Last Class – June 3

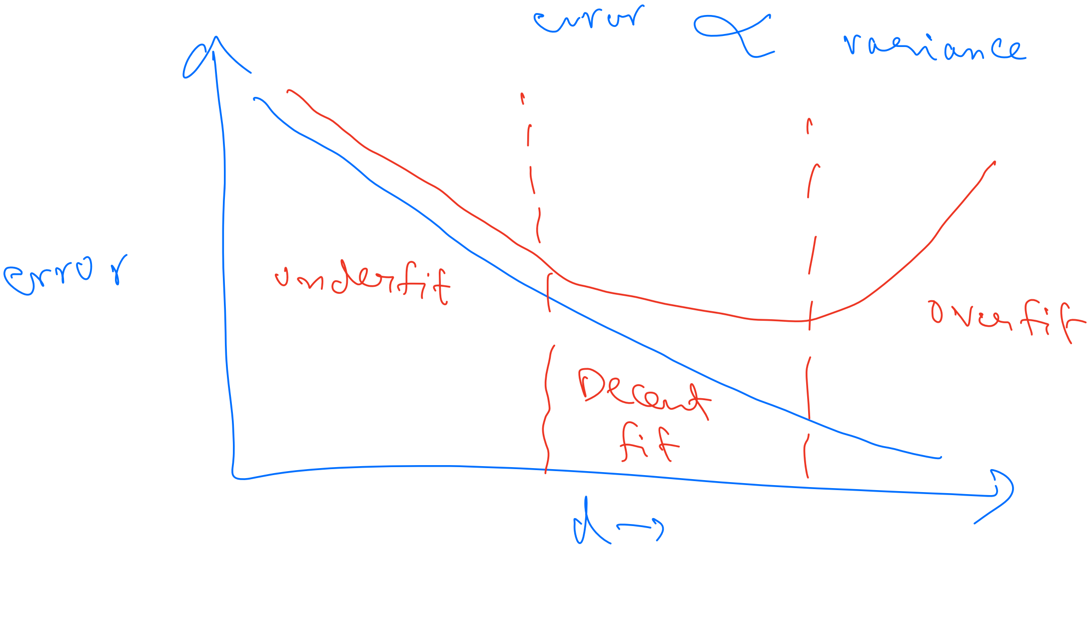
1) Generalization & Occam's Razor
2) Underfitting, Overfitting, Tradeoff
3) Bias – Variance Tradeoff
4) Regularization Overview

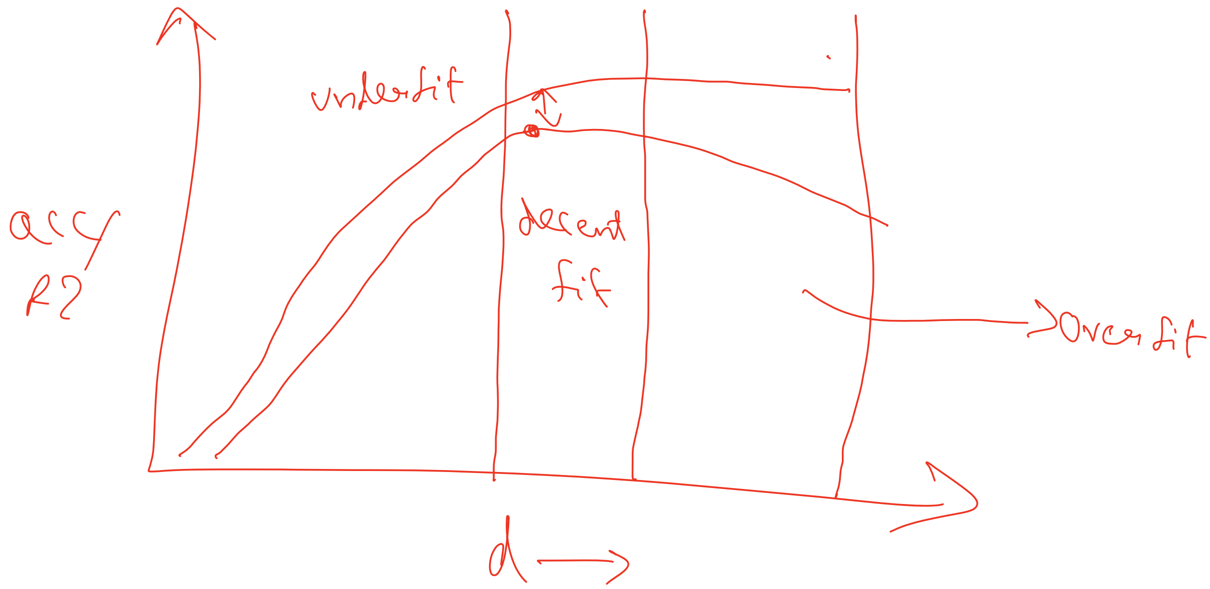
Today

1) Recap – Quizzes
2) L2, L1 & Elastic Net Regularization
3) Cross-Validation
4) Logistic Regression Overview

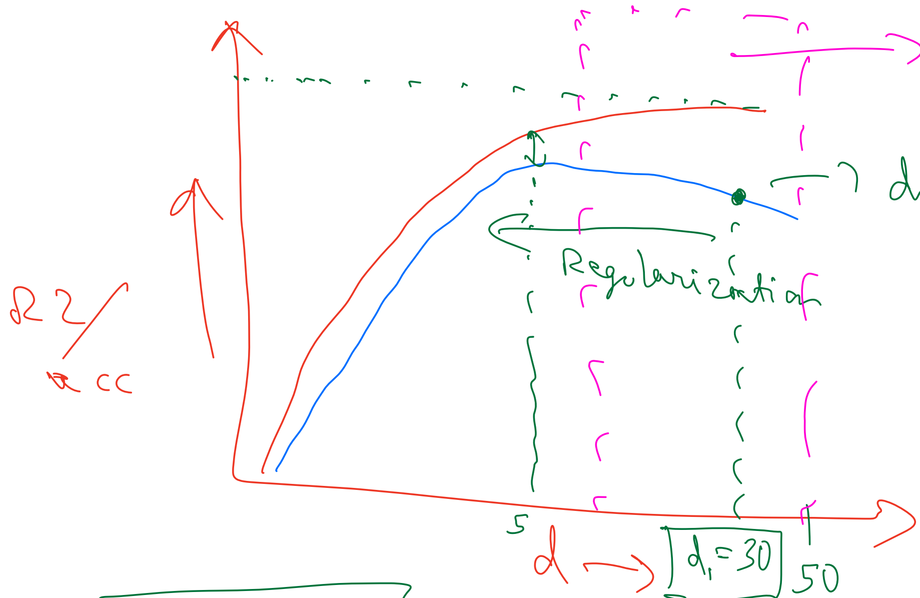
High or Low Bias → training error
- high → high bias
- low → low bias

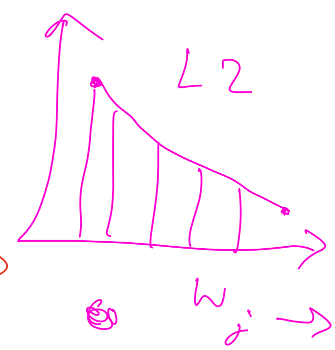
diff between train & validation error ∝ variance





Regularization





$\boxed{d_1 = 30}$

fit (d = 30)

train acc
val acc } → huge gap

Coeff of the imp. degrees → significant values

Coeff of the non-imp. degrees → close to zero data $\rightarrow$ degree 5

$\boxed{\begin{array}{l} \text{degree } 0 \text{ to } 5 \rightarrow \text{significant values} \\ \text{degree } 6 \text{ to } 30 \rightarrow \text{close to zero values} \end{array}}$

$\rightarrow$ Regularization

L2 $\rightarrow$ Ridge / Tekhonov

L1 $\rightarrow$ Lasso

$\boxed{\text{L2:}}$

$\downarrow$

$\boxed{\text{Ridge}}$

$$\sum_{i=1}^{N}\left(y_i - \sum_{j}^{D} w_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{D} w_j^2$$

regularization constant

$\downarrow$

hyperparameter

$\lambda_1 = 0 \Rightarrow$ no regularization (L2)
normal LR

$\boxed{\text{L1:}}$

$\downarrow$

$\boxed{\text{Lasso}}$

$$\sum_{i=1}^{N}\left(y_i - \sum_{j}^{D} w_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{D} |w_j|$$

$\lambda_2 = 0 \Rightarrow$ no L1 Reg
normal LR

Elastic Net

$$\text{MSE Error} + \boxed{\lambda_1} \underbrace{\sum_{j=1}^{D} |w_j|}_{} + \boxed{\lambda_2} \sum_{j=1}^{D} w_j^2$$

L1 L2

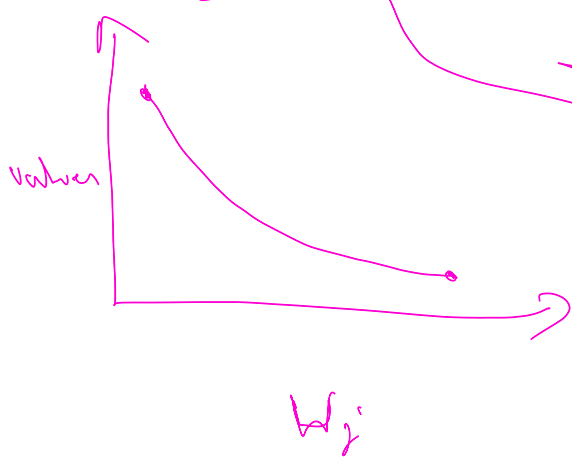
L2 : Loss fn = $\sum_{i=1}^{N}\left(y_i - \sum_j w_j x_{ij}\right)^2 + \lambda_2 \sum_{j}^{D} w_j^2$

$w_j = w_j - \eta \frac{\partial L}{\partial w_j}$ GD update

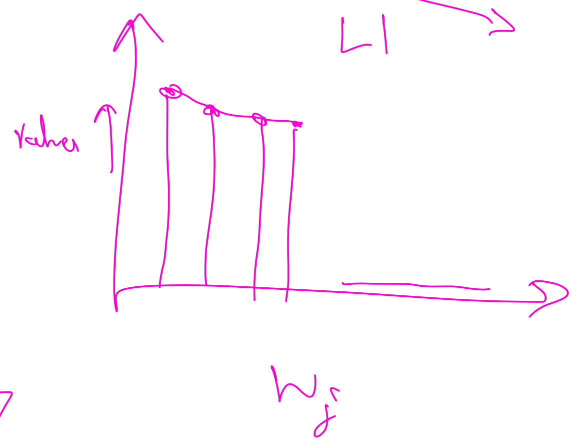
$\frac{\partial L}{\partial w_j} = 2\sum_{i=1}^{N}\left(y_i - \sum w_j x_{ij}\right)(-x_{ij})$ -ve ①

$+ \lambda_2 (2 w_j)$ ② +ve

L2



L1



$y = 7x + 8x^2 + \dots + 0.009x^{10} + 0.0009x^{11}$

coeff are imp

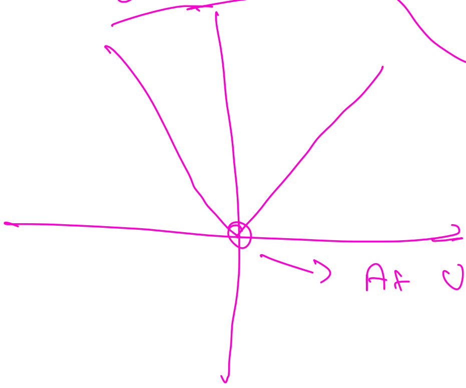
ignore

Lower degree polynomial

$$L1: \frac{\partial L}{\partial w_j} = 2\sum_{i=1}^{N}\left(y_i - \sum w_j x_{ij}\right)(-x_{ij}) + \lambda\,(0/+1/-1)$$

based on sign of w_j

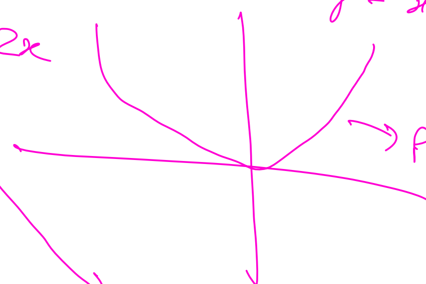
$y = |x|$



At 0, non-differentiable

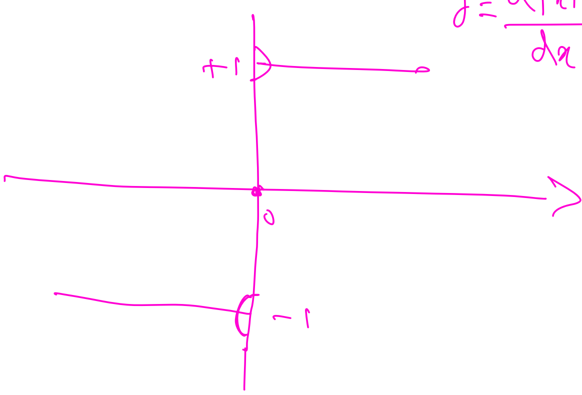
$$\frac{d(x^2)}{dx} = 2x$$

$y = x^2$

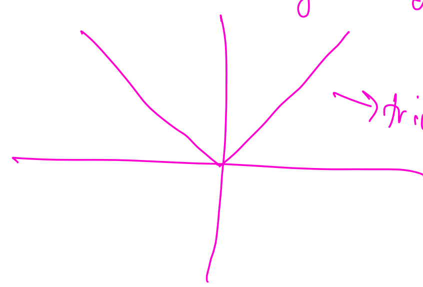


parabola

$$y = \frac{d|x|}{dx}$$



$$y = \frac{d(x^2)}{dx}$$



triangle

$$y = |x|$$

$$\frac{dy}{dx} = +1, \quad x > 0$$
$$= -1, \quad x < 0$$
$$= 0, \quad x = 0$$

L2 reg: Coeffs can be close to 0 but not exactly 0

L1 reg: some coeff will be 0

L2

$w_j = 0.5, \eta = 0.1$

$0.5 - 0.5 \times 0.1 = 0.45$

$0.45 - 0.45 \times 0.1 = 0.405$

$\vdots$

$0.405 - 0.405 \times 0.1 =$

L1

$w_j = 0.5, \eta = 0.1$

$0.5 - \boxed{1} \times 0.1 = 0.4$

$0.4 - 1 \times 0.1 = 0.3$

$0.3 - 1 \times 0.1 = 0.2$

$0.2 - 1 \times 0.1 = 0.1$

$0.1 - 1 \times 0.1 = 0$

1) L1 → less number of imp features

↳ feature selection

harmful for less number of features

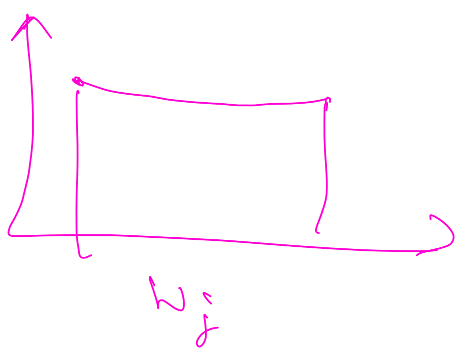
when we already have huge number of features

Gene data ($d >> n$)

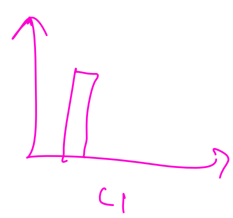
2) Outliers → Sensitive to outliers → L2

robust to outliers → L1

3)



want my weights to be uniform → L2



Q) $f_1, f_2, \dots f_{10}$

$f_1, f_2 \rightarrow 0.9$

$f_3, f_{10} \rightarrow 0.95$

L1 $\rightarrow$ make $\boxed{\text{one}}$ $\rightarrow$ few of them disappear (0 coeff)

$\boxed{\text{Elastic net}}$ $\rightarrow$ Combination of L1 & L2

How do we choose λ_1 & λ_2

Q) How did we choose optimum degree d for polynomial regression?

We used validation set

$\rightarrow$ We will use validation set

validation set $\longrightarrow$ optimum d, λ_1, λ_2

$$\rightarrow \left[\begin{array}{l} d = [0, 1, 2, \dots, D] \\ \lambda_1 = [10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 10, \dots] \\ \lambda_2 = [10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 10, \dots] \end{array}\right\} \text{Range of values}$$

$(\alpha_1, \lambda_{11}, \lambda_{12})$

$\leftrightarrow$ Triplets

iterate over all triplet combinations

Validation set — How is it defined?

Dataset of 1000 data points (X_i, y_i)

$X_i = [x_{i1}, x_{i2}, \dots x_{iD}]$

Build a ML model

→ Train 60%

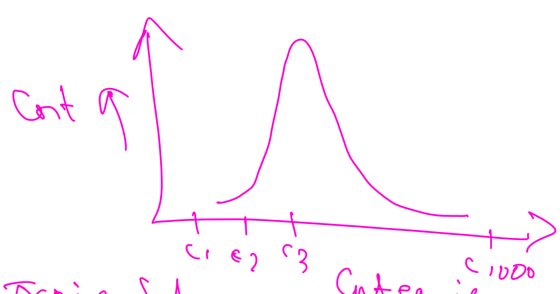
→ Val 20%

→ $\boxed{\text{Test}}$ 20%

random split

$\boxed{\text{Human help}}$ → Select the test data

Train + Val → 80 → 60, 20



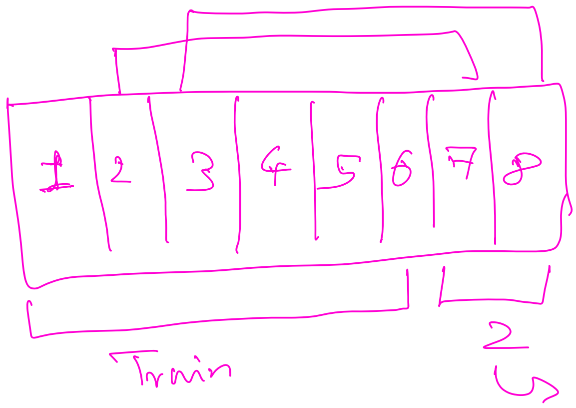
Train Set

Val set

Distributions are different

k-fold Cross validation

Train + Val → 80%



80 data points

60 → 6
20 → 2

1st iteration : Train 1 – 6, val – 7 – 8

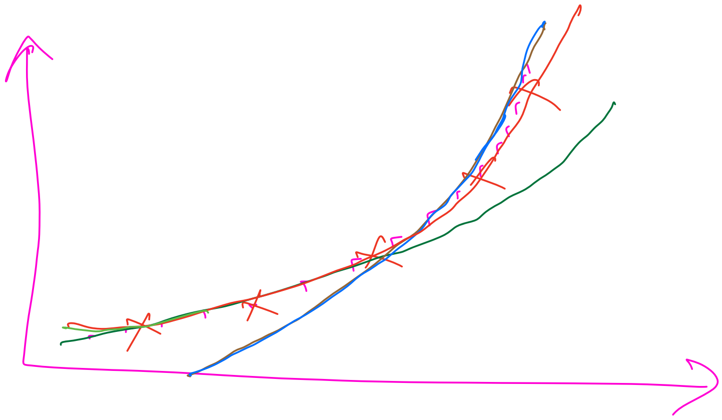
hyperparameters
accuracy

2nd iteration : Train 2 – 7, val – 8, 1

3rd iteration : Train 3 – 8, val – 1, 2

8 iterations :

ACC :
1st iteration → 90%
2nd iteration → 30%
avg acc → 60%



Logistic Regression Overview

Linear Regression → Linear Model for Regression Task

Logistic Regression → Linear Model for classification Task

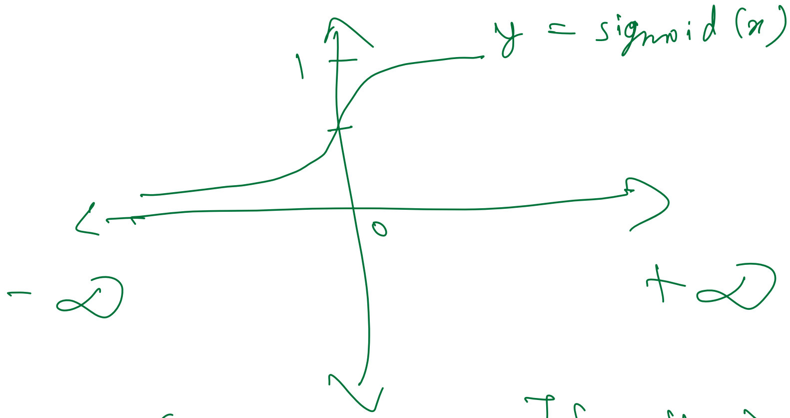
2 classes → 0
1

Binary classification

$$\underset{\text{output}}{y} = \boxed{w_0 + w_1 \underset{\text{input}}{x}} \rightarrow \text{Linear Regression}$$

$$z = w_0 + w_1 x$$

$$\underbrace{\boxed{\text{sigmoid}(z)}}_{\text{prob of falling into class 1}} = \boxed{\frac{e^z}{e^z + 1}} = \boxed{\frac{1}{1 + e^{-z}}}$$



$y = \text{sgmd}(x) \to 0$ when $x \to -\infty$

$y = \text{sgmd}(x) \to 1$ when $x \to +\infty$

Thresh (Th) = 0.5

If $y \geq 0.5$, o/p = 1

$y < 0.5$, o/p = 0

sgmd (x)

$$\text{sgmd}(x) = P(y = 1 \mid x)$$

$$P(y=0 \mid x) = 1 - P(y=1 \mid x)$$
$$= 1 - \text{sgmd}(x)$$

→ Multi-Class Logistic Regression

→ <u>Multi-Normial Logistic Regression</u>

class = 0, $\quad W_0 = (W_{01}, W_{02})$

$$z_0 = w_{01} + w_{02} \cdot x$$

class = 1, $\quad W_1 = (w_{11}, w_{12})$

$$z_1 = w_{11} + w_{12} \cdot x$$

Class = 2, $\quad w_2 = (w_{21}, w_{22})$

$$z_2 = w_{21} + w_{22} \cdot x$$

$$P(y=0|x) = \frac{e^{z_0}}{e^{z_0} + e^{z_1} + e^{z_2}}$$

$$P(y=1|x) = \frac{e^{z_1}}{e^{z_0} + e^{z_1} + e^{z_2}}$$

$$P(y=2|x) = \frac{e^{z_2}}{e^{z_0} + e^{z_1} + e^{z_2}}$$

Den

$\boxed{\text{Loss fn:}}$

3-class classification

$$\begin{array}{cccc} c_0 & c_1 & c_2 & \\ \rightarrow \ [1 & 0 & 0] & y_i: \\ \rightarrow \ [0 & 1 & 0] & y_i: \\ \rightarrow \ [0 & 0 & 1] & y_i: \end{array}$$

$$y_i = [c_{i0}, c_{i1}, c_{i2}]$$

$c_{i0} = (0, 1)$

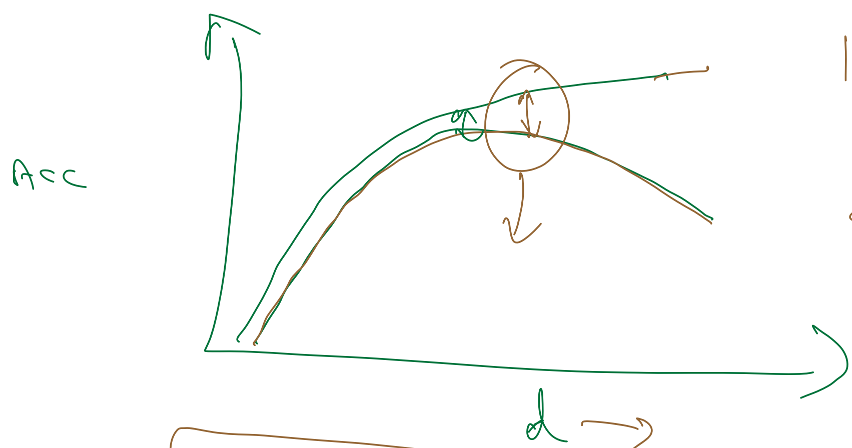
$c_{i1} = (0, 1)$

$c_{i2} = (0, 1)$

$$\underline{\text{Loss fn}} = -\sum_{i=1}^{N}\Big[c_{i0}\times\log\big(P(y=0|x)\big) + c_{i1}\times\log\big(P(y=1|x)\big) + c_{i2}\times\log\big(P(y=2|x)\big)\Big]$$

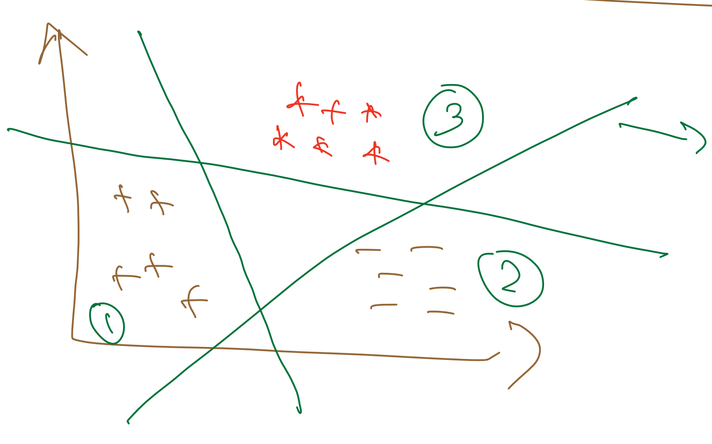
$$-\sum_i \underline{p_i}\log(p_i) \rightarrow \text{Entropy}$$

$$-\sum_i^{C} c_i\log(p_i) \quad \text{Cross-Entropy}$$



linear search
Binary search
Exponential search

$$\boxed{2,\ 4,\ 8,\ 16,\ 32,\ 64}$$



3 lines for 3 classes

w/o Reg

$$W = w - \frac{\partial L}{\partial w}$$ → −ve number → +ve

$$0 = 0 - (-0.1)$$

$$0.1 = 0.1 - (-0.2)$$

↓

0.3, 0.4, 0.5, . . .

−0.1, −0.2, −0.3, . . .

W Regularization

$$W = w - \frac{\partial L}{\partial W}$$

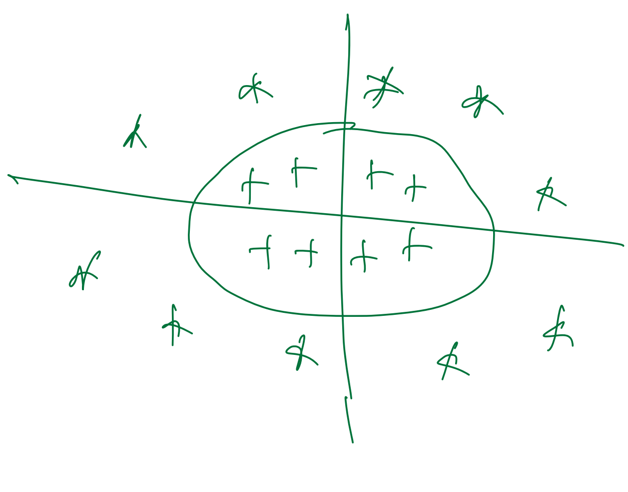
$$\frac{\partial L}{\partial w} = (+ve \quad -ve)$$

↓

len value

$$0 = 0 + 0.05$$

$$0.05 = 0.05 + 0.05 = 0.1$$



$$y = w_0 + w_1 x_1 + w_2 x_2 + \dots w_{10} x_{10}$$

$$X \quad \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_{10} \\ & & & & \\ & & & & \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{10} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$(N \times 11)$ $\quad$ (11×1) $\quad$ $(N \times 1)$

W

$$XW = Y$$

$\boxed{W} \rightarrow$ Find

$$XW = Y$$

$$X^T X W = X^T Y$$

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

I

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

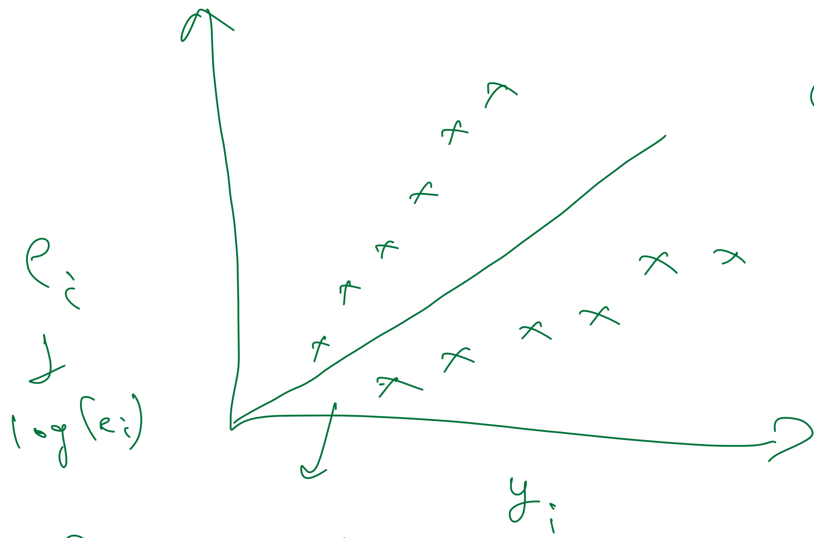
$$\boxed{W = (X^T X)^{-1} X^T Y}$$

Closed form solution of W

$\boxed{N = 1M}$

$$(X^T X)^{-1} \rightarrow \boxed{\text{not possible}}$$

if this exists or not



$e_i = f(y_i)$

Heteroskedasticity

$\sigma = f(y_i)$

<u>e_i is a linear fn of y_i</u>

$$e_i = y_i - \hat{y}_i$$

$$= y_i - (w_0 + w_1 x_i)$$

$\log(x_i)$

SGD Imm $\rightarrow$ <u>Assignment Q 7</u>