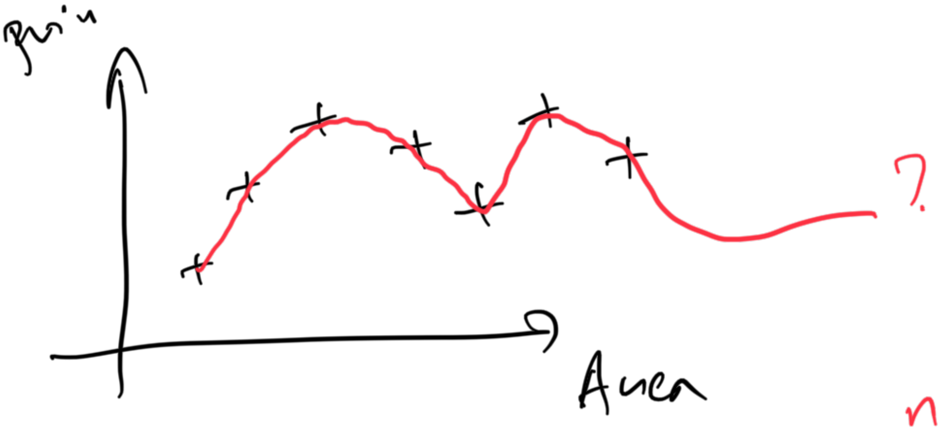
Regularization

Overfitting.

* Optimize the features ✓
* Data augmentation
* Remove Outliers ✓
* Simpler models.
* Regularization.



$$\min_{\theta} \underline{\underline{\mathcal{L}(\theta)}} = \frac{1}{n} \sum_{i=1}^{n} (\underline{y_i} - \underline{\hat{y}_i})^2$$

$\hat{y}_i = f(\theta)$

$$\boxed{\min \left[\mathcal{L}(\theta) + \lambda R(\theta) \right]}$$

θ L

$$\begin{cases} \hat{y}_1 = 1000\, x_1^{(i)} + 100\, x_2^{(i)} + 10 x_3 \\ \hat{y}_2 = 5\, x_1^{(i)} + 2 x_2 + 10 x_3 \end{cases}$$

$$\theta_1 = \begin{bmatrix} 1000 \\ 100 \\ 10 \end{bmatrix} \qquad \theta_2 = \begin{bmatrix} 5 \\ 2 \\ 10 \end{bmatrix}$$

$$\hat{y}_1 = X\theta_1$$

we want $\|\theta\|$ to be small

$$\|\theta\|_2 = \sqrt{\sum_{i=1}^{3} \theta_i^2}$$

$$\|\theta\|_1 = \sum_{i=1}^{3} |\theta_i|$$

$$\min_{\theta} \mathcal{L}(\theta) + \lambda R(\theta)$$

with $R(\theta) = \|\theta\|$

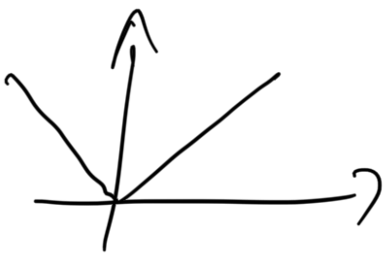
if $\boxed{R(\theta) = \|\theta\|_1 = \sum_{i=1}^{m} |\theta_i| \quad \text{LASSO}}$

$\boxed{\text{if } R(\theta) = \|\theta\|_2 = \sqrt{\sum_{j=1}^{m} \theta_i^2} \quad \text{Ridge}}$

$$\begin{cases} \theta^* = (X^T X)^{-1} X^T y & \lambda = 0 \\ \min_{\theta} \mathcal{L}(\theta) \end{cases}$$

$$\min_{\theta} \underbrace{\mathcal{L}(\theta) + \lambda R(\theta)}_{H(\theta)}$$

$$\frac{\partial H}{\partial \theta} = 0$$



Ridge

$$\hat{\theta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$