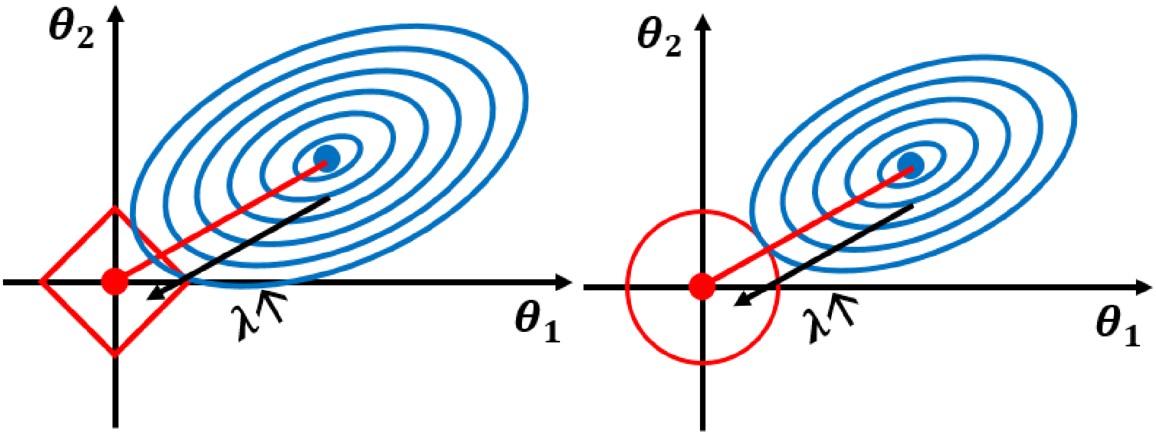
(a) $p = 0.5$

(b) $p = 1$

(c) $p = 2$

(d) $p = 4$

Figure 5: Various regularization functions



(a) LASSO ($p = 1$)

(b) Ridge ($p = 2$)

Figure 6: Ridge vs. LASSO

$$\begin{aligned} \min_{\theta} \quad & \|h_\theta(x) - y\|_2^2 \\ \text{s.t.} \quad & \|\theta\|_p^p \leq c \end{aligned} \tag{5}$$

where

$$\|\theta\|_p = \left(\sum_i |\theta_i|^p \right)^{1/p} \tag{6}$$

and c is a positive real number.
The above optimization problem is constrained, and we can find the expression corresponding to unconstrained one by constructing the corresponding Lagrangian function. That is,

$$\mathcal{L}(\theta) = \frac{1}{2} \sum_{i=1}^{I} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2 + \lambda \left(\|\theta\|_p^p - c \right) \tag{7}$$

From the Lagrangian function, we can formulate the unconstrained version of the above optimization problem as

$$\min_{\theta} \quad \frac{1}{2} \sum_{i=1}^{I} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2 + \lambda \|\theta\|_p^p \tag{8}$$

In particular, for the Ridge problem, we have

$$\min_{\theta} \quad \frac{1}{2} \sum_{i=1}^{I} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2 + \lambda \|\theta\|_2^2 \tag{9}$$

And, for the LASSO problem

$$\min_{\theta} \quad \frac{1}{2} \sum_{i=1}^{I} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2 + \lambda \|\theta\|_1^1 \tag{10}$$

For the Ridge problem, we can have a closed form solution

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \right)^{-1} X^T y \tag{11}$$

Suppose that $N \leq I$, then

$$\theta = \left(X^T X \right)^{-1} X^T y \tag{12}$$

is not invertible. But, the regularization makes that the matrix in question be invertible.