Linear Regression

$x \longrightarrow y$

x is a vector

$x \in \mathbb{R}^d \rightarrow$ x is a vector of length d

y is a scalar

$y \in \mathbb{R}$

Prediction or Regression

Predict future income

Current GPA, # AI courses taken

$y =$ 70000

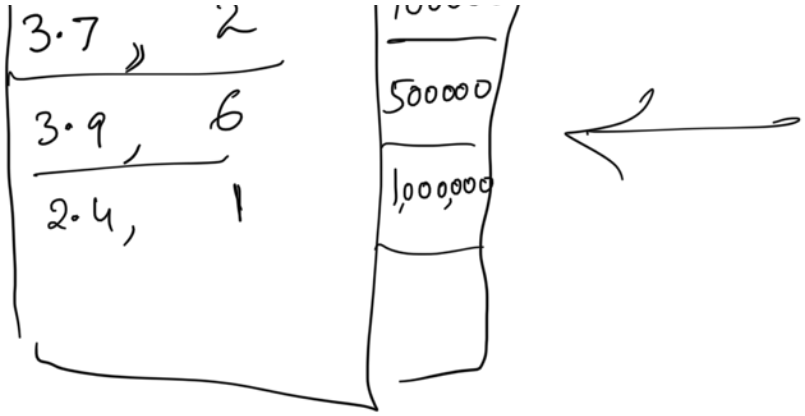
$y =$ 300000

3.8	4

y ?

Training data

GPA	# AI	Income
		100000



3.5, 2	?

Functional models.

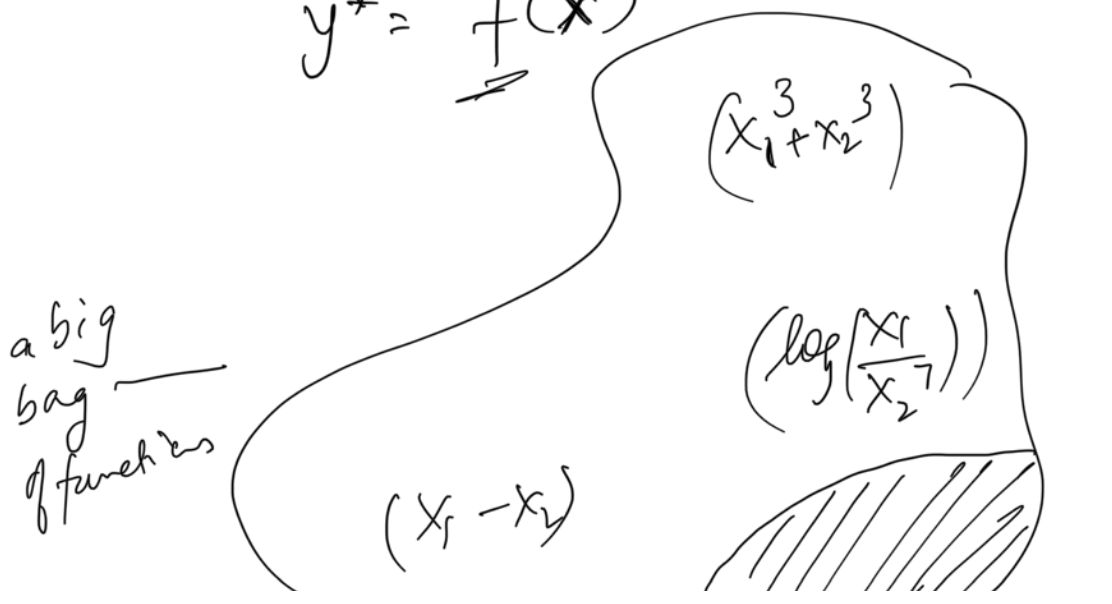
$$y = f(x)$$

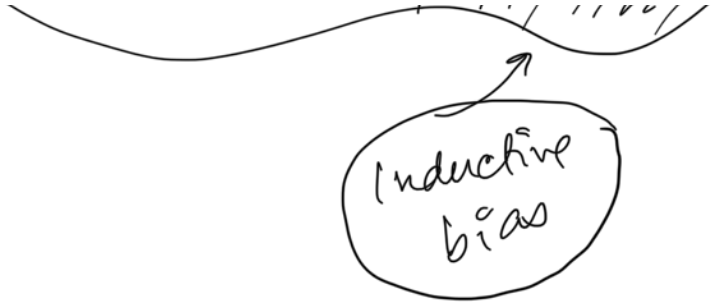
↑ some function.

If we learn $f()$,

then for a new x^* $x = [x_1 \; x_2]$

$$y^* = f(x^*)$$





Monday Feb 8

$X \rightarrow Y$

Functional models	Probabilistic Models.
$y = f(x)$	$p(x, y)$
	$p(y \mid x)$ ← Bayes Rule

X_1

$X \rightarrow$ [x_1 | x_2 | | | | | | x_D]

x, y, z

x_1 x_2

[x_{11} x_{12} ---]

$\overset{D}{\leq}$

$$x \cdot y = \sum_{i=1} x_i y_i$$

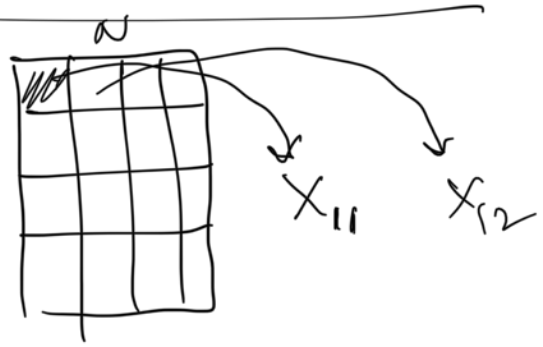
$$= x_1 y_1 + x_2 y_2 + \cdots + x_D y_D$$

$$|x| = \sum_{i=1}^{D} |x_i|$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{D} x_i^2} \quad l_2 \text{ norm}$$

$$\|x\|_p = \left(\sum_{i=1}^{D} x_i^p \right)^{1/p}$$

Matrix



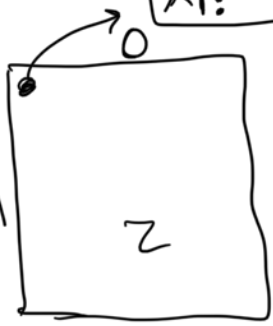
$$X \in R^{M \times N}$$

$$Y = X^T \quad (N \times M)$$

$$cX = \begin{bmatrix} cX_{11} & cX_{12} & \cdots \\ \vdots & \ddots & \vdots \end{bmatrix}$$

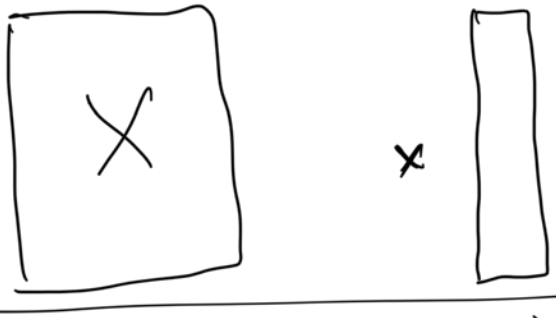
XY N O

$X_{1:} \cdot Y_{:1}$



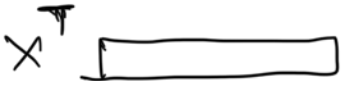
$Z = XY$

$Z_{ij} = X_{i:} \cdot Y_{:j}$



A D-length vector is a (D×1) matrix

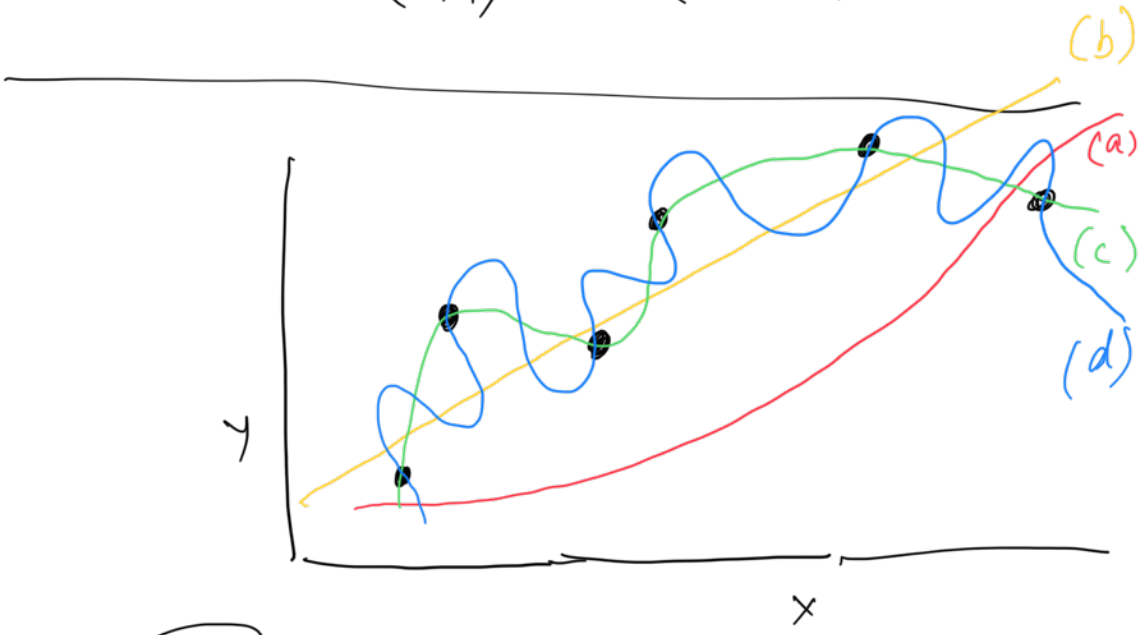
$$x \cdot y = x^T y$$



$1 \times D$ $\quad$ y

$D \times 1$

$z = x^T y = (1 \times 1) = (x^T) \cdot y$



error

(b) → $y = mx + c$

(c) → $y = ax^3 + bx^2 + cx + d$

(d) → $y = ax^{10} + bx^9 + \cdots + \cdots$

Linear Regression

$y = f(x)$

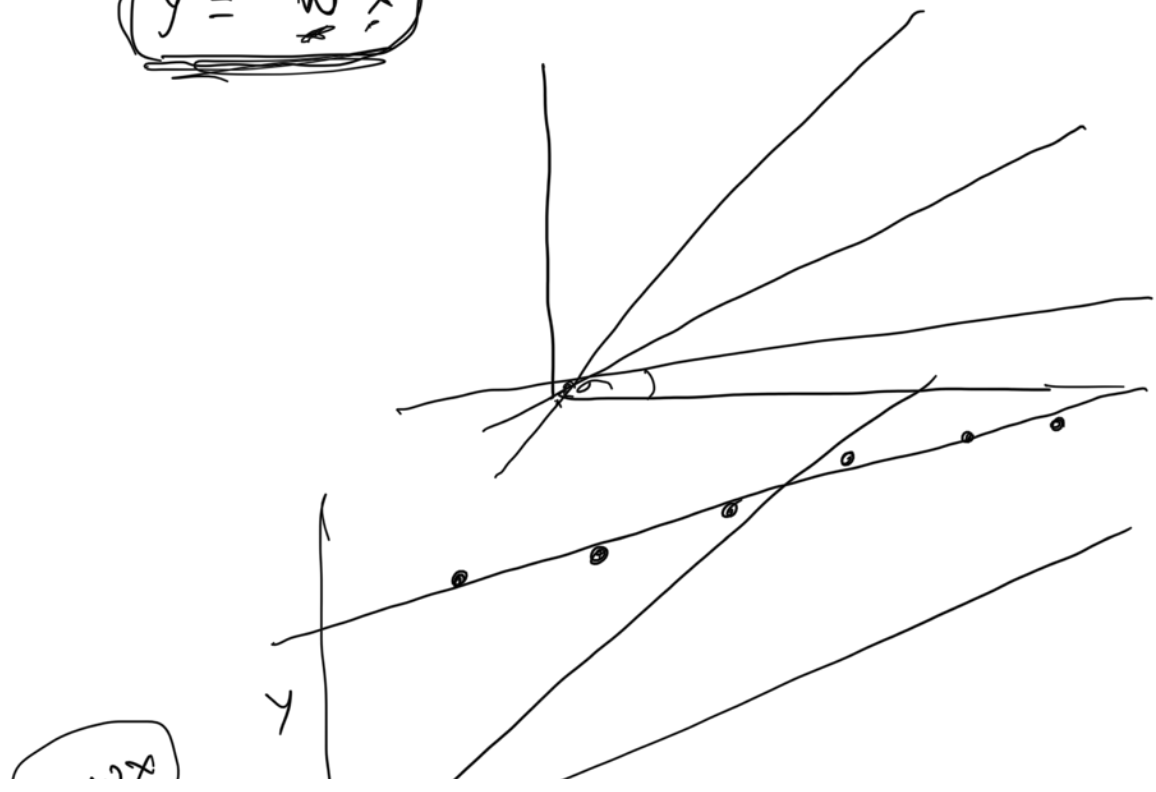
$y = w^T x \rightarrow D \times 1$

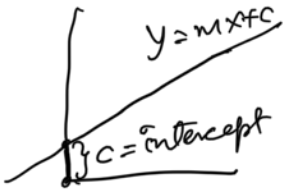
weight vector $(D \times 1)$

$= w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$

For the flu example

$y = w x$





$$y = w_0 + w_1 x$$

w_0: bias-term

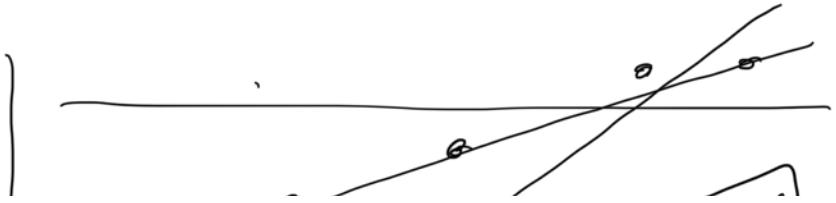
Given some data:

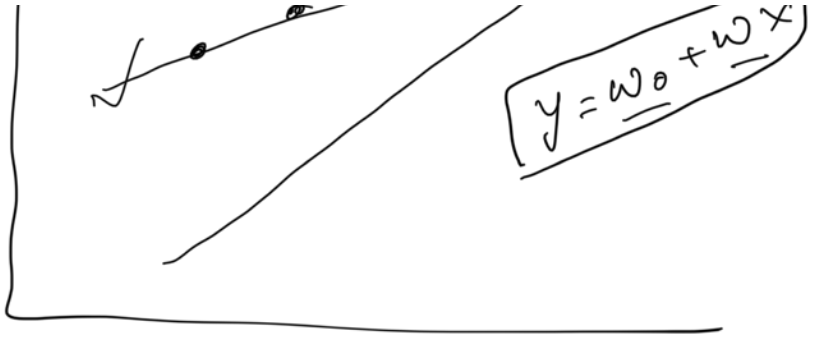
x	y
12	9
28	15
15	11
48	21
56	22

Find the best w_0, w_1 which "fits" the data best.

Wed, Feb 10

Resources on Piazza





Find w_0, w that does what?

X	Y	Pred. $\bar{Y}$
x_1	y_1	$\bar{y}_1$
x_2	y_2	$\bar{y}_2$
x_3	y_3	$\bar{y}_3$
⋮	⋮	⋮
x_N	y_N	$\bar{y}_N$

For a given w, w_0

$$\bar{y}_1 = w_0 + w x_1$$

$$\vdots$$

$$\bar{y}_i = w_0 + w x_i$$

Error: $e_i = y_i - \bar{y}_i$

$$J = \frac{1}{N} \sum e_i^2$$

$$J(w_0, w) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - (w_0 + w x_i)\right)^2$$

Squared loss function

$\frac{1}{2}$ → just for mathematical convenience

What if x is not 1-D

$\Rightarrow x \in \mathbb{R}^d$ where $d > 1$

$y = w_0 + \mathbf{w}^T x$

$\mathbf{w} \in \mathbb{R}^d$

Squared loss function

$$\underline{J(\mathbf{w})} = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T x_i\right)^2$$

absorbed

added a 1 to x_i

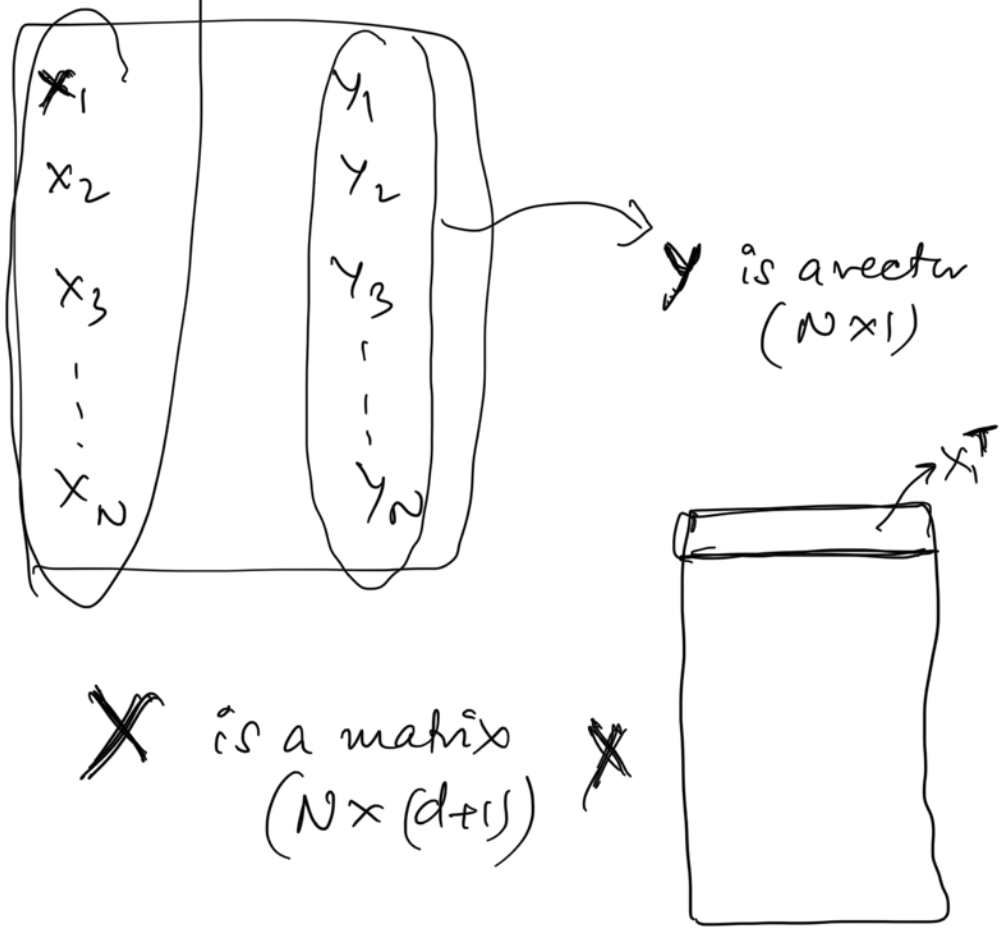
w_0 into w

Find w that minimizes $J(w)$

$$w^T x_i = \sum_{j=0}^{d} (w_j \, x_{ij})$$

$w \cdot x_i$

$w^T x_i$

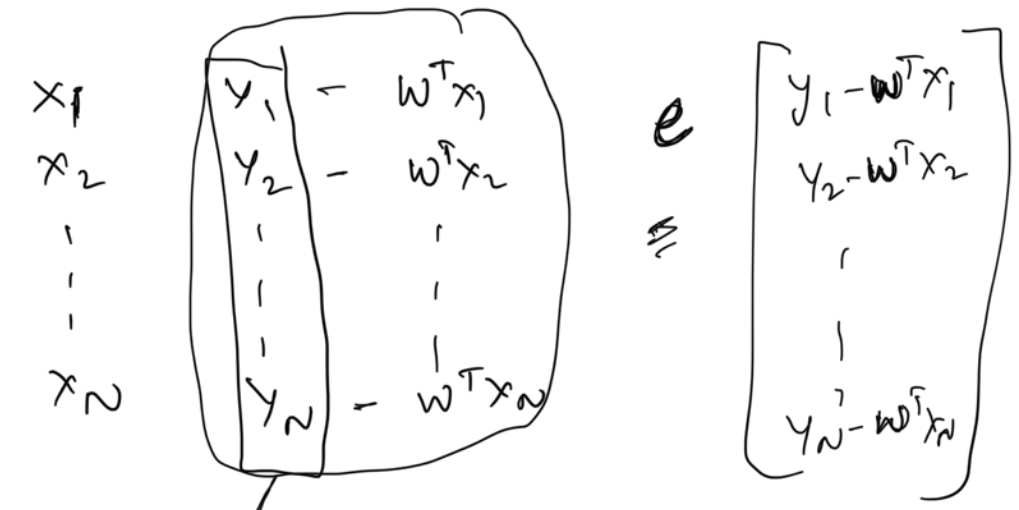


Training data: X, y — target vector

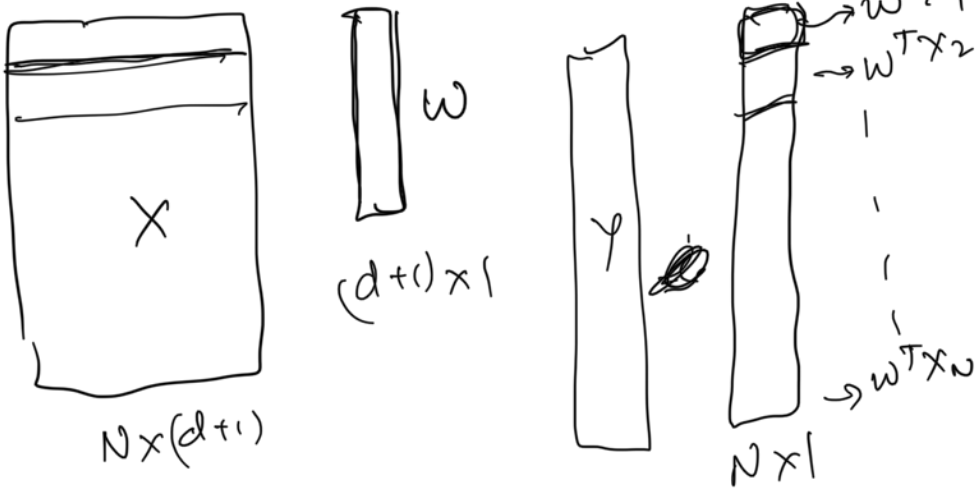
$(N \times (d+1))$ $(N \times 1)$

data matrix

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



$$e = (y - Xw)$$



e

$$\sum (y_i - w^T x_i)^2 \equiv (y - Xw)^T (y - Xw)$$

$$e_i = y_i - w^T x_i$$

$$e = \begin{bmatrix} y_1 - w^T x_1 \\ \vdots \\ y_N - w^T x_N \end{bmatrix}$$

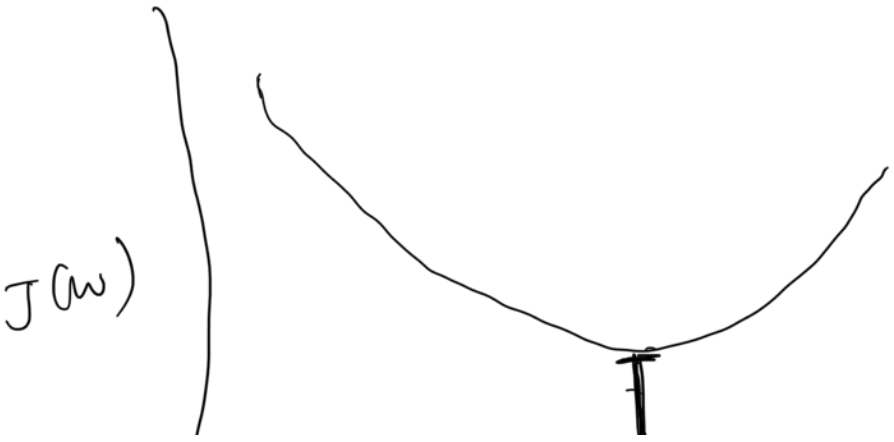
any vector p

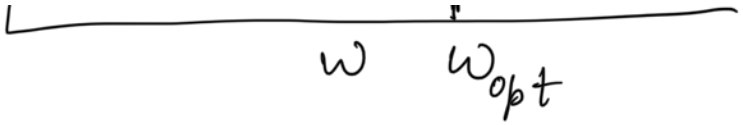
$$p^T p = \sum p_i^2$$

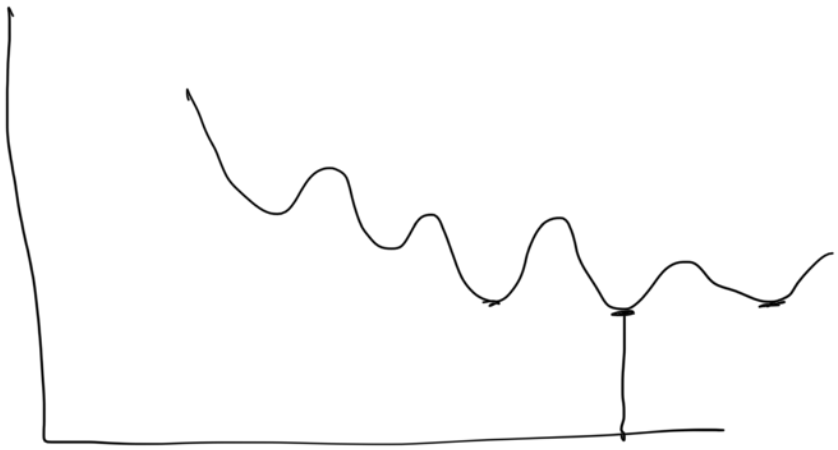
$$\sum (y_i - w^T x_i)^2 = (y - Xw)^T (y - Xw)$$

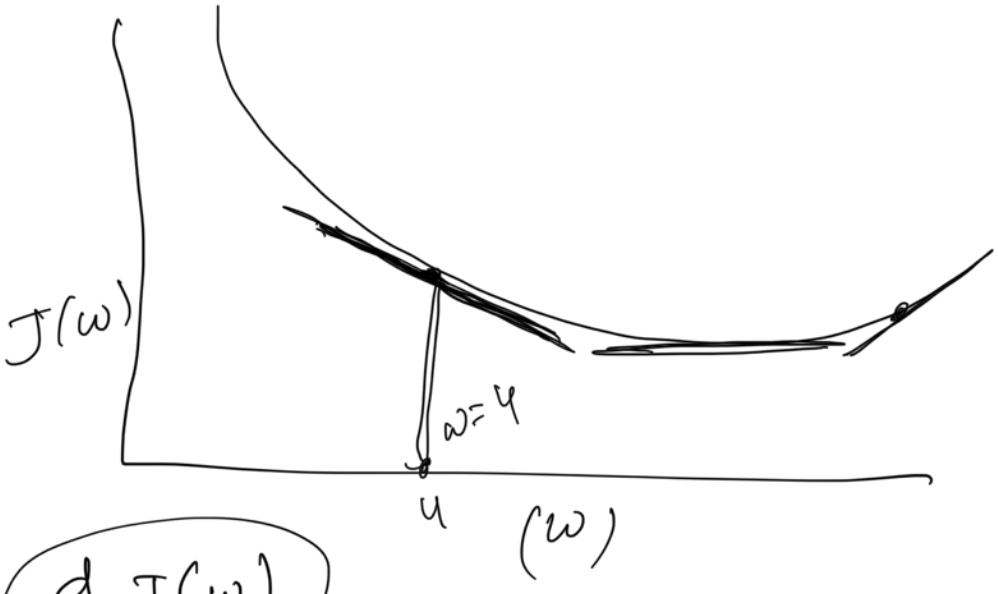
$$J(w) = \frac{1}{2} (y - Xw)^T (y - Xw)$$

How do we find w that minimizes $J(w)$









$\frac{d}{dw} J(w)$

e.g $J(w) = 3w^3 + 4w$

$\frac{dJ(w)}{dw} = 9w^2 + 4$

at $w = 4$

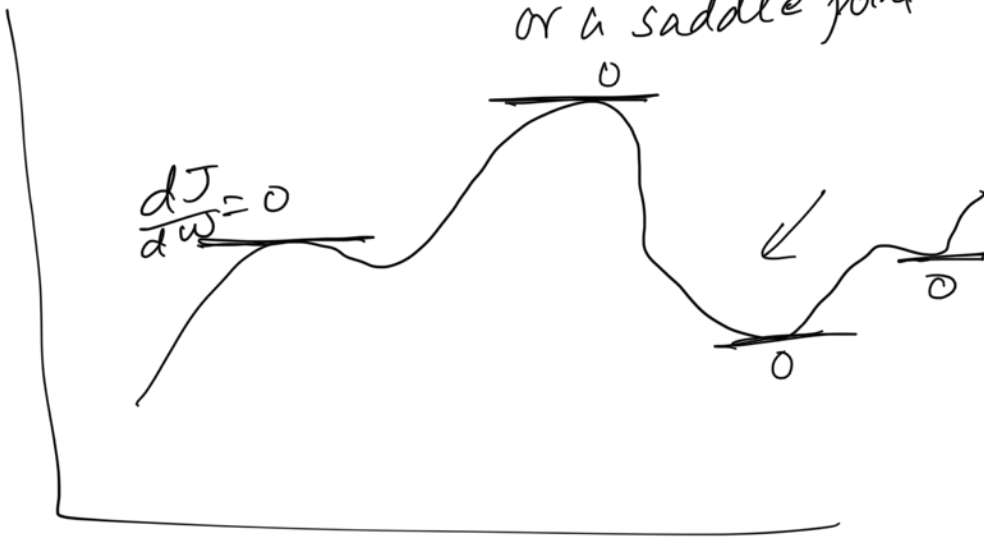
$\frac{dJ(w)}{dw} = \underline{148}$

if $\frac{d(J(w))}{dw} = 0$ at a given w

that means $w \rightarrow$ point of minima

or maxima

or a saddle point



If $J(w)$ is convex

then $\boxed{\frac{d\,J(w)}{dw} = 0}$

solution for this will give us w for which $J(w)$ is minimum

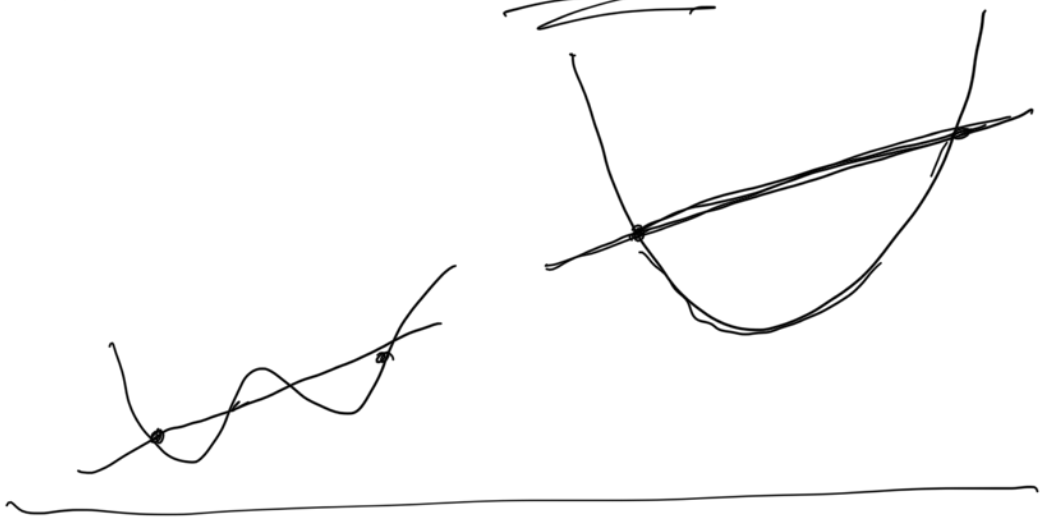
Feb 12

$$f(w) = 7w^3 - 4w^2 + 8$$

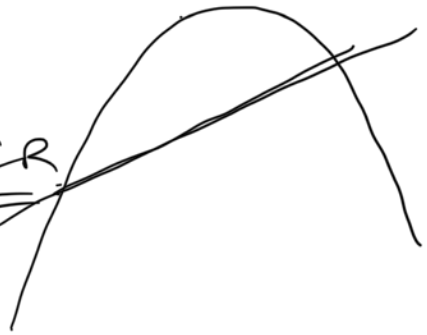
$f(w)$: scalar

$$\frac{d}{dw} f(w) = f'(w) = 21w^2 - 8w$$

$$\frac{d^2}{dw^2} f(w) = \frac{d}{dw}\left(\frac{df(w)}{dw}\right)$$

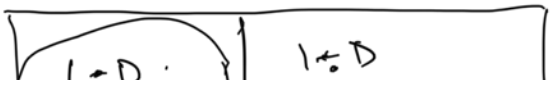


Squared loss function for LR is convex



$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

Output



$$\frac{1}{2}\sum (y_i - w^T x_i)^2$$

$$\frac{d}{dw} J(w)$$

$$= \nabla J(w)$$

gradient

$$f(w) = 3w_0^2 + 7w_1 + 8w_0 + 9$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\frac{\partial}{\partial w_0} f(w) = 6w_0 + 8$$

$$\frac{\partial}{\partial w_1} f(w) = 7$$

$$\frac{d}{dw} f(w) = \begin{bmatrix} \frac{\partial}{\partial w_0} f(w) \\ \frac{\partial}{\partial w_1} f(w) \end{bmatrix}$$

Gradient

$$\nabla f(w)$$

$$\frac{d^2}{dw^2} f(w)$$

$$\frac{d}{dw} \nabla f(w)$$

$$f(w) = \begin{bmatrix} 7w_0^2 + 8w_1 \\ w_0 w_1 + 9w_1^3 \end{bmatrix}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\frac{d}{} f(w)$$

$= H$

Hessian of $f(w)$

$dw = \begin{bmatrix} 14w_0 & 8 \\ w_1 & w_0 + 27w_1^2 \end{bmatrix}$

Jacobian

$J(w)$

Calculate $\nabla J(w) \equiv \frac{d}{dw} J(w)$

and then solve for w : $\boxed{\nabla J(w) = 0}$

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

$$J(w) = \frac{1}{2}\left[(y^T - (Xw)^T)(y - Xw)\right]$$

$$= \frac{1}{2}\left[(y^T - w^T X^T)(y - Xw)\right]$$

$$= \frac{1}{2}\left[(y^T y) - y^T Xw - w^T X^T y + w^T X^T X w\right]$$

$$= \frac{1}{2}\left[y^T y - 2 w^T X^T y\right]$$

$(A+B)^T = A^T + B^T$

$(AB)^T = B^T A^T$

$a^T b = b^T a$

$y^T Xw$: $z_{(N \times 1)}$

$+ w^T X^T X w \Big]$ → $y^T z$, $z^T y$

$$\frac{d}{dw} J(w) = \frac{1}{2}\left[\frac{d}{dw} y^T y - 2 \frac{d}{dw} w^T X^T y + \frac{d}{dw} w^T X^T X w \right]$$

$$= \frac{1}{2}\left[-2 X^T y + 2 X^T X w \right]$$

$$= -X^T y + X^T X w$$

Set $\frac{d}{dw} J(w) = 0$ $\quad \left[\nabla f(x) = 0 \right]$

$$-X^T y + X^T X w = 0$$

$$X^T X w = X^T y$$

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

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

X - $N \times (d+1)$

$A^{-1} A = I$

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

y — $N \times 1$

$X^T X$ = $(d+1) \times (d+1)$

Scalability

A $d \times d$ $\quad A^{-1}$ $\quad O(d^3)$

$\frac{1}{x} \rightarrow x = 0$

A^{-1} → is not easy to calculate

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

$A^{-1} \rightarrow Z$

Mon, Feb 15

W

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

Evaluation Metric

$(y_i - \hat{y}_i)^2$

$$+ (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2 + (y_5 - \hat{y}_5)^2$$

Root Mean Squared Error

$$\text{(RMSE)} \quad \sqrt{\frac{1}{N} \sum (y_i - \hat{y}_i)^2} \qquad \hat{y}_i = w^T x_i$$

unit — cm

Squared Loss $\quad \frac{1}{2} \sum (y_i - \hat{y}_i)^2 \qquad \hat{y} = w^T x_i$

What data to measure error on?

Training data?

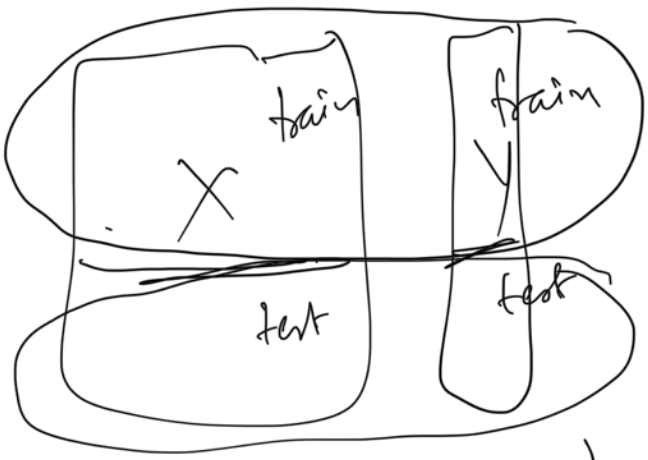
Generalization Error

Testing error

MAE: $\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

Scaling ⟷

Price = − 0.108 CRIM
+ 0.0464 ZN
+ ----



GD - LR

Choose some w (init)

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

Calculate ∇J

$w \leftarrow w - \eta \nabla J$

Choose some w (init)

for $i = 1$ to N:

while not converged

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

∇J

$w \leftarrow w - \eta \nabla J$

Stochastic Gradient Descent

SGD