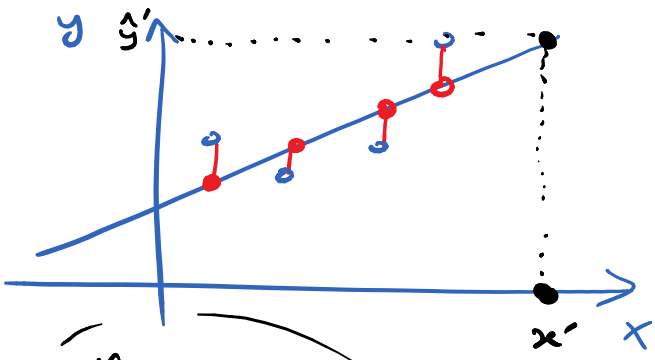
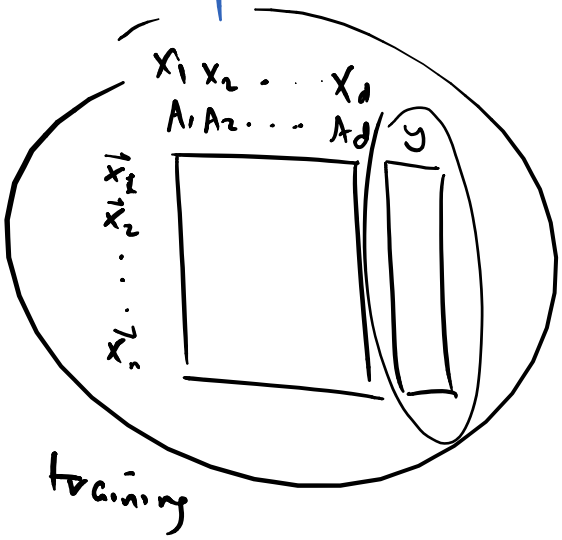
Linear Regression



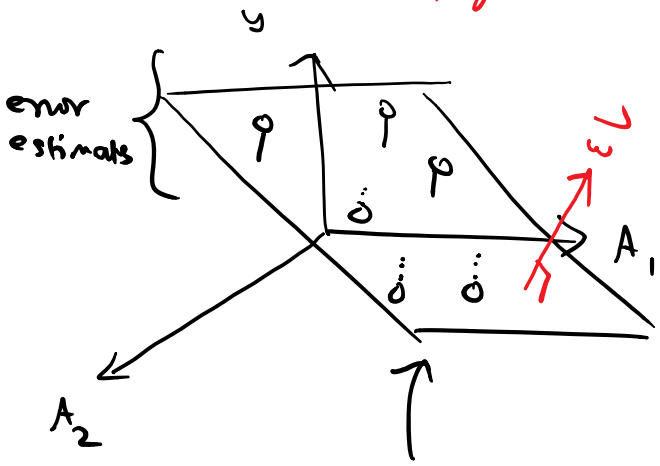
dependent Var. — Independent Var

$\hat{y} = wx + b$

slope — equation of a line — bias / offset



$\vec{x}_i \in \mathbb{R}^d$



error : $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$\hat{y}_i = \sum_{j=1}^{d} w_i x_i + w_0$$

bias $w_0 = b$

$$\vec{x}_i = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_0 = b \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \Bigg\} \mathbb{R}^{d+1} \text{ space}$$

$$\hat{y}_i = \vec{w}^T \vec{x}$$

$$E = \sum_{i=1}^{n} (y_i - \vec{w}^T \vec{x}_i)^2$$

← find the best $\vec{w}$ to minimize the squared errors

n diff components

n diff Components

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{pmatrix} \qquad \vec{\omega}$$

$$\epsilon = \|\vec{y} - X\vec{\omega}\|^2$$

$$\epsilon = \|\vec{y} - \hat{\vec{y}}\|^2$$

$$\hat{\vec{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

$$\epsilon = (\vec{y} - X\vec{\omega})^T (\vec{y} - X\vec{\omega})$$

$$\frac{\partial \epsilon}{\partial \omega} \Rightarrow (X^T X)\omega = X^T y$$

$$\omega = (X^T X)^{-1} X^T y$$

$d+1$ dimensions

$X \in R^{n \times d}$

$$X = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_n \end{pmatrix}$$

$d \times d$ $\quad d \times n$ $\quad n \times 1$

$d \times 1$

$$\hat{y} = X\omega$$

$$= X\left((X^T X)^{-1} X^T y\right)$$

$$\hat{y} = \left(X (X^T X)^{-1} X^T\right) y$$

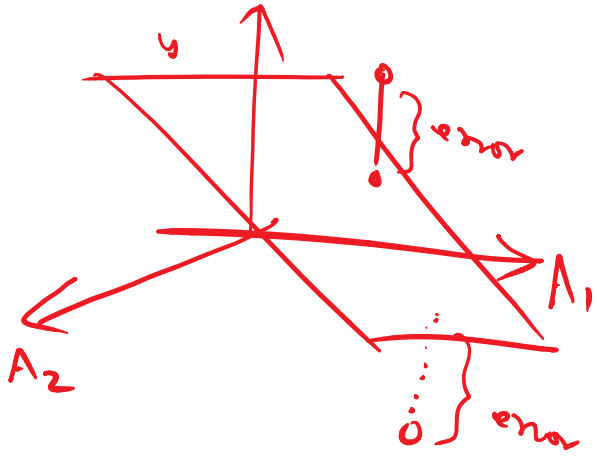
Predicted values

Hat Matrix

Projection matrix

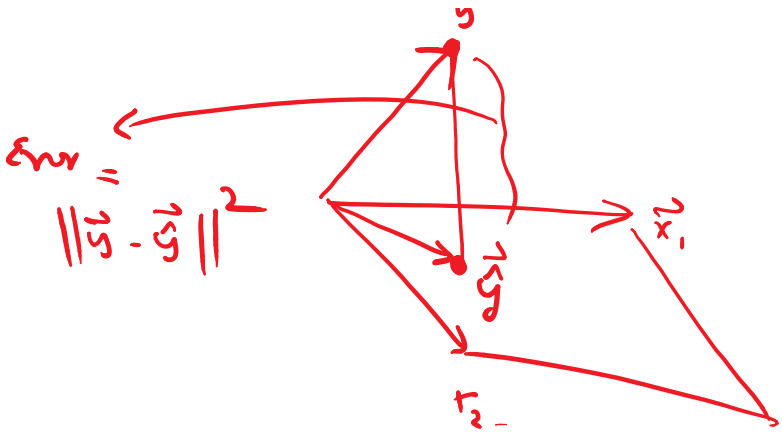
H

true value

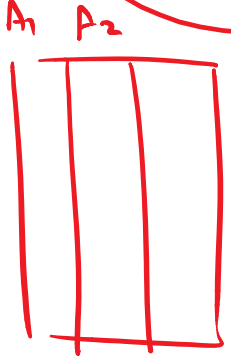


$d+1$ dimensions

H



$y \in \mathbb{R}^n$



$A_1 \in \mathbb{R}^n$

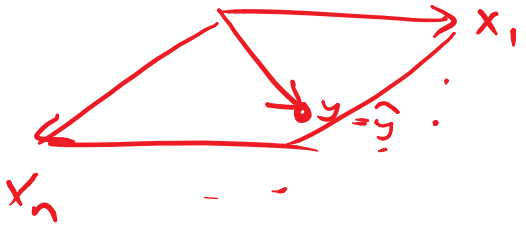
$A_2 \in \mathbb{R}^n$

$\vdots$

$$y = Xw$$

y could be expressed as a linear combination of attributes

$$y = A_1 w_1 + A_2 w_2 \cdots + A_d w_d$$



$\hat{y} = Xw \neq y$

$\vec{y}$ has to be projected onto the space spanned by the column of X (attributes)

$\hat{y}$ is the orthogonal projection of $\vec{y}$ onto $(1, A_1, A_2 \cdots, A_d)$

the projected point $\hat{y}$ is the best we can do

bias w_0 w_1 w_2 $\cdots$ w_d

$$\hat{y} = Xw$$

$$\sum_{i=1}^{n} x_i^T w$$

$$\sum_{i=1}^{n} \left(\sum_{j=0}^{d} x_{ij} w_j \right)$$

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{a} x_{ij}\,\omega_j\right)$$

weights

Regularized linear regression

ω_i should not be too large

$$\sum_{i=1}^{d} \omega_i^2 = \omega^T\omega$$

Regularized objective

$$\min_{\vec{\omega}} E = \|\vec{y} - \hat{\vec{y}}\|^2 + C\,\omega^T\omega \qquad C \geq 0$$

error — size/norm of weight vector

Small errors with small weights

$$E = (\vec{y} - X\vec{\omega})^T(\vec{y} - X\vec{\omega}) + \omega^T(C \cdot I)\omega$$

original — identity matrix

$$= \vec{y}^T\vec{y} - 2(X\vec{\omega})^T\vec{y} + \vec{\omega}^T(X^TX)\omega + \omega^T(CI)\omega$$

$$E = \vec{y}^T\vec{y} - 2(X\vec{\omega})^T\vec{y} + \omega^T(X^TX + CI)\omega$$

$$\frac{\partial E}{\partial \vec{\omega}} = 0 \Rightarrow 2\vec{\omega}(X^TX + cI) = 2XX^Ty$$

ridge regression

$$\vec{\omega} = \left(X^T X + cI\right)^{-1} X^T y$$

$X^T X$

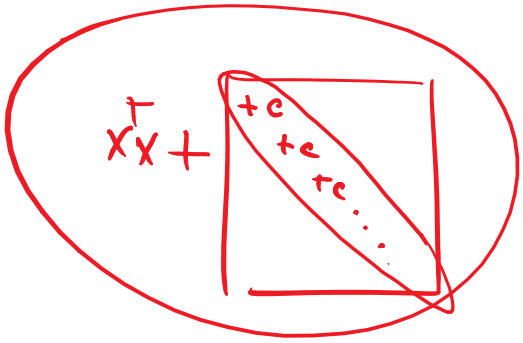
Covariance

$d \times d$

Symmetric

PSD

$\lambda_i \geq 0$



Inverse always exists

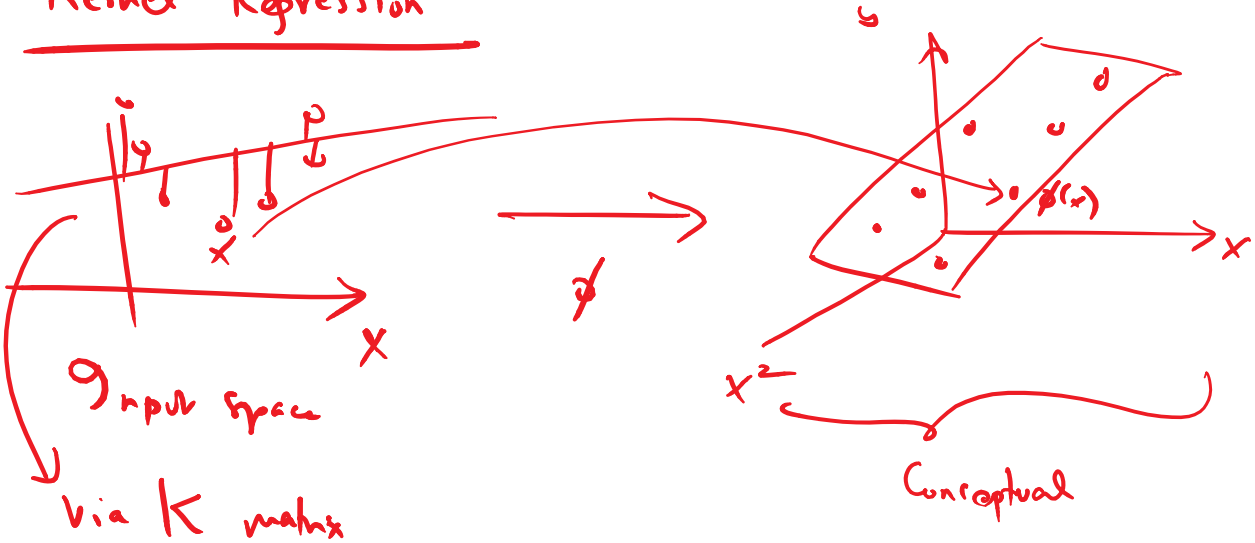
if λ_i is eigenvalue of $X^T X$

$\lambda_i + c$ is eigenvalue of $X^T X + cI$

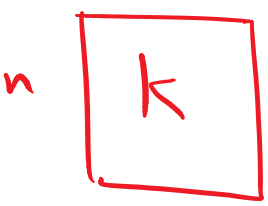
$$\left(X^T X + cI\right)$$

$\lambda_i' \geq c$

Kernel Regression



via K matrix



Kernel Ridge Regression

cI

$$\vec{\omega} = \sum \alpha_i \phi(x_i)$$

$$\vec{\omega} = \sum \alpha_i \phi(x_i)$$

linear case / kernel

$$\phi(x_i) = x_i$$

we cannot reconstruct the weight vector

$$\vec{\omega} = \sum \alpha_i x_i$$

$$\vec{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$$

$$\hat{y}_j = \phi(x_j)\,\vec{\omega}$$

$$= \sum_{i=1}^{n} \alpha_i \phi(x_j)^T \phi(x_i)$$

$$= \sum_{i=1}^{n} \alpha_i K(j,i)$$

$$\hat{y}_j = \vec{\alpha} \cdot K_j$$

$\vec{\alpha}$: alpha vector; K_j: j-th row of K

---

linear kernel: $\phi(\vec{x}_i) = \vec{x}_i$

ridge objective $E = \|y - X\omega\|^2 + c\,\omega^T\omega$

$$\frac{\partial E}{\partial \omega} = 0 \Rightarrow (X^TX + cI)\,\vec{\omega} = X^Ty$$

$$c\vec{\omega} = X^Ty - X^TX\omega$$

$$\vec{\omega} = X^T\left(\frac{y - X\vec{\omega}}{c}\right)$$

$\vec{\alpha}$ (1 c)

$\vec{w} = X^T \vec{\alpha} = \sum_{i=1}^{n} x_i \alpha_i$

$\vec{\alpha} = \frac{\vec{y} - X\vec{w}}{c} \Rightarrow c\vec{\alpha} = y - X\vec{w}$

$\Rightarrow c\vec{\alpha} = y - (XX^T)\alpha$

$\Rightarrow c\vec{\alpha} + (XX^T)\alpha = y$

$c I \vec{\alpha} + (XX^T)\alpha = y$

$X \in \mathbb{R}^{n \times d}$

$XX^T = K$

n.d.d.n

$n \times n$

linear kernel

$K(i,j) = x_i^T x_j$

$\phi(x_i)^T \phi(x_j)$

$(XX^T + cI)\vec{\alpha} = \vec{y}$

$\vec{\alpha} = (XX^T + cI)^{-1} \vec{y}$

$\vec{\alpha} = (K + cI)^{-1} \vec{y}$

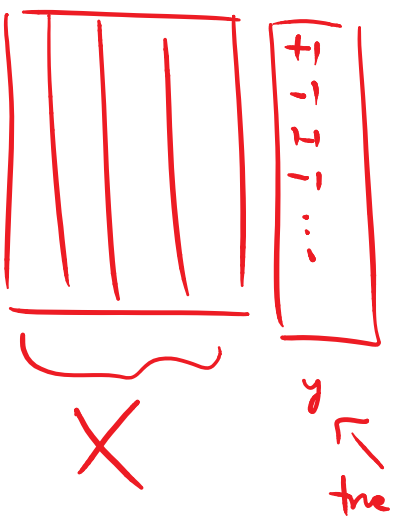
kernel ridge

$\vec{w} = \sum_{i=1}^{n} \alpha_i \phi(x_i)$

$d \times d \rightarrow n \times n$ space

$\hat{y} = \vec{w}^T \phi(z) = \sum_{i=1}^{n} \alpha_i k(i,z)$ for test case z.

Logistic Regression

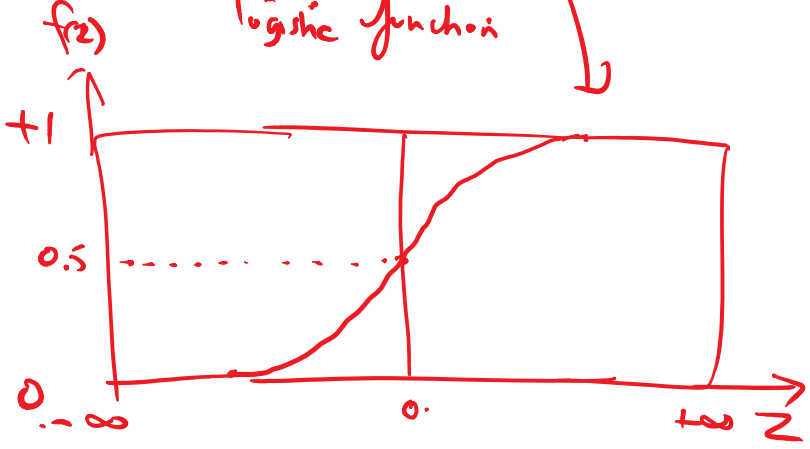


$$z = \sum_{i=1}^{d} w_i x_i + w_0$$

squash the value z in the range [0, 1]

$$f(z) = \frac{1}{1+e^{-z}}$$

logistic function



find the best w with logistic function

y classes : $\{+1, -1\}$

$$+1 \begin{cases} \hat{y} = P(y=1|x) = f(z) \\ -1 \quad 1-\hat{y} = 1 - P(y=1|x) = 1 - f(z) \end{cases}$$

$$f(-z) = 1 - f(z)$$ ← property of logistic function

$$1 - f(z)$$

$$1 - \frac{1}{1+e^{-z}} = \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{e^{-z}/e^{-z}}{(1+e^{-z})/e^{-z}} = \frac{1}{e^{z}+1}$$

$$f(y \cdot z)$$

$$f(-z) = \frac{1}{1+e^{-(-z)}} = \frac{1}{1+e^{z}}$$ ///

$+1 \equiv f(z)$

$-1 \equiv f(-z) = 1 - f(z)$

likelihood function for all data points

$$L = \prod_{i=1}^{n} f(y_i z) = \prod_{i=1}^{n} f(y_i w^T x_i)$$

maximize $\ln L = \sum_{i=1}^{n} \ln f(y_i w^T x_i)$

finding the parameter $\vec{w}$ so that we maximize the log of the likelihood (MLE)

how probable is X, y given the parameters $\vec{w}$

likelihood

$$w^T = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix}$$

$$L = \sum_{i=1}^{n} \ln f(y_i w^T x_i)$$

$$f(z_i) = \frac{1}{1+e^{-z_i}}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial f(z_i)} \cdot \frac{\partial f(z_i)}{\partial z_i} \cdot \frac{\partial z_i}{\partial w}$$

$$= \frac{1}{f(z_i)} \cdot [f(z_i)(1-f(z_i))] \cdot y_i x_i$$

$$\nabla_w = f(-z_i) y_i x_i$$

where $z_i = y_i w^T x_i$

Gradient descent optimization for $\vec{\omega}$, per component $\vec{\omega}_j$

$\vec{\omega} = (0 \ldots 0)$

repeat

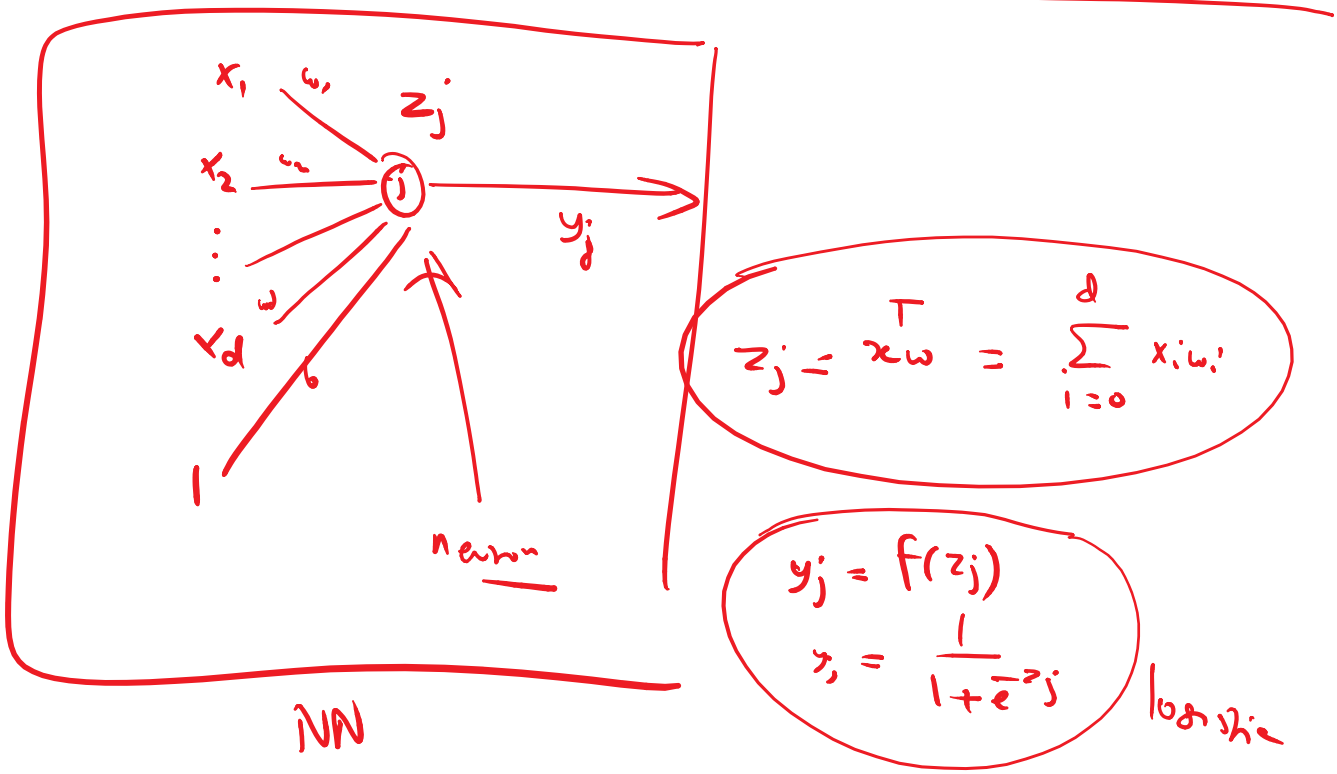
  For all points i in random order

    $z_i = y_i \omega^T x_i$

    $\nabla_i = f(-z_i) \cdot y_i x_i$

    $\omega = \omega + \eta \nabla_i$

until $\|\omega - \omega_{prev}\|^2 \leq \varepsilon$

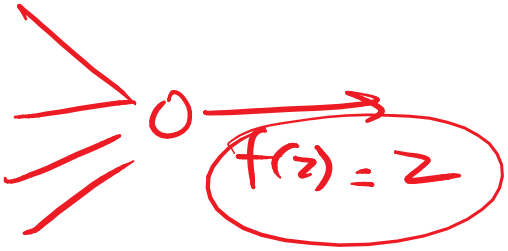


$$z_j = x\omega^T = \sum_{i=0}^{d} x_i \omega_i$$

$y_j = f(z_j)$

$y_j = \frac{1}{1+e^{-z_j}}$ logistic

NN



$f(z) = z$ linear