Regression, Regularization

2/17/2020
ECE 4200

Regression

We are given:

$$(\bar{X}_1, y_1), (\bar{X}_2, y_2), \dots, (\bar{X}_n, y_n)$$

Now the y_i are real valued, the $\bar{X}_i$'s are still d dimensional

Example:

$d = 1$, (Area, house prices)

https://www.zillow.com/promo/zillow-prize/

Linear Regression

$d = 1$, suppose we get n examples:

$$(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$$

$$X_i, y_i \in \mathbb{R}$$

Linear Models. Assume the generative process as:

$$y = w^0 + w^1 X$$

Least Squares Regression

$$J(w^0, w^1) = \sum_{1 \le i \le n} (y_i - w^0 - w^1 X_i)^2$$

Optimize:

$$\arg\min_{w^0, w^1} J(w^0, w^1)$$

Exercise: Show that $J(w^0, w^1)$ is convex.

Therefore making gradient equal to zero is enough!

Least Squares Regression

$$\nabla J(w^0, w^1) = 0$$

$$\frac{\partial J(w^0,w^1)}{\partial w^0} = 0, \frac{\partial J(w^0,w^1)}{\partial w^1} = 0$$

Two linear equations in w^0 and w^1.

$$w^1 = \frac{n(\sum_i X_i y_i) - (\sum_i y_i)(\sum_i X_i)}{n \cdot (\sum_i X_i^2) - (\sum_i X_i)^2}, \; w^0 = \frac{(\sum_i y_i) - w^1 (\sum_i X_i)}{n}$$

Least Squares Regression

Consider the high dimensional regression problem, $\bar{X}_i \in R^d$

Model: $(\bar{X}, y)$, where $\bar{X} = \left(\bar{X}^1, \bar{X}^2, \dots, \bar{X}^d\right) \in R^d, y \in R$:

$$y = w^0 + w^1\bar{X}^1 + w^2\bar{X}^2 + \cdots + w^d\bar{X}^d$$

Let $\bar{X}' = (1, \bar{X})$, and $\bar{w}' = \left(w^0, w^1, \dots, w^d\right)$, the model is

$$y = w^0 + w^1\bar{X}^1 + w^2\bar{X}^2 + \cdots + w^d\bar{X}^d = \bar{w}' \cdot \bar{X}'$$

Least Squares Regression

$$J(\overline{w}') = \sum_{1 \le i \le n} (y_i - \overline{w}' \cdot \bar{X}')^2$$

Let $Y = [y_1, \ldots, y_n]^T$, and $\overline{w}' = [w^0, \ldots, w^d]^T$

Let X be the $n \times (d+1)$ matrix whose ith row is

$\bar{X}_i{}' = (1, \bar{X}_i), = \left(1, \bar{X}_i^1, \bar{X}_i^2, \ldots, \bar{X}_i^d\right).$

Then,

$$J(\overline{w}') = \| Y - X \cdot \overline{w}' \|_2^2$$

Least Squares Regression

$$J(\overline{w}') = \| Y - X \cdot \overline{w}' \|_2^2$$

$$= (Y - X \cdot \overline{w}')^T \cdot (Y - X \cdot \overline{w}')$$

$$= Y^T Y - 2 \cdot Y^T X \cdot \overline{w}' + (\overline{w}')^T X^T X \overline{w}'$$

Taking the gradient with respect to $\overline{w}'$,
$\nabla J(\overline{w}') = -2X^T Y + 2\, X^T X \overline{w}' = 0$

giving

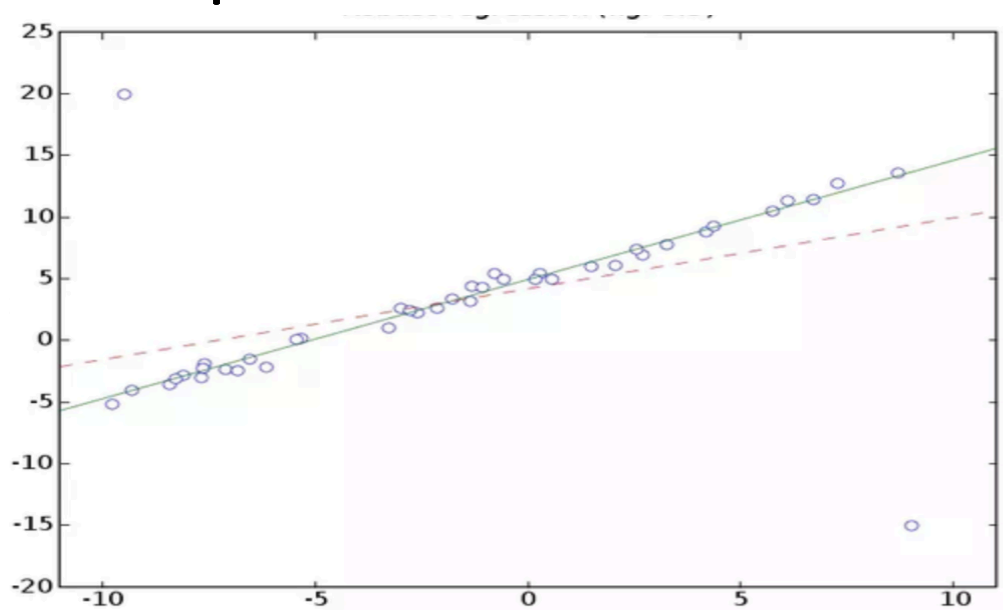
$$\overline{w}' = (X^T X)^{-1} \cdot\ X^T Y$$

MLE interpretation

Maximum Likelihood with Gaussian Noise

Outliers

Suppose one of the labels gets corrupted.



What happens?

One point can affect a lot.

Why?

Squared loss. One term can dominate

Overfitting

Overfitting usually means LARGE coefficients.

Penalty for weights:

$$J(\overline{w}') = \| Y - X \cdot \overline{w}' \|_2^2 + \frac{\lambda}{2} \| \overline{w}' \|_2^2$$

Prove:

$$\overline{w}' = (X^T X + \lambda I)^{-1} \cdot \; X^T Y$$

This is called as Ridge Regression.

Proof.

$$J(\overline{w}') = Y^T Y - 2 \cdot Y^T X \cdot \overline{w}' + (\overline{w}')^T X^T X \overline{w}' + \frac{\lambda}{2} (\overline{w}')^T (\overline{w}')$$

$$\nabla J(\overline{w}') = -2X^T Y + 2\, X^T X \overline{w}' + \lambda \overline{w}' = 0$$

Taking $\overline{w}'$ *to one side, we obtain*

$$\overline{w}' = (X^T X + \lambda I)^{-1} \cdot\ X^T Y$$

Ridge Regression.

Ridge Regression

Does it actually regularize? How do we know that it regularizes.

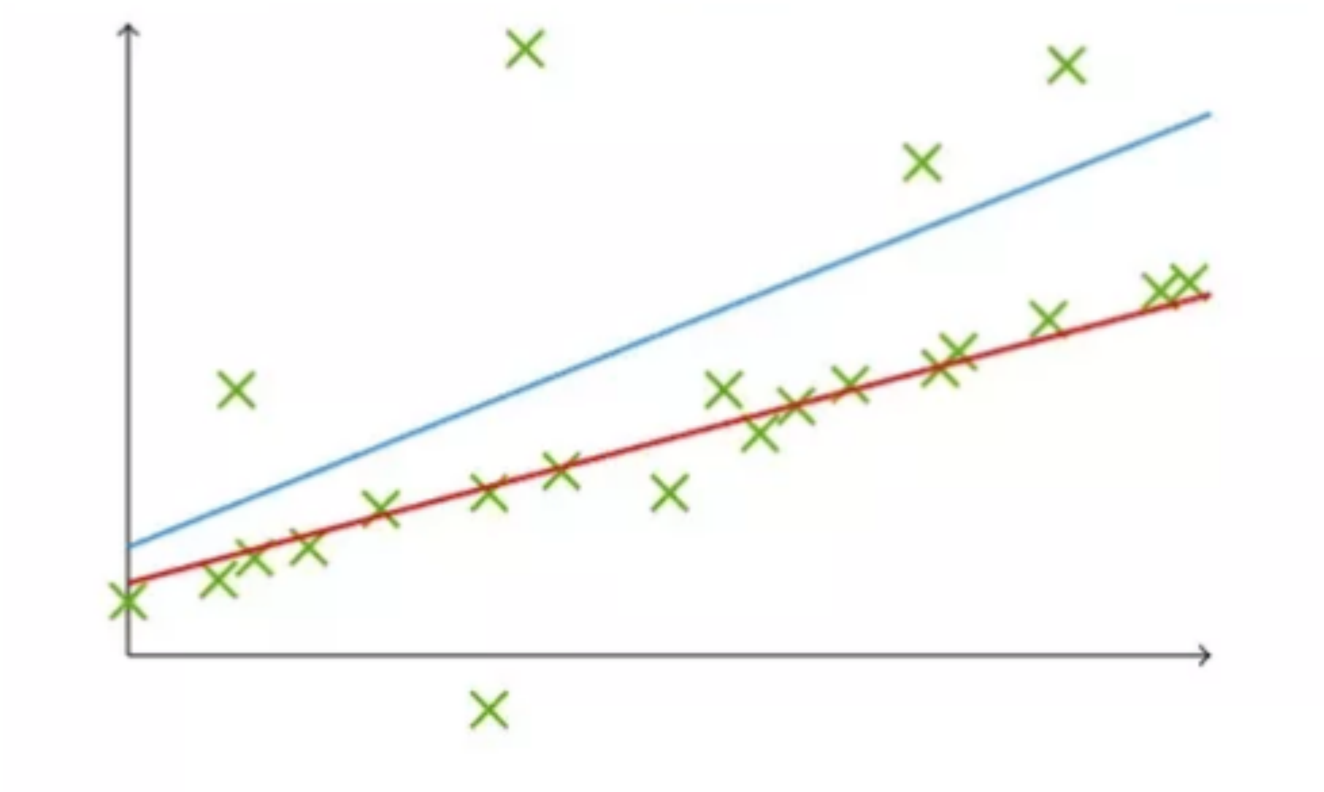
- Does increasing λ reduce the norm of w?
- Does the training error always grow with λ?

Yes, and Yes, which you will show in the next assignment!

LASSO

$$J(\overline{w}') = \| Y - X \cdot \overline{w}' \|_2^2 + \frac{\lambda}{2} \| \overline{w}' \|_1$$

This is called as LASSO Regression.

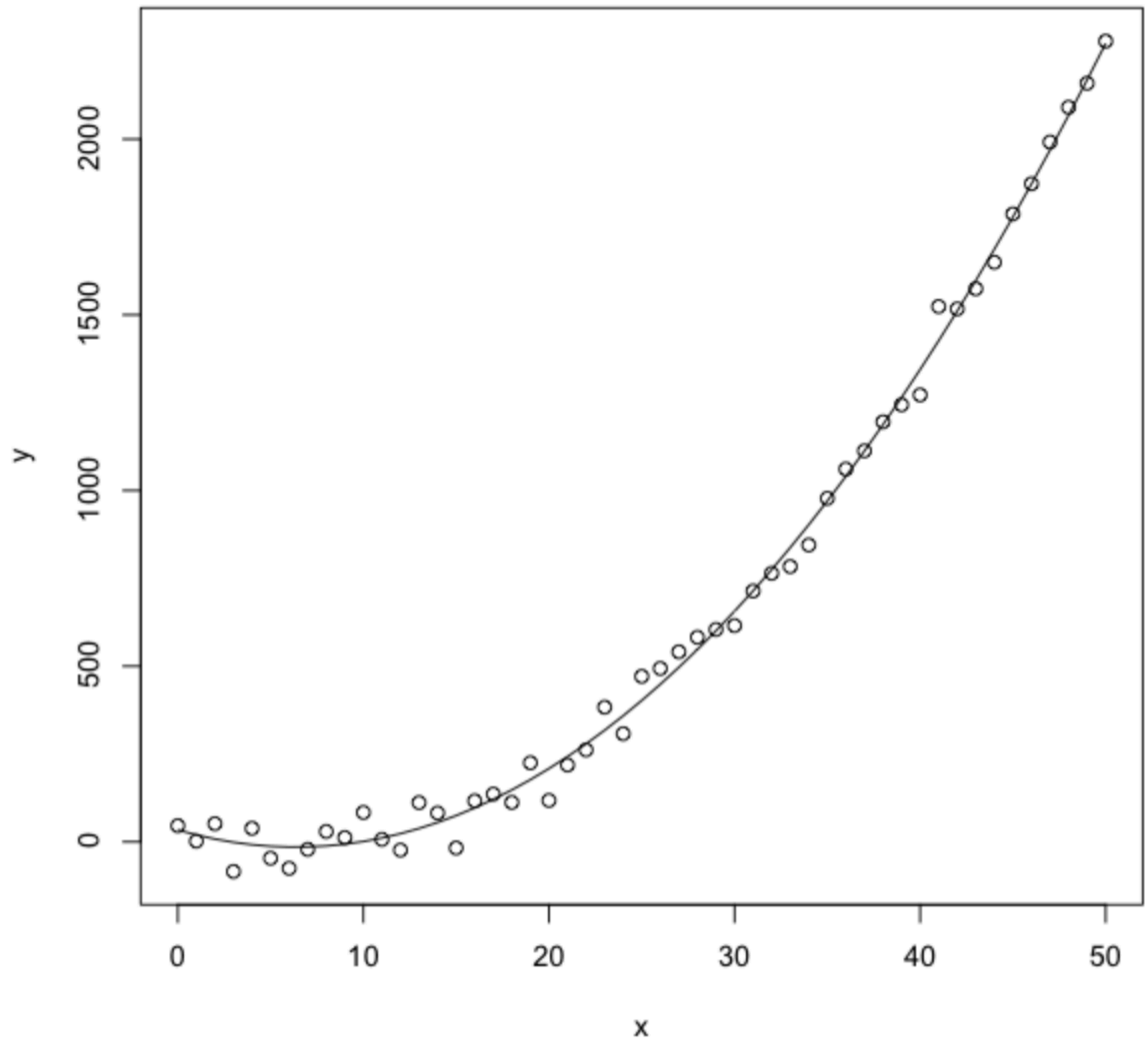


Further Readings

http://eniac.cs.qc.cuny.edu/andrew/gcml/lecture5.pdf

Polynomial Regression

Linear Regression may not be enough



Polynomial regression for $d = 1$

$d = 1$, suppose we get n examples:

$$(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$$

$$X_i, y_i \in \mathbb{R}$$

Linear Models. Assume the generative process as:

$$y = w^0 + w^1 X$$

Polynomial regression for $d = 1$

$d = 1$, suppose we get n examples:

$$(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$$

$$X_i, y_i \in \mathbb{R}$$

Polynomial models of degree p, generative process as:

$$y = w^0 + w^1 X + \cdots + w^p X^p$$

Polynomial regression for $d = 1$

$d = 1$, and degree $p = 2$

$$y = w^0 + w^1 x + w^2 x^2$$

This fits a parabola to the data

Polynomial regression for $d = 1$

$d = 1$, and degree $p = 2$

How to do polynomial regression on

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)?$$

We will use linear regression to do polynomial regression.

Polynomial regression for $d = 1$

We will use linear regression to do polynomial regression.

Map the feature into higher dimension. When $p = 2$:

$$(X_i, y_i) \to \left((X_i, X_i^2), y_1\right)$$

Let $\bar{X}_i = \left(X_i, X_i^2\right)$, then do linear regression for the two dimensional features!

$$(\bar{X}_1, y_1), (\bar{X}_2, y_2), \ldots, (\bar{X}_n, y_n)$$

Polynomial regression for $d = 1$

We will use linear regression to do polynomial regression.

Map the feature into higher dimension. General p

$$(X_i, y_i) \to \left((X_i, X_i^2, \dots, X_i^p), y_1\right)$$

Let $\bar{X}_i = \left(X_i, X_i^2, \dots, X_i^p\right)$, then do linear regression for the p-dimensional features!

Summary

One dimensional degree p polynomial regression was reduced to a linear regression with p features!!