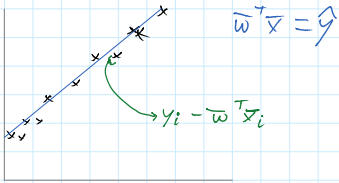


Linear Regression



$$\bar{w}^T \bar{x} = \bar{y}$$

MLE

$$\underset{\bar{w} \in \mathbb{R}^d}{\operatorname{argmin}} - \sum_{i=1}^N \log p(y_i | \bar{x}_i, \bar{w}) \Rightarrow \text{negative log likelihood}$$

$\bar{w}$ is parameter, d by 1

$$\begin{aligned} \text{NLL}(\bar{w}) &= \frac{1}{2} (Y - X\bar{w})^T (Y - X\bar{w}) + c \\ &= \frac{1}{2} (Y^T - \bar{w}^T X^T) (Y - X\bar{w}) \\ &= \frac{1}{2} [Y^T Y - Y^T X \bar{w} - \bar{w}^T X^T Y + \bar{w}^T X^T X \bar{w}] \end{aligned}$$

$$\nabla \text{NLL}(\bar{w}) = 0 - \cancel{X^T} Y + X^T X \bar{w}$$

scalar, so $[a]^T = [a]$ because a is 1 by 1
 $\bar{w}^T X^T Y = Y^T X \bar{w}$

$$X^T X \bar{w} = X^T Y$$

d by d matrix d by 1 vector

How to solve $\underbrace{X^T X}_{A} \bar{w} = \underbrace{X^T Y}_{b}$?

$$A \bar{w} = b$$

option: $A^{-1} *$

"Gauss Jordan"

$$\bar{w} = (X^T X)^{-1} X^T Y$$

$O(d^3)$

* $X^T X \succ 0$
columns of X are independent

option: SVD $O(d^3)$ okay as inverse doesn't have to exist

$d = \text{number of columns (features)}$

option: QR $\bar{w} = (X^T X)^{-1} (X^T Y)$ $O(d^3)$

option: choleski: * $X^T X \succ 0$ positive eigenvalues $\checkmark \Rightarrow O(d^3)$

LU?

$$X \bar{v}$$

option: conjugate gradient * $O(Nd)$

$\checkmark \Rightarrow$

$$\bar{x}_i^T \bar{v}$$

option: SAG

$O(d)$

$\checkmark \checkmark$

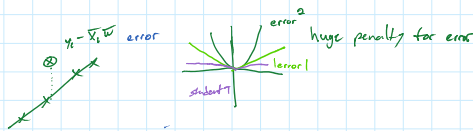
Problems with least squares

- "non robust" (outliers)
- linear assumption
- non-uniqueness

L_1 Regression

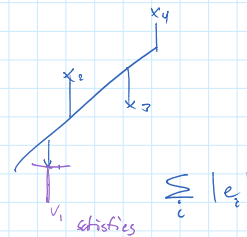
$$\underset{\bar{w}}{\operatorname{argmin}} \sum_{i=1}^N |y_i - \bar{x}_i^T \bar{w}|$$

abs value



from laplace distribution

at the origin, not differentiable



$$\underset{\bar{w}, \bar{v}}{\operatorname{argmin}} \sum_{i=1}^N v_i$$

$$\bar{w}^T x = x^T \bar{w}$$

subject to $v_i \geq |y_i - \bar{w}^T \bar{x}_i|$ bigger than our error (at minimizer its equal)

$$\downarrow$$

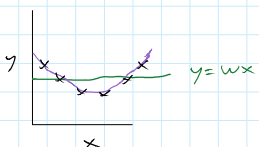
$$-v_i \leq y_i - \bar{w}^T \bar{x}_i \leq v_i$$

linear programming

linear constraints, to avoid NP hard problem

Basis Expansion

(linear parameterization) + (non-linear transformation of variables (features $\bar{x}_i$))
 $\rightarrow$ non-linear regression



$$y = w^T [1, x, x^2]$$

$$y = w_1 + w_2 x + w_3 x^2$$

polynomial

Bases:
- general polynomials: $w^T (1, x, x^2, \dots, x^m)$

- RBE

$$x_i \mapsto \exp\left(\frac{-(x_i - x_j)^2}{2\sigma^2}\right)$$

- measure distance to data point, use those as features
- go through every data point if error = 0 (universal approximator)

$-\sin(x)$ periodic 
 $[x, \sin(x)]$ periodic, (in/de)creasing 

Model Selection

KNN: parameter k

Naïve bayes: "add one" α

Polynomial basis: parameter 'm' (degree of polynomial)

Which features $\bar{x}_i^j$ to include?

- cross validation

Akaike information criterion } put penalty on "degrees of freedom"
Bayesian information criterion } (-not recommended)

for $k \in \{1, 2, 3 \dots K\}$

partition into folds

$$X = \begin{bmatrix} X_{\text{train}} \\ X_{\text{validation}} \end{bmatrix}$$

- train with K on X_{train}
- test on $X_{\text{validation}}$

2 fold

$$X = \begin{bmatrix} X_{\text{validation}} \\ X_{\text{train}} \end{bmatrix} \quad \left. \begin{array}{l} \text{ie. } N/2 \\ \\ \end{array} \right\} \quad \left. \begin{array}{l} \\ \\ N/2 \end{array} \right\}$$

- train $X_1 \rightarrow X_q$, validate on X_{10}

- combinatorial, "leave one out"

10 fold

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{10} \end{bmatrix}$$

Ridge Regression

$$\operatorname{argmin}_{\bar{w} \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^N (y_i - \bar{w}^T \bar{x}_i)^2$$

many solutions, pick one.

L2. Make solution unique.

$$+ \frac{\lambda}{2} \sum_{j=1}^d w_j^2 \quad \} = w^T w = \|w\|^2 \quad (\text{euclidean norm squared})$$

why? (the norm of)

- encourage w to be smaller
- smoother (more regular)