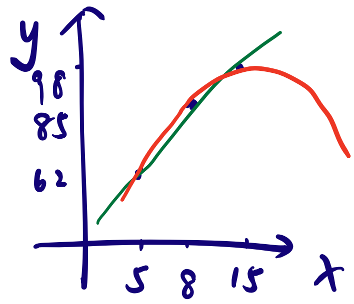
Sep 17, 2024

$X = [15, 5, 8]$

$y = [98, 62, 85]$



loss function $J(w) = \sum_{i=1}^{N} (y_i - \underline{w^T x_i})^2$ $\quad x_i = \begin{bmatrix} 1 \\ 15 \end{bmatrix}_{(d+1)\times 1}$

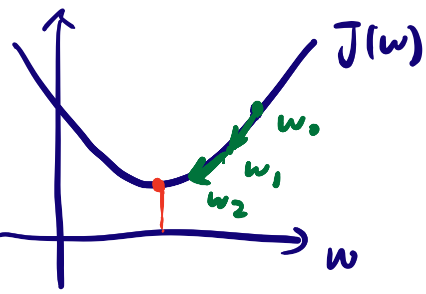
scalar $\underline{J(w)} = (y - X^T w)^T (y - X^T w)^T$

$X = \begin{bmatrix} 1 & 15 \\ 1 & 5 \\ 1 & 8 \end{bmatrix}_{n\times(d+1)}$ $\quad w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}_{(d+1)\times 1}$ $\quad y = \begin{bmatrix} 98 \\ 62 \\ 85 \end{bmatrix}_{n\times 1}$

bias

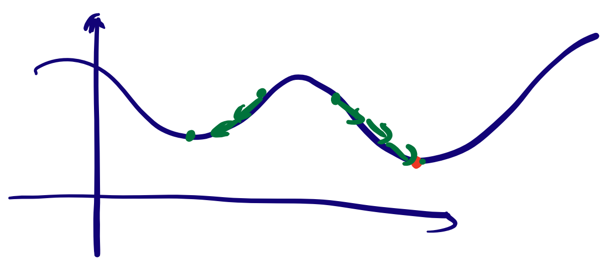
$$\hat{w} = \underbrace{(X^T X)}_{(d+1)\times(d+1)}^{-1} X^T y$$

Gradient Descent



w_0 random initialization

$$w_{i+1} = w_i - \eta \frac{\partial J(w_i)}{\partial w_i}$$



$\phi(x) = [1, x, x^2]$

$[1, x, x^2 \dots x^d]$

$$X = \begin{bmatrix} 1 & 15 & 225 \\ 1 & 5 & 25 \\ 1 & 8 & 64 \end{bmatrix}_{N \times (d+1)} \qquad w \qquad y$$

Ridge Regression

$$J(w) = \frac{1}{2}(y - Xw)^T(y - Xw) + \frac{1}{2}\lambda w^T w$$

L^2 regularization

prevent overfitting

reduces impact of correlated inputs

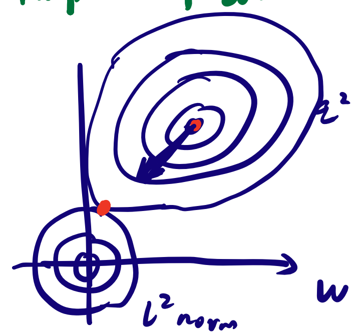
$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

adding a prior to w

$y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$

$w \sim \mathcal{N}(0, \tau^2)$

$$\lambda = \frac{\sigma^2}{\tau^2}$$

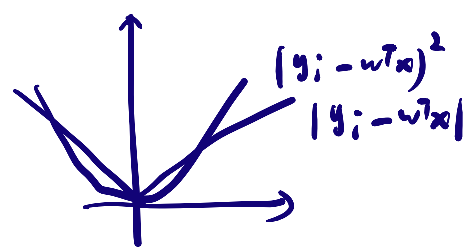


Robust Regression

$$J(w) = \sum_{i=1}^{N} |y_i - w^T x_i|$$

prevent outliers

use gradient descent



Laplace $\quad y_i \sim \text{laplace}(w^T x_i, b)$

LASSO

$$J(w) = \sum_{i=1}^{N} \frac{1}{2}(y_i - w^T x_i)^2 + \lambda \|w\|_1$$

L_1 norm regularization

prevent overfitting

encourage sparse w

help feature selection

$y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$

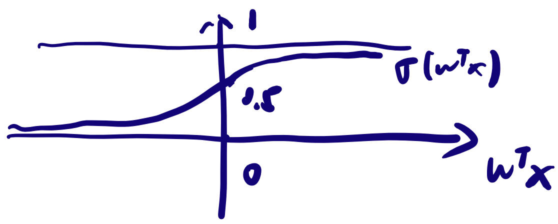
$w \sim \text{Laplace}(0, b)$

Logistic Regression classifier

training examples $(x_i, y_i)_{i=1}^{D}$, learn w

$P(y|x) \sim \text{Ber}(\theta)$

$$\theta = \text{sigmoid function} = \sigma(w^T x) = \frac{1}{1+\exp(-w^T x)}$$



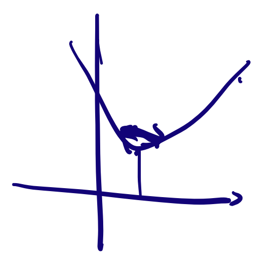
likelihood $P(D|\theta) = \prod_{i=1}^{N} \theta_i^{y_i} (1-\theta_i)^{1-y_i}$

$$NLL = -\log P(D|\theta)$$

$$= \sum_{i=1}^{N} -y_i \log \theta_i - (1-y_i)\log(1-\theta_i)$$

Cross entropy loss

log loss



Gradient Descent

initialize w_0

$$w_{i+1} = w_i - \eta \frac{dNLL}{dw}$$

$$= w_i - \eta \sum_{i=1}^{N} (\theta_i - y_i) x_i$$

ADAM

Mulitple classes

$p(y|x) \sim \text{multinoulli}(\theta_k)$

$$\theta_k = \frac{\exp(w_k^T x)}{\sum_{k=1}^{C} \exp(w_k^T x)}$$ C classes

Softmax function

$C = 2$

$$\theta_1 = \frac{\exp(w_1^T x)}{\exp(w_1^T x) + \exp(w_2^T x)}$$

$$= \frac{1}{1 + \exp((w_2 - w_1)^T x)}$$

$$\hat{w} = -(w_2 - w_1)$$

$$= \frac{1}{1 + \exp(-\hat{w}^T x)}$$