Sep 12, 2024

Robust Regression

$y \sim \text{laplace}(w^T x, b)$

$y = w^T x + \varepsilon, \quad \varepsilon \sim \text{laplace}(0, b)$

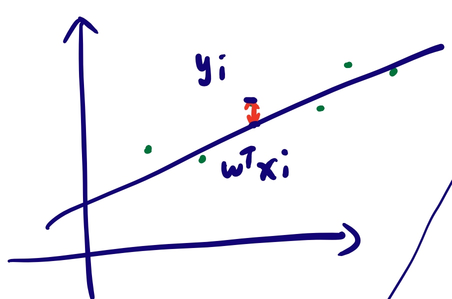
MLE. probabilistic interpretation

$$LL(w) = \log \prod_{i=1}^{N} P(y_i \mid w, b)$$

$$= \log \prod_{i=1}^{N} \frac{1}{2b} \exp\left(-\frac{|y_i - w^T x_i|}{b}\right)$$

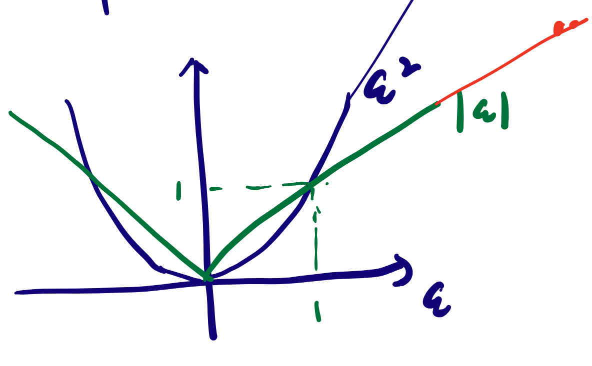
$$= \frac{N}{2b} - \underline{\frac{1}{b} \sum_{i=1}^{N} |y_i - w^T x_i|}$$

Geometric interpretation

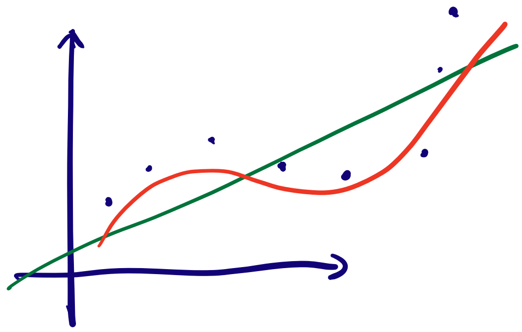


$$J(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

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



outlier, large ε



$y = w^T x$

basis function $\phi(x) = [1, x, x^2 \dots x^d]$

$y = w_0 + w_1 x + \underbrace{w_2}_{\approx 0} x^2 + \cdots \underbrace{w_d}_{\approx 0} \cdot x^d$

Linear to w, non-linear to x

$\phi(x) = [1, x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_1 x_2, \cdots]$

Ridge Regression

$\theta(w) = J(w) + \underline{\lambda \|w\|_2^2}$

l_2 norm regularization
prevent overfitting

$\|w\|_2^2 = w_1^2 + w_2^2 + \cdots + w_d^2$

$\theta(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \underline{\lambda \|w\|_2^2}$

Setting $\frac{\partial \theta(w)}{\partial w} = 0$

$\theta(w) = (y - Xw)^T (y - Xw) + \lambda w^T w$

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

Correlated variables

$X = [x_1, x_2] \quad x_2 = x_1 + \epsilon$

$y = w_0 + w_1 x_1 + w_2 x_2$

$y = w_0 + 2 w_1 x_1 + 0 w_2 x_2$

$y = w_0 + 1.5 w_1 x_1 + 0.5 w_2 x_2$

$\vdots \qquad \vdots$

adding L_2 norm regularization

$\min \|w\|_2^2 \Rightarrow y = w_0 + w_1 x_1 + w_2 x_2$

LASSO

l_2 norm $\|w\|_2 = (w_1^2 + \cdots w_d^2)^{1/2}$

$\|w\|_2^2 = w_1^2 + \cdots + w_d^2$

L_1 norm $\|w\|_1 = |w_1| + |w_2| + \cdots + |w_d|$

l_∞ norm ..

l_0 norm ..

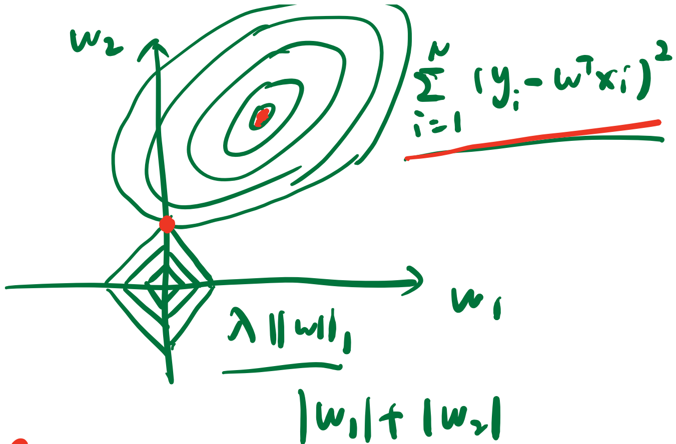
$J(w) = \sum_{i=1}^{N} (y - w^T x_i)^2 + \lambda \|w\|_1$

l_1 norm regularization

prevent overfitting

leads sparsity in w

many w will 0



$\begin{cases} w_1 = 0 \\ w_2 \neq 0 \end{cases}$

LASSO

$\begin{cases} w_1 \neq 0 \\ w_2 \neq 0 \end{cases}$

Ridge

True Bayesian

prior w

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

$$\begin{bmatrix} \tau^2 & & 0 \\ & \ddots & \\ 0 & & \tau^2 \end{bmatrix}$$

$$P(w|D) \propto \prod_{i=1}^{N} \mathcal{N}(y_i | w^T x_i, \sigma^2) \cdot p(w)$$

$$\hat{w}_{MAP} = \arg\max_{w} \sum \left(-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2 - \frac{1}{2\tau^2} w^T w \right)$$

$$= \arg\min_{w} \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \frac{\sigma^2}{\tau^2} w^T w$$

Ridge Regression l_2 norm

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

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