

The background of the slide is a light gray surface with a complex network of thin, dark gray lines. These lines connect various colored dots (red, green, blue, black) scattered across the frame. The dots are of different sizes and are distributed in a way that suggests a complex, interconnected system or a data network. The overall aesthetic is technical and abstract.

Gradient descent with Regularization



$$\min_{\vec{w}, b} J(\vec{w}, b) = \min_{\vec{w}, b} \left\{ \frac{1}{2m} \sum_{i=1}^m [f_{\vec{w}, b}(\vec{x}^{(i)}) - y^{(i)}]^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2 \right\}$$

Gradient descent

repeat {

$$w_j = w_j - \alpha \frac{\partial J(\vec{w}, b)}{\partial w_j}$$

$$\frac{1}{m} \sum_{i=1}^m (f_{\vec{w}, b}(\vec{x}^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} w_j$$

$$b = b - \alpha \frac{\partial J(\vec{w}, b)}{\partial b}$$

$$\frac{1}{m} \sum_{i=1}^m [f_{\vec{w}, b}(\vec{x}^{(i)}) - y^{(i)}]$$



$$w_j = \underbrace{\left(1 - \frac{\alpha d}{m}\right) w_j}_{\text{Usual update}} - \underbrace{\frac{\alpha}{m} \sum_{i=1}^m \left[f_{\vec{w}, b}(\vec{x}^{(i)}) - y^{(i)} \right] x_j^{(i)}}_{\text{Usual update}}$$

Usual update

$$\alpha = 0.01$$

$$d = 1$$

$$m = 100$$

$$0.9999$$



$$\frac{\partial}{\partial w_j} J(\vec{w}, b) = \frac{\partial}{\partial w_j} \left[\frac{1}{2m} \sum_{i=1}^m [f(\vec{x}^{(i)}) - y^{(i)}]^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2 \right]$$

$$f(\vec{x}^{(i)}) = \vec{w} \cdot \vec{x}^{(i)} + b$$

$$= \frac{1}{2m} \sum_{i=1}^m 2 \cdot [\vec{w} \cdot \vec{x}^{(i)} + b - y^{(i)}] \cdot \vec{x}_j^{(i)} + \frac{\partial \lambda}{2m} w_j$$

$$= \frac{1}{m} \left[\sum_{i=1}^m [f(\vec{x}^{(i)}) - y^{(i)}] \cdot \vec{x}_j^{(i)} + \lambda w_j \right]$$



Types of Regularization

- Ridge Regression (L2 Regularization)

Effect: Shrinks coefficients smoothly, making the model more robust to noise.

Use Case: When all features are relevant, but we want to prevent large coefficients.

$$\text{Loss} = \sum (y - \hat{y})^2 + \lambda \sum w^2$$

- Lasso Regression (L1 Regularization)

Effect: Shrinks some coefficients to exactly zero, making it useful for feature selection.

Use Case: When we suspect that some features are unnecessary.

$$\text{Loss} = \sum (y - \hat{y})^2 + \lambda \sum |w|$$

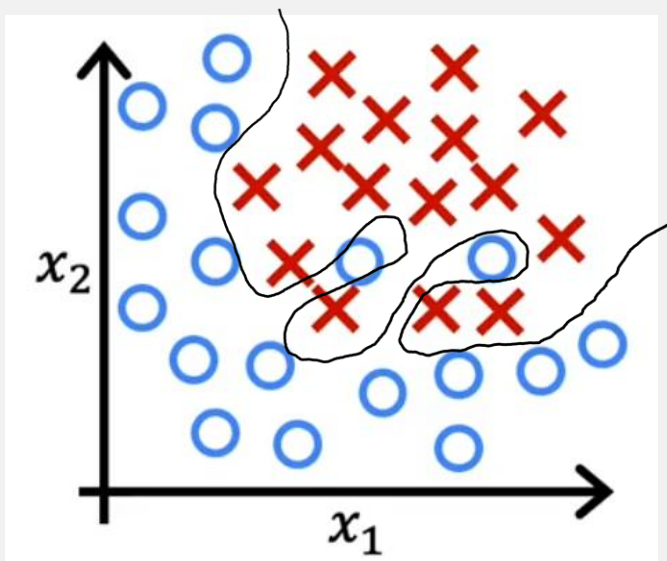


Dataset

x	y
1	2
2	4
3	6
4	8
5	10
6	13
7	15
8	20
9	25
10	30



Logistic Regression



$$J(\vec{w}, b) = -\frac{1}{m} \sum_{i=1}^m \left[y^{(i)} \log(f_{\vec{w}, b}(\vec{x}^{(i)})) + (1 - y^{(i)}) \log(1 - f_{\vec{w}, b}(\vec{x}^{(i)})) \right] \\ + \frac{1}{2m} \sum_{j=1}^n w_j^2$$



Gradient Descent L.R.

$$w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(\vec{w}, b)$$

$$\frac{1}{n} \sum_{i=1}^n \left[\left[f_{\vec{w}, b}(x^{(i)}) - y^{(i)} \right] x_j^{(i)} \right] + \frac{\lambda}{n} w_j$$

$$b = b - \alpha \frac{\partial}{\partial b} J(\vec{w}, b)$$

$$\frac{1}{n} \sum_{i=1}^n \left[f_{\vec{w}, b}(\vec{x}^{(i)}) - y^{(i)} \right]$$

