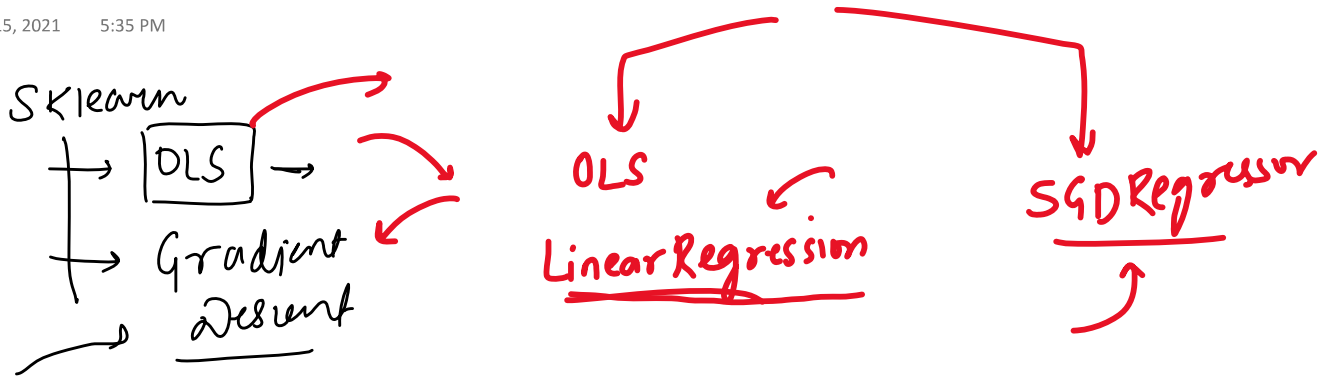
Why Gradient Descent

Saturday, May 15, 2021 5:35 PM



Code From Scratch

Monday, May 17, 2021 12:15 PM

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

(100,3) (100)

100 → (100,4)

X → matrix

	cgpa	iq	gender	lpa
1	—	—	—	→
1	—	-	-	—
				—

X → X_train

y → y train

diabetes → sklearn → R2- → same

→ my class → R2 →

Xtest β0 β1 → βn

3

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$= \beta_0 + np.dot(\beta, x\_test)$$

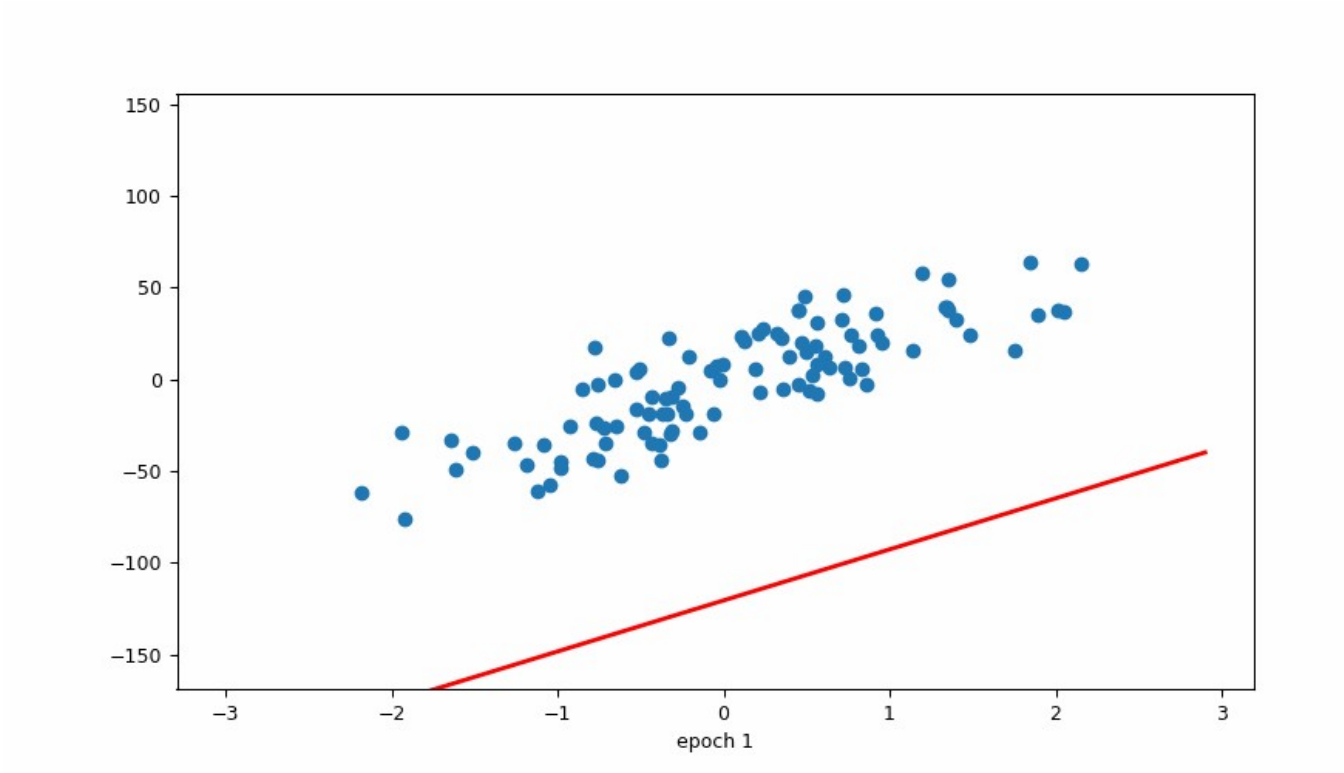
X_test coeff

(89,10) (10,1)

(89,1) + β0

89 → 1 → (89,1)

y pred



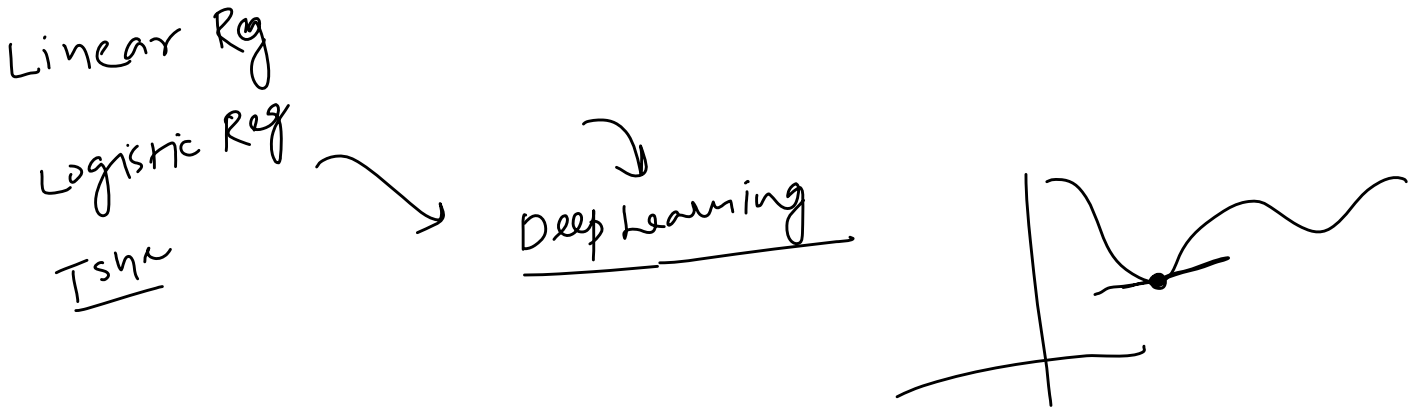
$$\sum (y_i - mx_i - b)^2$$

$$\sum -2 (y_i - mx_i - b)\, x_i$$

$$-2$$

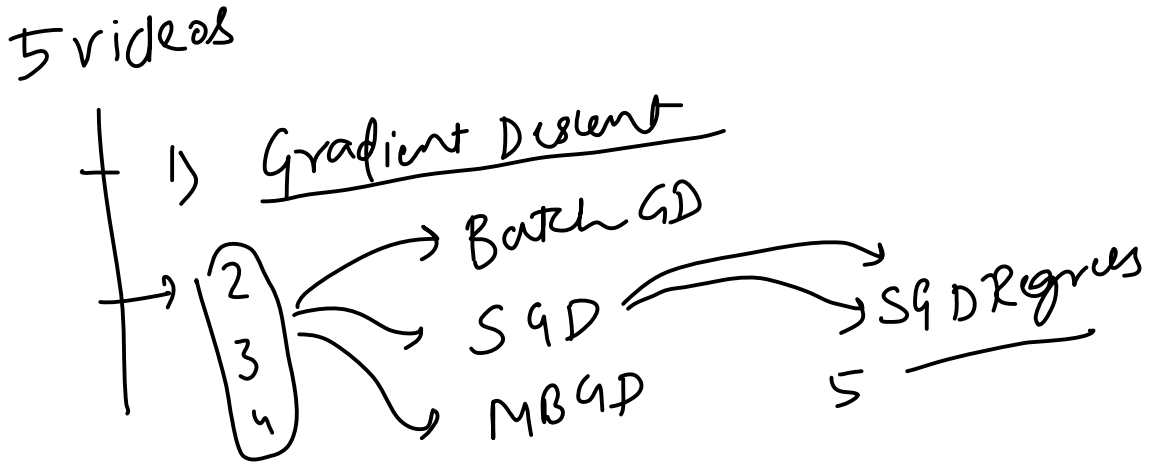
What is Gradient Descent?

Thursday, May 20, 2021 1:46 PM



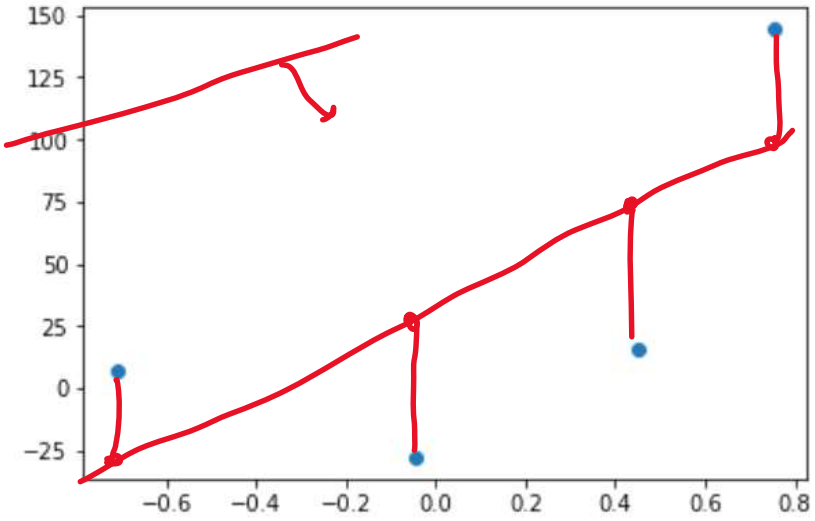
The Plan

Thursday, May 20, 2021 1:46 PM



Intuition

Thursday, May 20, 2021 1:46 PM



2 cols
4 rows

cgpa | lpa

OLS →

$\hat{y}_i = mx_i + b$

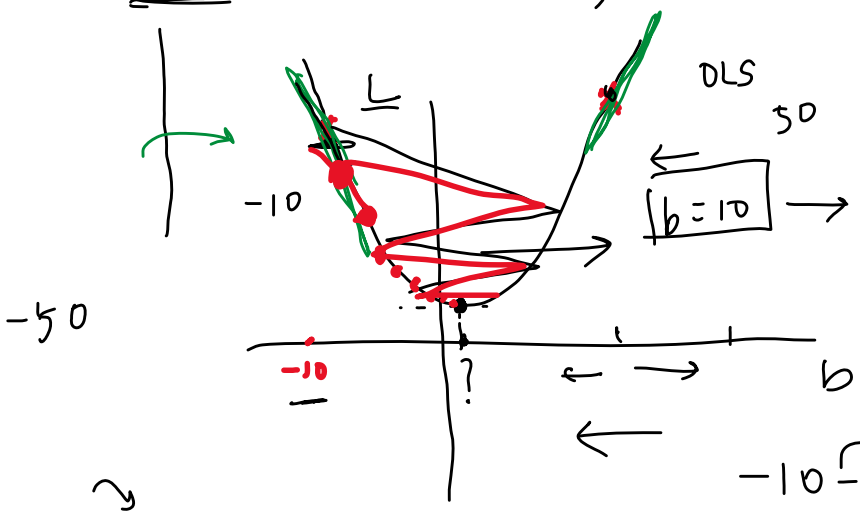
$m = 78.35$

$L = \sum_{i=1}^{4} (y_i - \hat{y}_i)^2$

$L = \sum_{i=1}^{4} (y_i - mx_i - b)^2$

$L \to b^2$

L(b)



$\min_{i=1}^{4} (y_i - 78.35 * x_i - b)^2$

step 1 - select a random b_{min}

b increment

b decrement

L_{min}

b = −10

x = 5

b = −10 → Slope = −ve

$b_{new} = b_{old} - slope$

$b_{new} = b_{old} - \eta \, slope$

−10 − (−50)

40

−40

10 − (56)

0.01

0.1, 0.001

$y = x^2 + 2$

$\frac{dy}{dx} = 2x$

−40

=10

−9.5

x = 50

−10

$b_{new} = -10 + (0.01 \times 50)$

$= -10 + 0.5 = -9.5$

$b_{new} = -9.5 - (0.01 * -40) = -9.5 + 0.4 = -9.1$

When to stop

1) diff b_{old} b_{new} => 0.0001

$b_{new} - b_{old} = 0.0001$

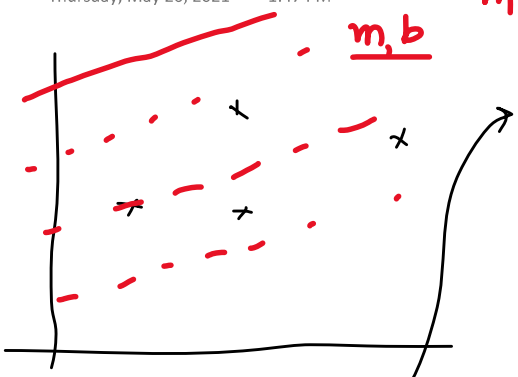
=> 0.0001

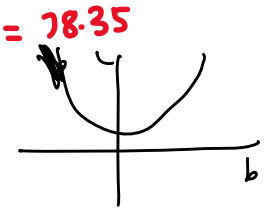
2) Iteration => 1000, 100,
epochs

$b_{new} - b_{old} = 0$

Mathematical Formulation

Thursday, May 20, 2021 1:47 PM





Step → Start with a random $b = b$

for i in epochs; 1000, 100

$i = 0$ 0, 1,

$\eta = 0.01$

$b = 0$

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

$$b_{new} = b_{old} - \eta * slope \quad (b = 0)$$

$$\frac{dL}{db} = \frac{d}{db}\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right) = 2\sum_{i=1}^{n} (y_i - mx_i - b)(-1)$$

$b = 0$

$$\frac{d}{db}\sum_{i=1}^{n} (y_i - mx_i - b)^2 \quad \text{slope} = -2\sum_{i=1}^{n} (y_i - mx_i - b)$$

$$= -2\sum_{i=1}^{n} (y_i - 78.35 * x_i - 0)$$

$i = 1$

epoch

slope ($b = 0$)

b_{new}

$$b_{new} = b_{old} - \eta \, slope \quad b = b_{old}$$

step size

Example

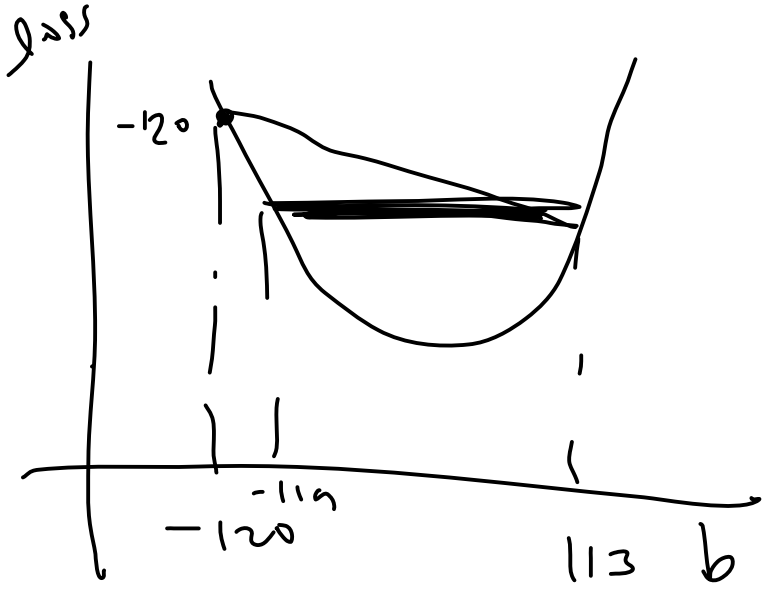
Thursday, May 20, 2021 1:47 PM

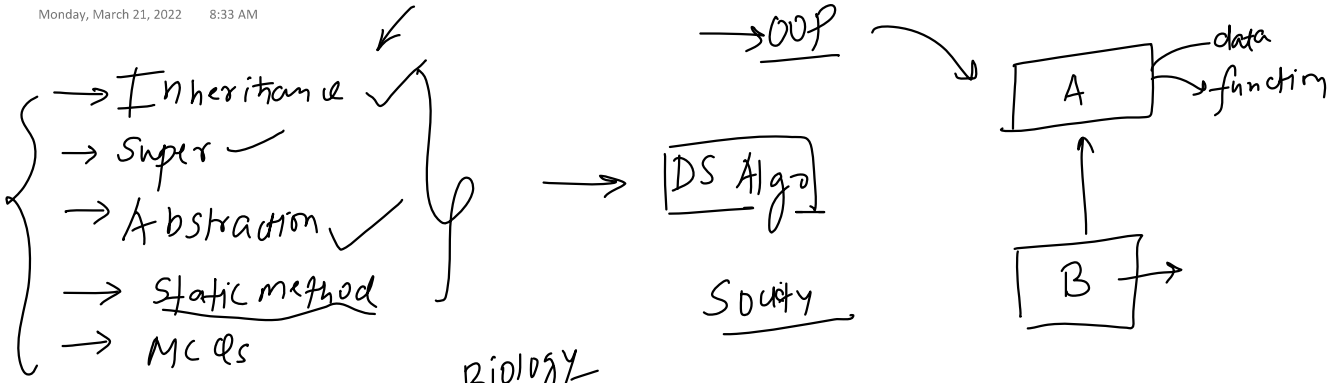
Code from Scratch

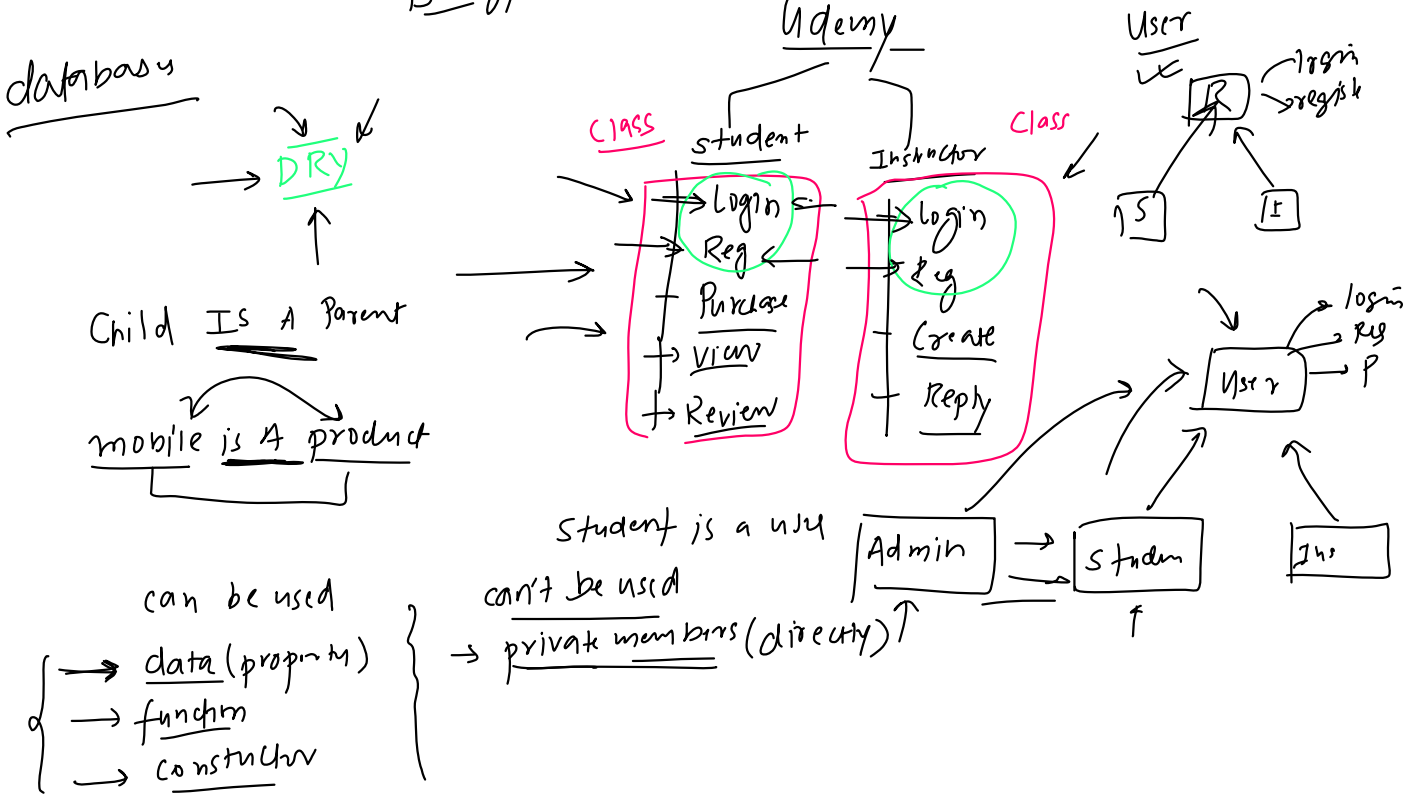
Thursday, May 20, 2021 1:48 PM

Visualization 1

Thursday, May 20, 2021 1:52 PM

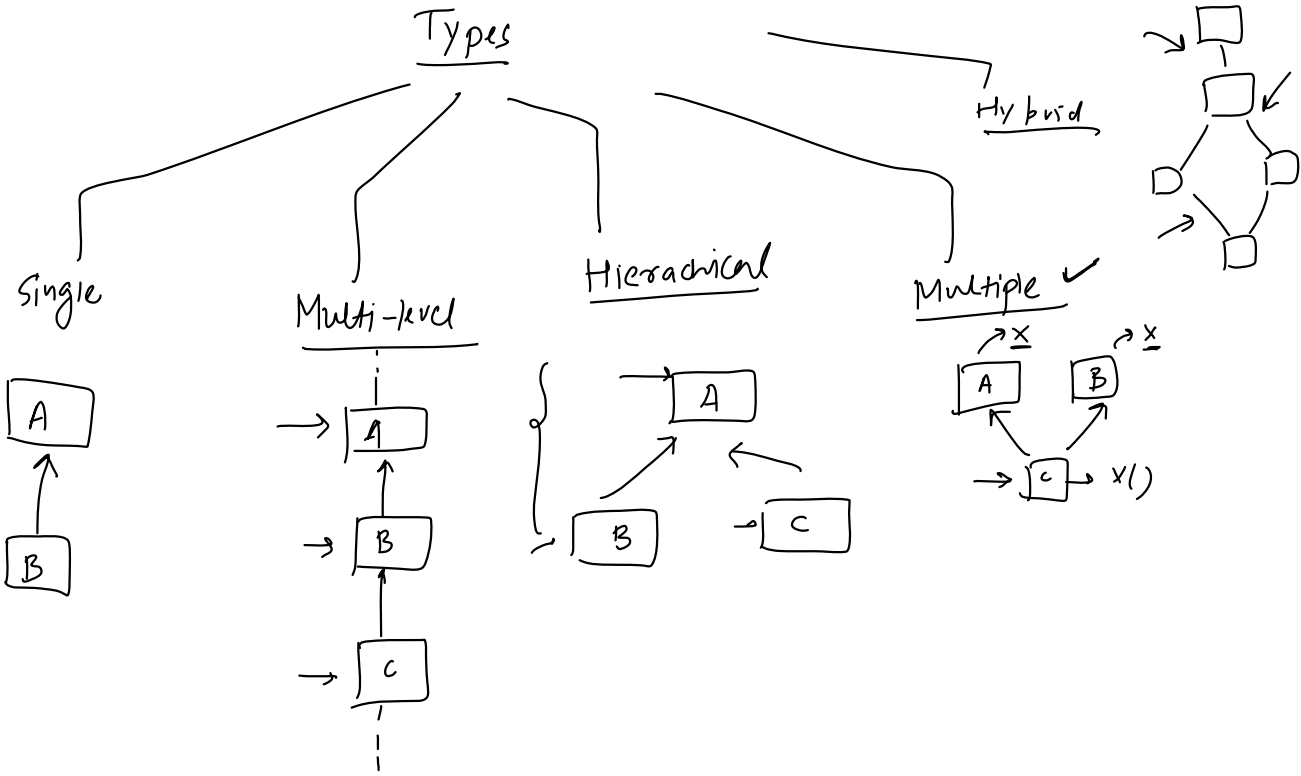


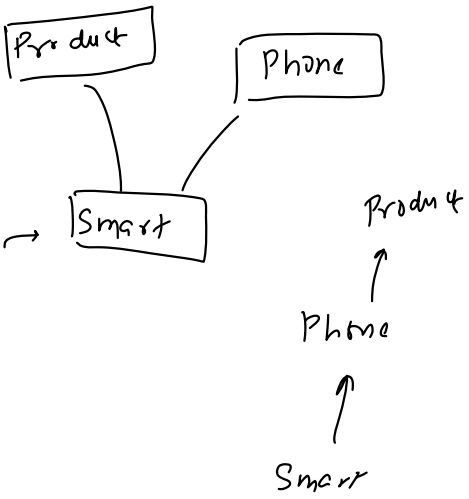


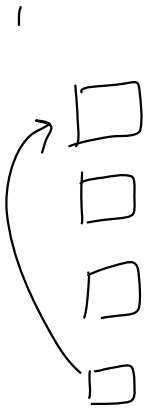


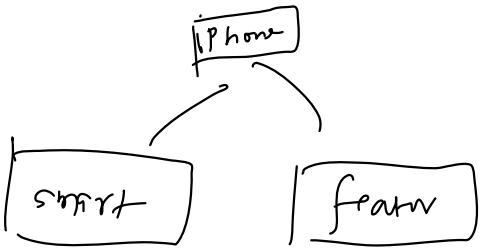
can be used
- data (property)
- function
- constructor

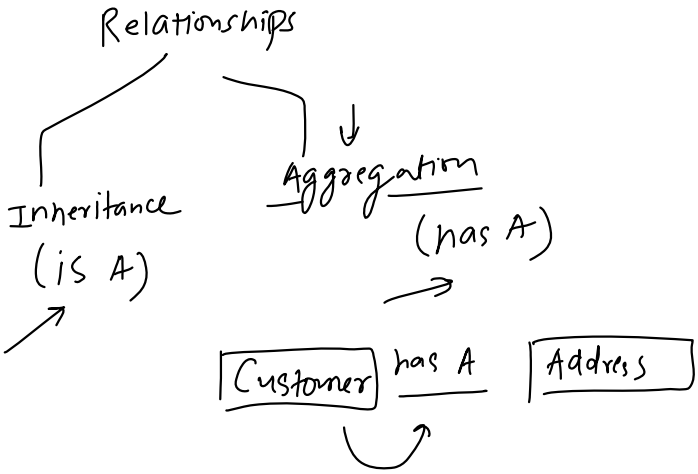
can't be used
- private members (directly)



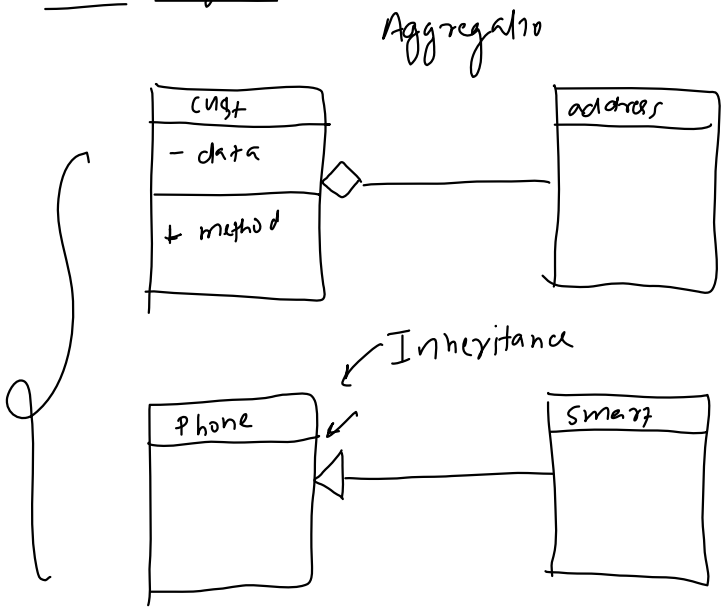




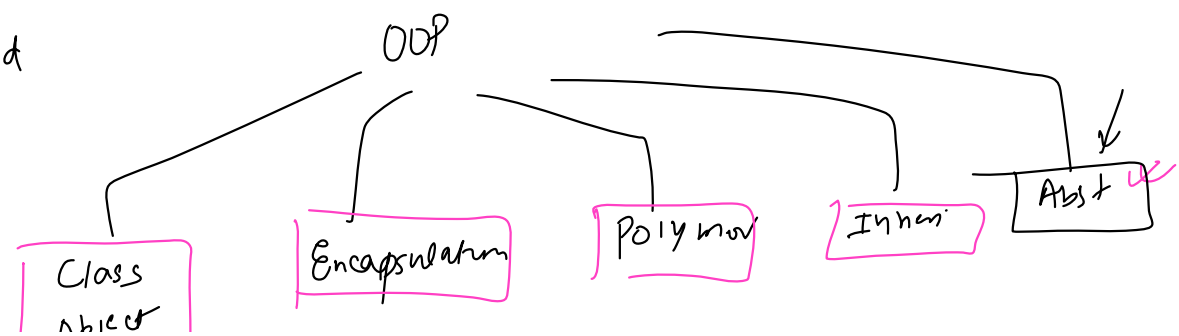




Class diagram



Construct
static method
inst vs class
super



Class
Object

Encapsulation

20 → 9 → DS Algo → 8 classes

Class 9 → Time Complexity
Class 10 → Dynamic Arrays (List)
Class 11 → LL
Class 12 → Stacks + Queues
Class 13 → Hashing → searching and sorting
Class 14 → Tree + Graph
Class 15 → Dynamic + greedy
Class 16 → Exam + MCQ + qns

Python → Excepn
→ file
→ yield

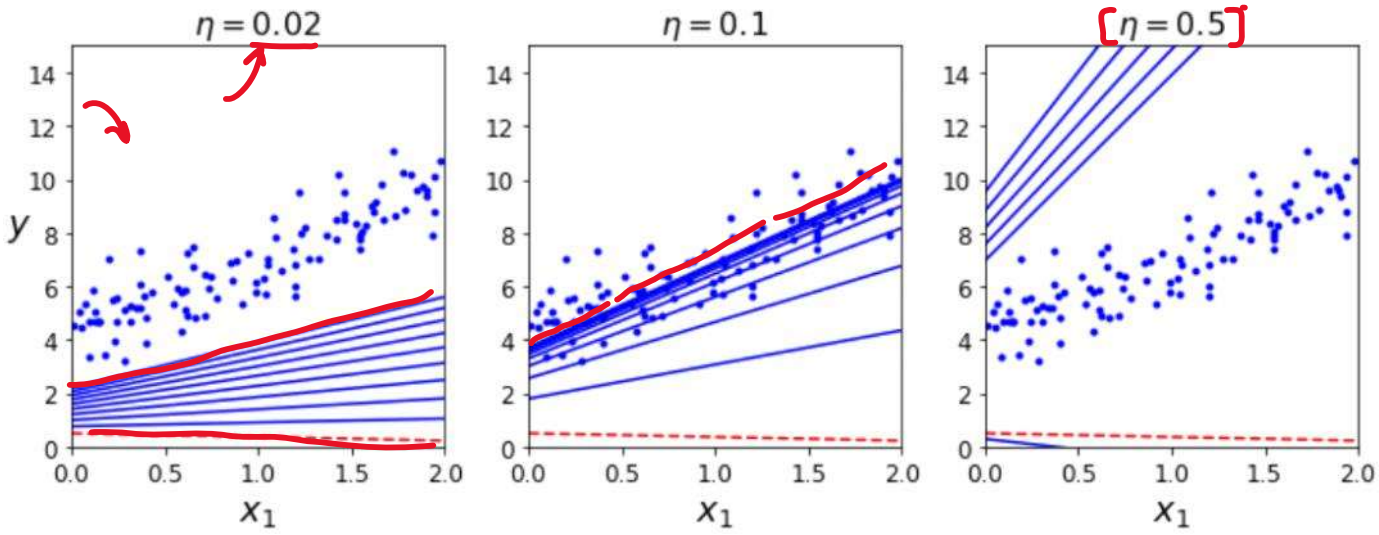
{ 17 18 19 20 }
SQL

Few Discussions

Thursday, May 20, 2021 4:47 PM

1. Effect of Learning rate
2. The universality of Gradient Descent

epoch = 10



$b = 0$

$b = b_{old} - \eta \cdot \text{slope}$

$\frac{dL}{db} = [\quad]$

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

Deep →

LOR → L function

Adding m into the mix

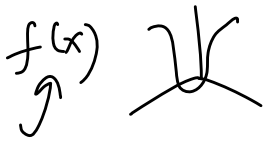
Thursday, May 20, 2021 1:48 PM

steps

1) init random vals for m and b
   $m = 1$ and $b = 0$

2) epochs = 100, lr = 0.01
   for i in epochs:
   $b = b - \eta$ slope
   $m = m - \eta$ slope

$z = f(x, y)$



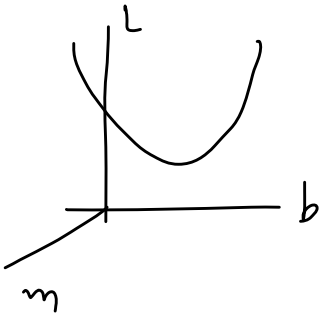
$b = 0$

$L = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - mx_i - b)^2$

$L(m, b)$

$\frac{dL}{db}$

b_slope = $\frac{\partial L}{\partial b}$



m_slope = $\frac{\partial L}{\partial m}$

$\sum (y_i - mx_i - b)^2$

2D gparlepa (m, b)

$\frac{\partial L}{\partial b} = -2 \sum (y_i - mx_i - b)$

$1 \quad 0 \quad 0 \quad -1$

$= -2 \sum (y_i - mx_i - b)$

= slope_b at $b = 0$

$\frac{\partial L}{\partial m} = 2 \sum (y_i - mx_i - b)$

$g \quad \nearrow \quad \nearrow \quad \nearrow$

$0 \quad -x_i \quad 0$

$\rightarrow$ (+)

$= -2 \sum (y_i - mx_i - b) x_i$

slope_m at $m = 1$

Code

Thursday, May 20, 2021 1:48 PM

Visualization 2

Thursday, May 20, 2021 1:52 PM

Effect of Learning Data

Thursday, May 20, 2021 1:53 PM

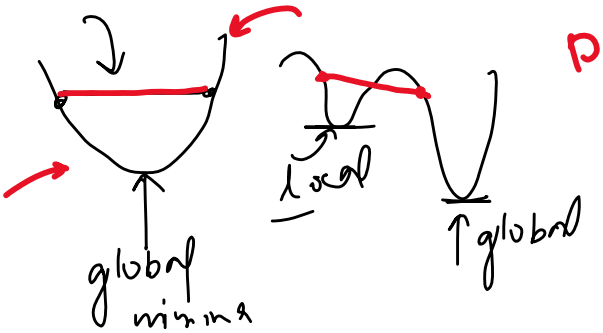
Effect of Loss Function

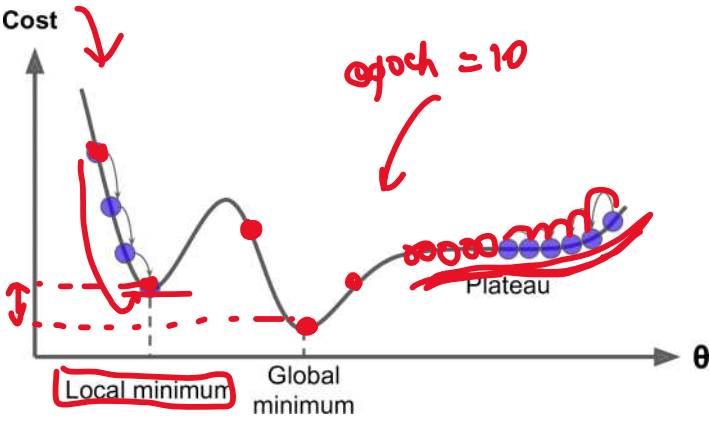
Thursday, May 20, 2021 1:53 PM

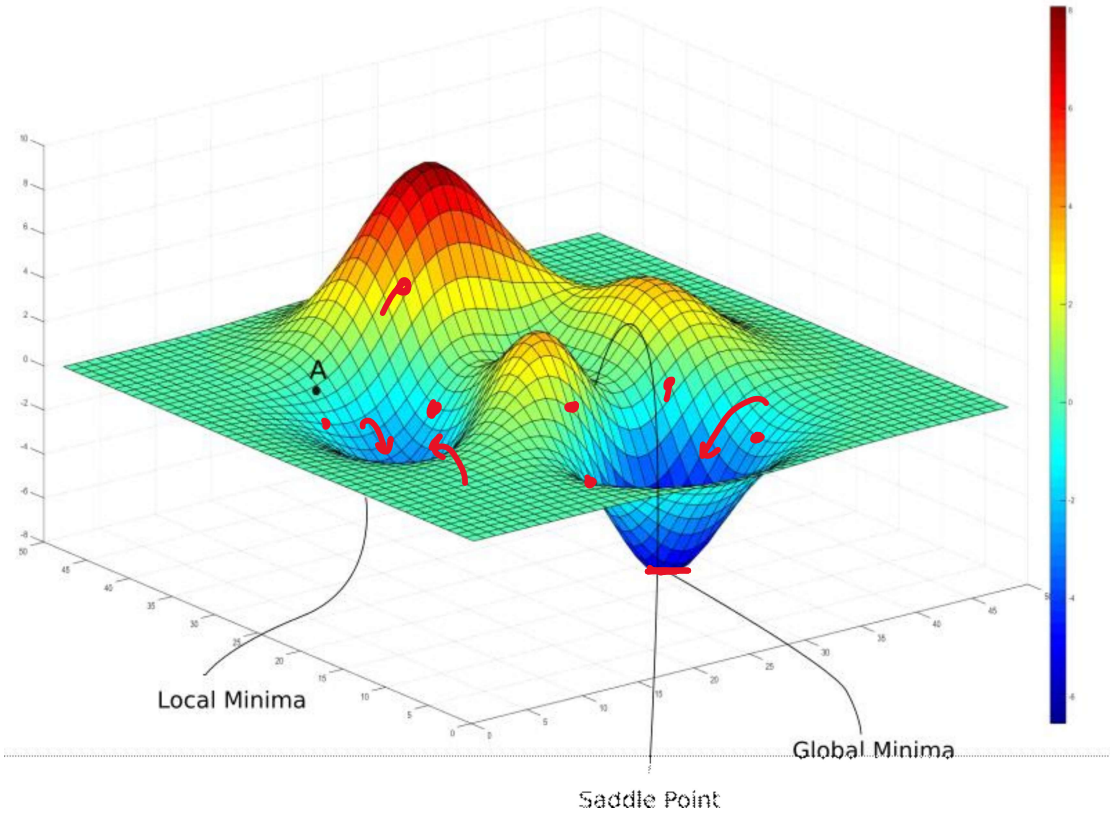
$L = \sum (y_i - \hat{y}_i)^2$ → convex function

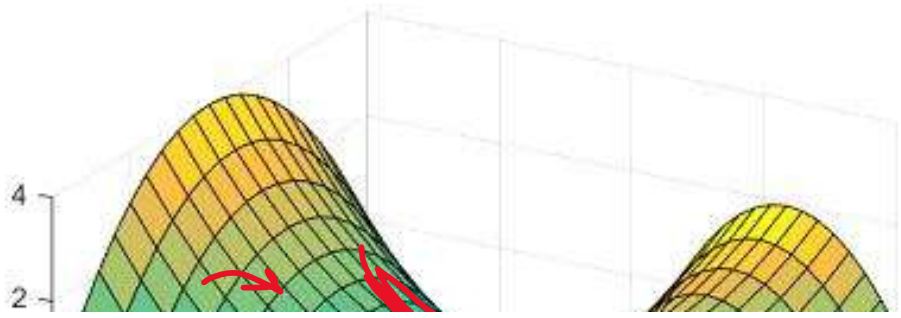
10 cols

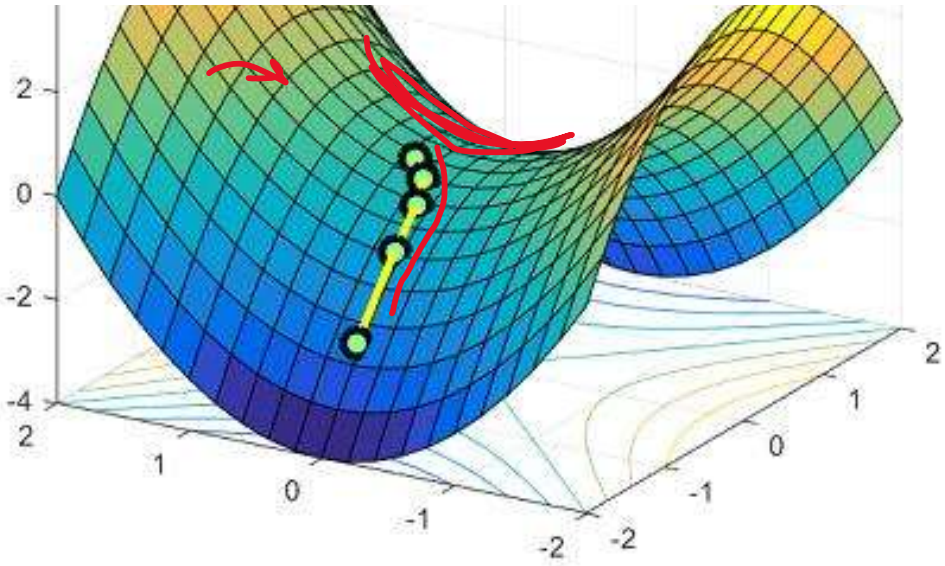
deep learn





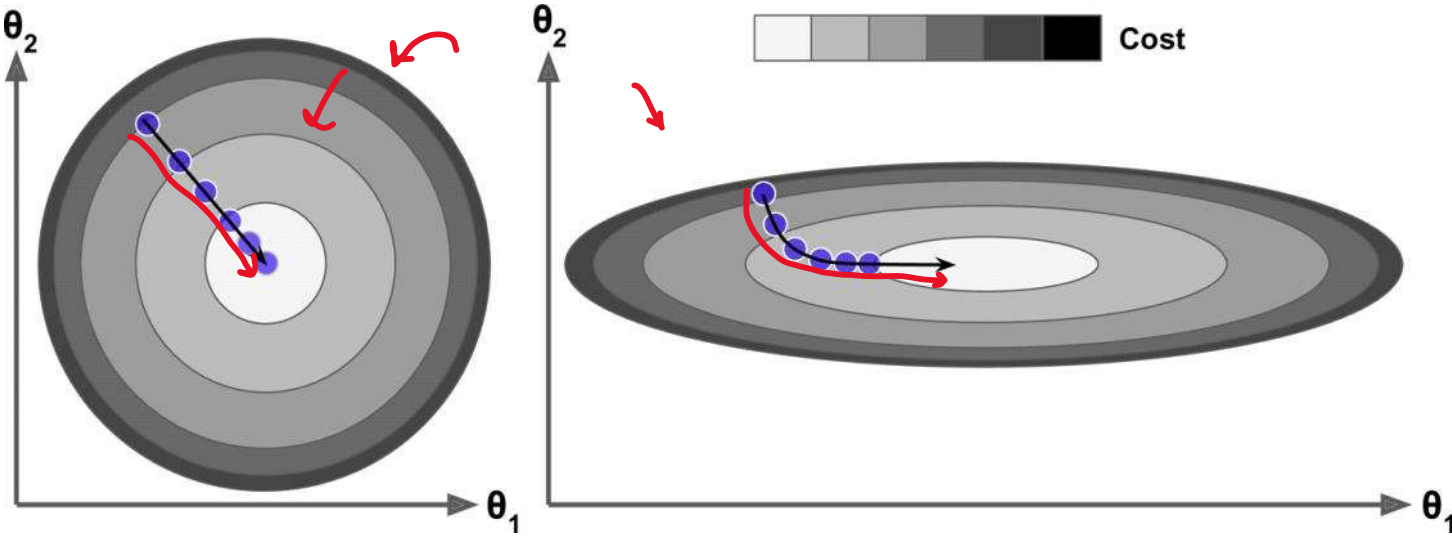





2
0
-2
-4
2
1
0
-1
-2
-2
-1
0
1
2

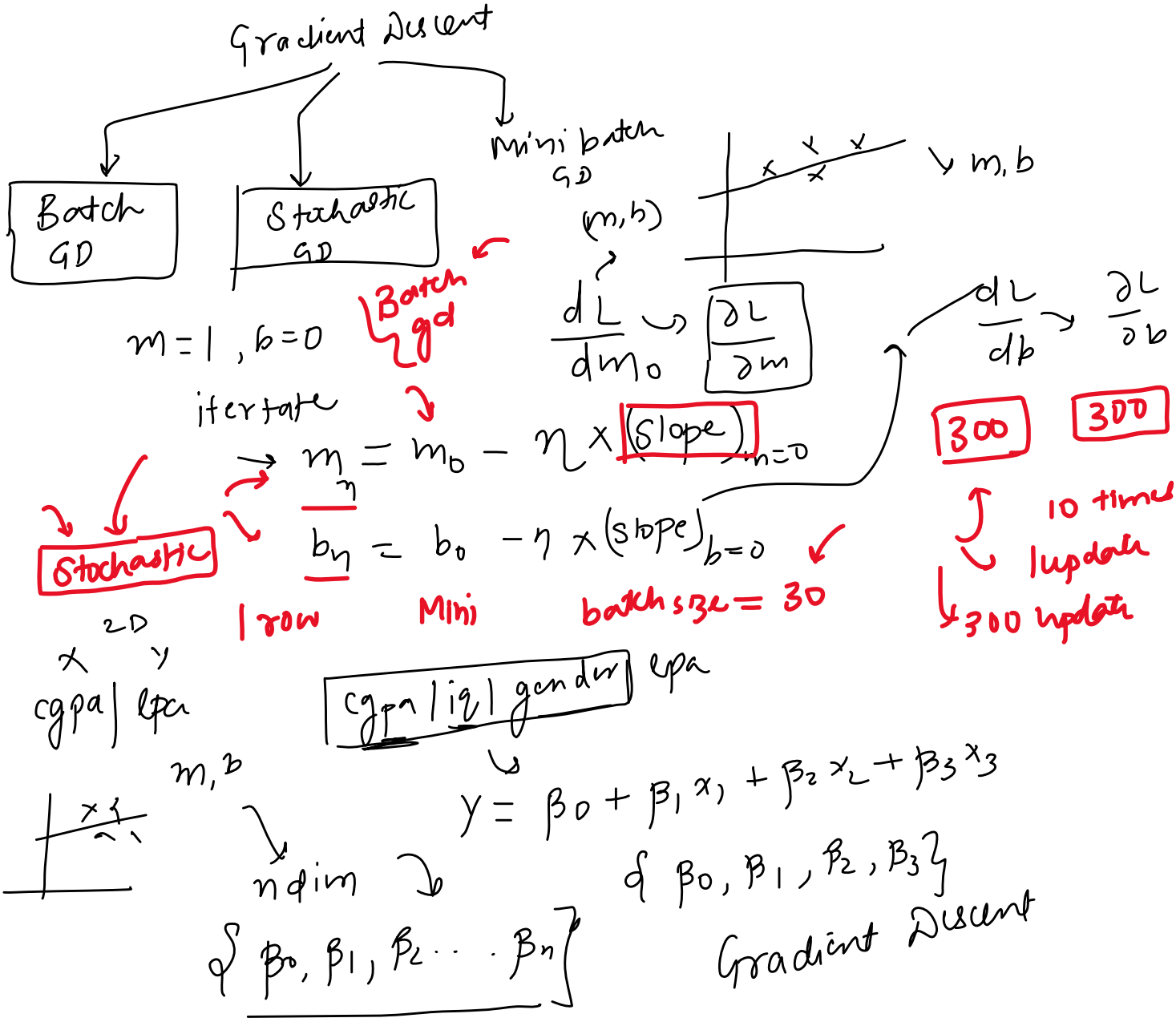
Effect of Data

Thursday, May 20, 2021 1:53 PM



Types of Gradient Descent

Saturday, May 22, 2021 1:30 PM



Batch GD

Saturday, May 22, 2021 3:38 PM

x cgpa | iq | lpa

(2,3) .

n-dim-dataset 3-cols

$\hat{y}_1$ x_{11} | x_n

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ x_{21} | x

(lpa) (cgpa) (iq)

{m, b}

{ $\beta_0, \beta_1, \beta_2$ } L ($\beta_0, \beta_1, \beta_2$)

1) Random values

$\beta_0 = 0$, $\beta_1, \beta_2 = 1$

2) epoch = 100, $\alpha = 0.1$

$\beta_0 = \beta_0 - \eta$ slope

$\beta_1 = \beta_1 - \eta$ slope

$\beta_2 = \beta_2 - \eta$ slope

$\frac{\partial L}{\partial \beta_0}$ $\frac{\partial L}{\partial \beta_1}$ $\frac{\partial L}{\partial \beta_2}$

$\frac{\partial L}{\partial \beta_0}$ β_1 β_n

n-dim

(n+1) $\beta_0 - \beta_n$

MSE

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

$= \frac{1}{2} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2]$

{ row = 2, cols = 2+1}

$\hat{y}_1 = \beta_0 +$

x_1	x_2	y
8.1	93	3.2
7.5	95	3.5

$L = \frac{1}{2} [(y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 +$

$(y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2]$

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12}$

$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22}$

$\frac{\partial L}{\partial \beta_0} = \frac{1}{2} [2(y_1 - \hat{y}_1)(-1) + 2(y_2 - \hat{y}_2)(-1)]$

$\frac{\partial L}{\partial \beta_0} = -\frac{2}{2} [(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2)]$

$= \frac{-2}{n} [(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + (y_3 - \hat{y}_3) + \dots + (y_n - \hat{y}_n)]$

y → array

y_i [...]

$= \frac{-2}{n} \left[(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + \ldots \right]$

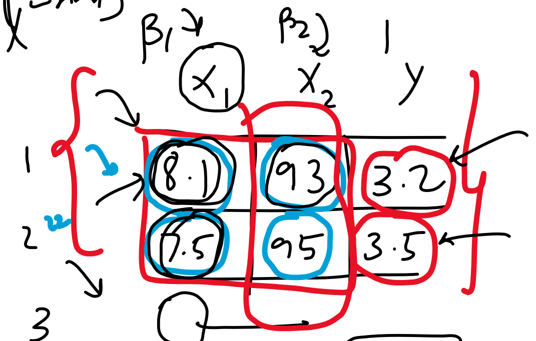
$$= \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) = \frac{\partial L}{\partial \beta_0}$$

$y_i \rightarrow$ [- - -] 353

353 x ← model

X_train

353 → y-hat



$$L = \frac{1}{2} \sum_{i=1}^{2} (y_i - \hat{y}_i)^2 \qquad \frac{\partial}{\partial \beta} \; \frac{1}{n} \sum_{i=1}^{n} x_i \qquad -\beta_1 x_{11} = -x_{11}$$

$$L = \frac{1}{2} \left[(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 \right]$$

$$L = \frac{1}{2} \left[(y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 \right]$$

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{2} \left[2(y_1 - \hat{y}_1)(-x_{11}) + 2(y_2 - \hat{y}_2)(-x_{21}) \right]$$

x_{i1} — 11, 21, 31 . . . n1

$x_{11}, x_{21}, x_{31}, x_{n1}$

$$\frac{\partial L}{\partial \beta_1} = \frac{-2}{n} \left[(y_1 - \hat{y}_1)(x_{11}) + (y_2 - \hat{y}_2)(x_{21}) + (y_3 - \hat{y}_3)(x_{31}) + \ldots + (y_n - \hat{y}_n)(x_{n1}) \right]$$

$$\frac{\partial L}{\partial \beta_1} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{i1}$$

$x_{i1} \rightarrow$ 1 col data

$\beta_1 \rightarrow$ values of 1 col.

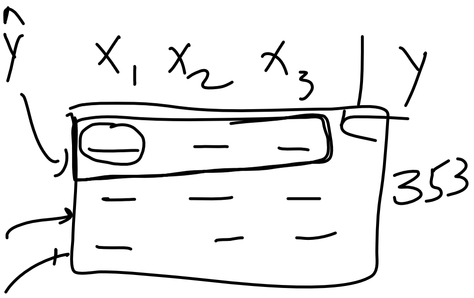
$$\frac{\partial L}{\partial \beta_2} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{i2}$$

m cols

$\beta_0 - \beta_m$

$$\frac{\partial L}{\partial \beta_m} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{im}$$

code



$$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}$$

$$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23}$$

$\beta_0 + np.dot(X\_train, coef)$ $\hat{y} = \beta_0 + [x_{11}\ x_{12}\ x_{13}] \begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}$

(353,10) (10,1)

$\hat{y} = np.dot(coef, X\_train) + \beta_0$

$\beta_0 + (353,1)$

$+ (353,1)$

y_pred

x_1	x_2	y	$\hat{y}$
1	2	5	6
3	4	7	8

$y \quad \hat{y}$

$[5\ \ 7]\ [6\ \ 8]$

$y - \hat{y} = [-1\ \ -1]$

$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,x_{11} = -\frac{2}{n}\left[[-1\ -1]\begin{bmatrix}1\\3\end{bmatrix}\right]$

$[-4 + -4] \rightarrow -8 = 8$

$\frac{\partial L}{\partial \beta_2} = [y - \hat{y}]\begin{bmatrix}2\\4\end{bmatrix}$

$= [-1\ -1]\begin{bmatrix}2\\4\end{bmatrix} \times -\frac{2}{n}$

(1,2) (2,2) (1,2)

$\left[[y - \hat{y}]\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\right] \times -\frac{2}{n}$

$[x\ \ y]$

$\frac{\partial L}{\partial \beta_1} \cdots \frac{\partial L}{\partial \beta_{10}} = \left[(y_1 - \hat{y}_i)\,X\_train\right] \times -\frac{2}{n}$

y_train

$y_i = 353$

353

(353,1) (353,10) (1,353)

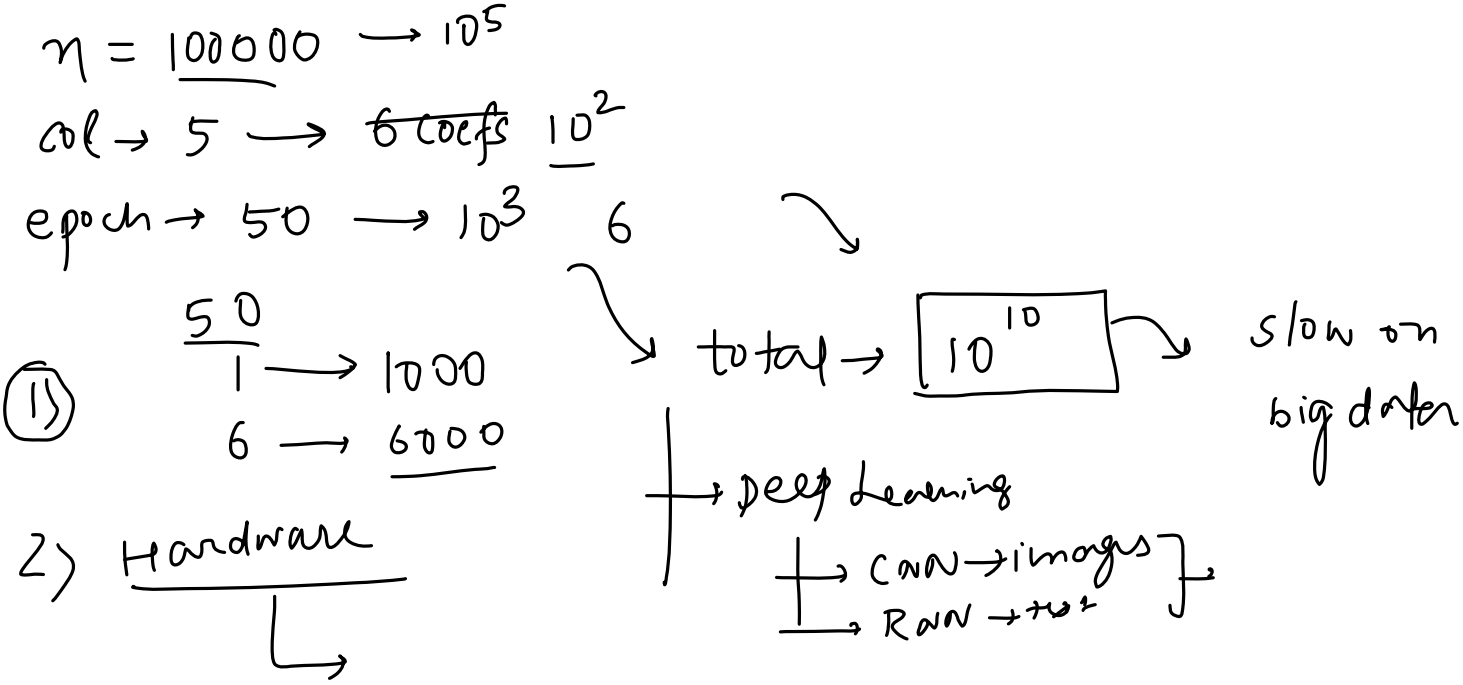
coef_cal

(1,353) (353,10)

$(1,10) \times \frac{-2}{n} = (1,10)$

The Problem with Batch GD

Tuesday, May 25, 2021 6:43 AM



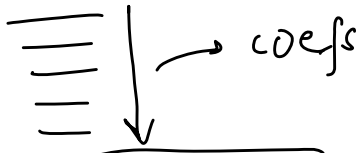
Stochastic GD

Tuesday, May 25, 2021 6:44 AM

5, 10

100 epochs

Batch



→ coefs

Stochastic → —— row → update

n rows

1 epoch

n updates

faster

1 epoch

1 update → random

→ faster convergence

→ Stotastk

steady soln

Code

Tuesday, May 25, 2021 6:44 AM

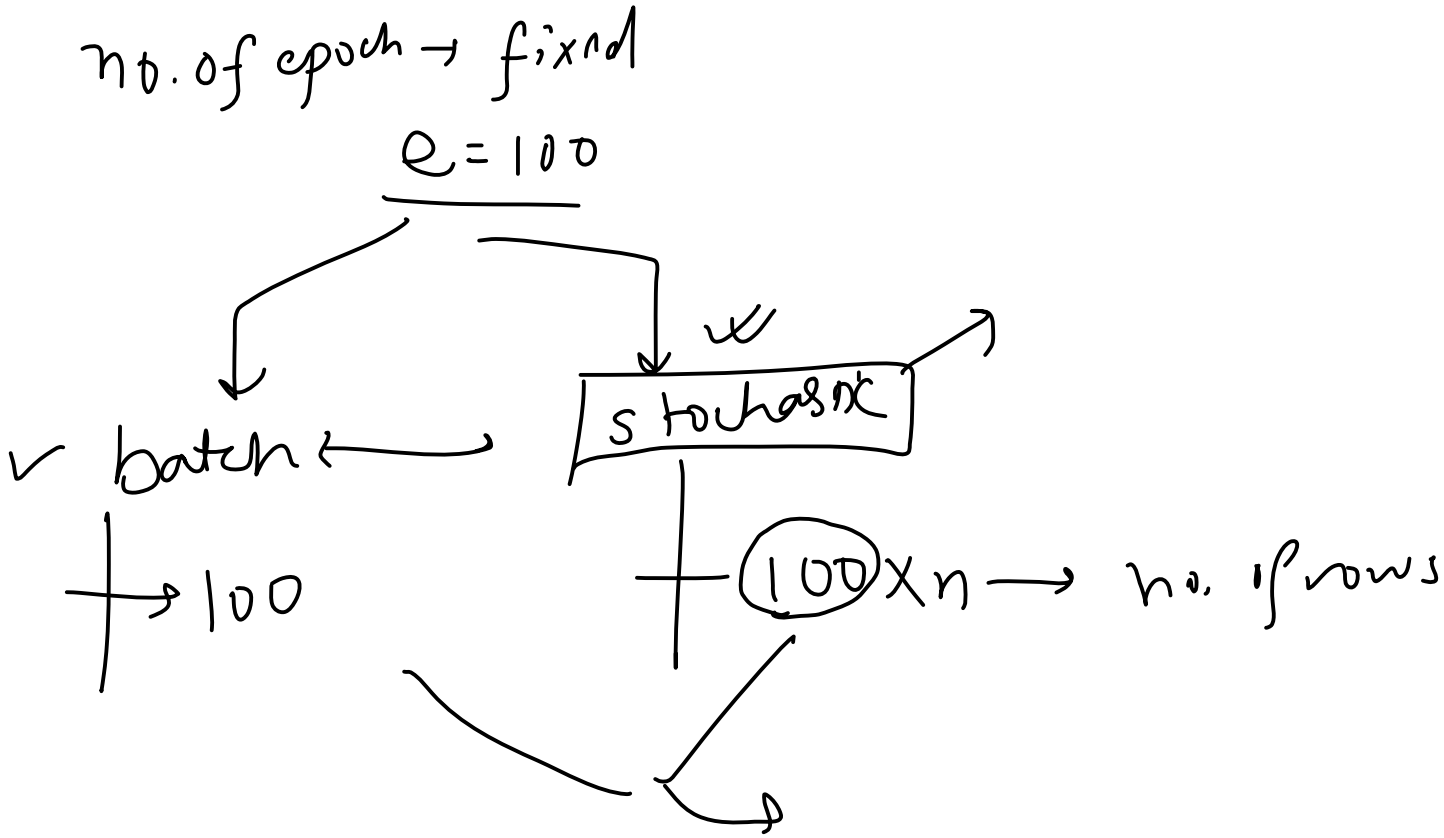
$$\frac{\partial L}{\partial \beta_0} = \frac{-2}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)$$

$i = idx$

$$= -2 \left(\underline{y_i} - \hat{y}_i \right)$$

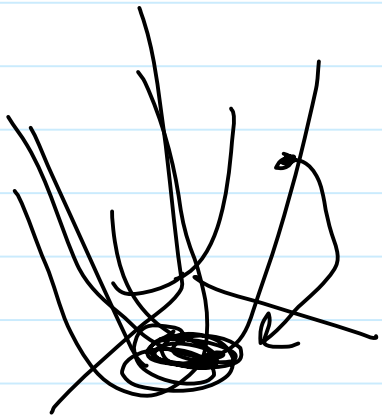
Time Comparison

Tuesday, May 25, 2021 12:39 PM



Visualizations

Tuesday, May 25, 2021 6:44 AM

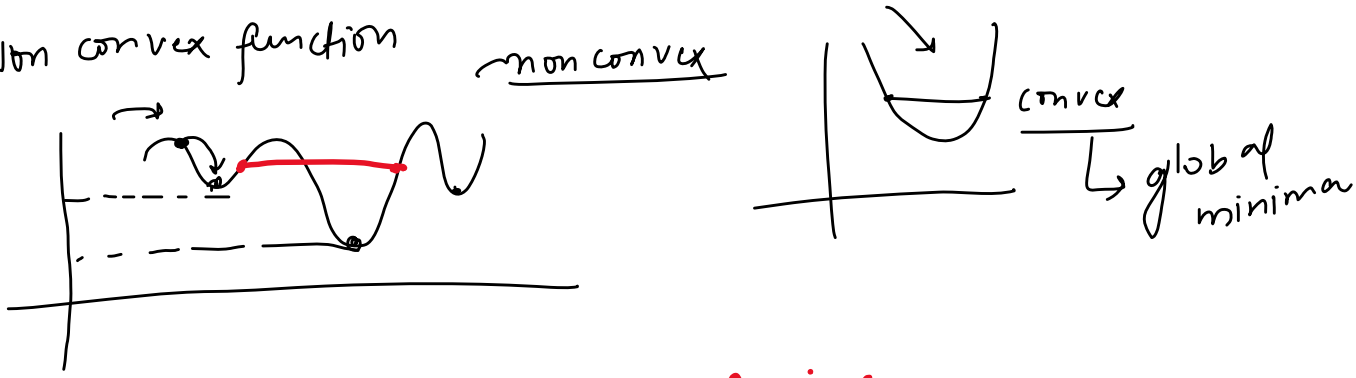


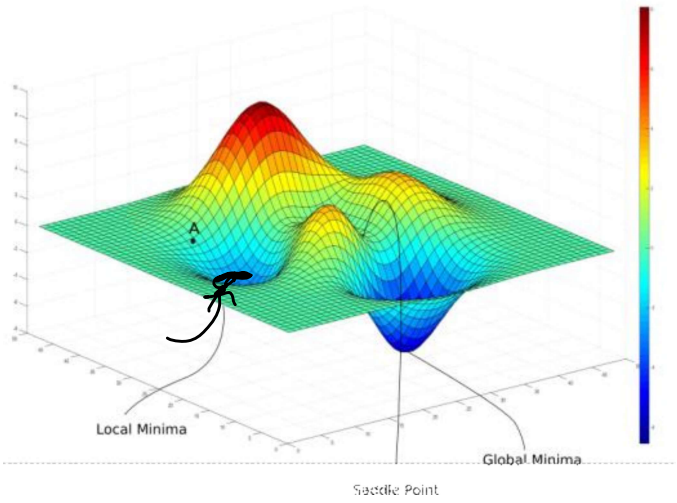
When to use Stochastic GD

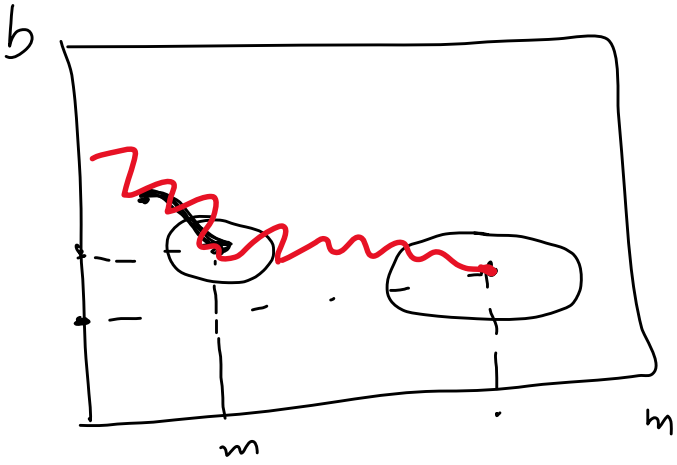
Tuesday, May 25, 2021 6:44 AM

1) Big data → sgd

2) Non convex function

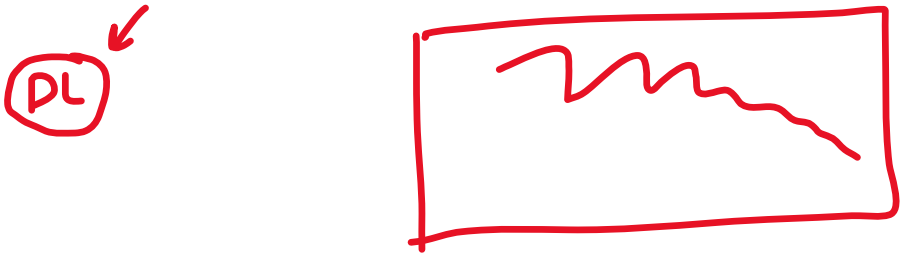


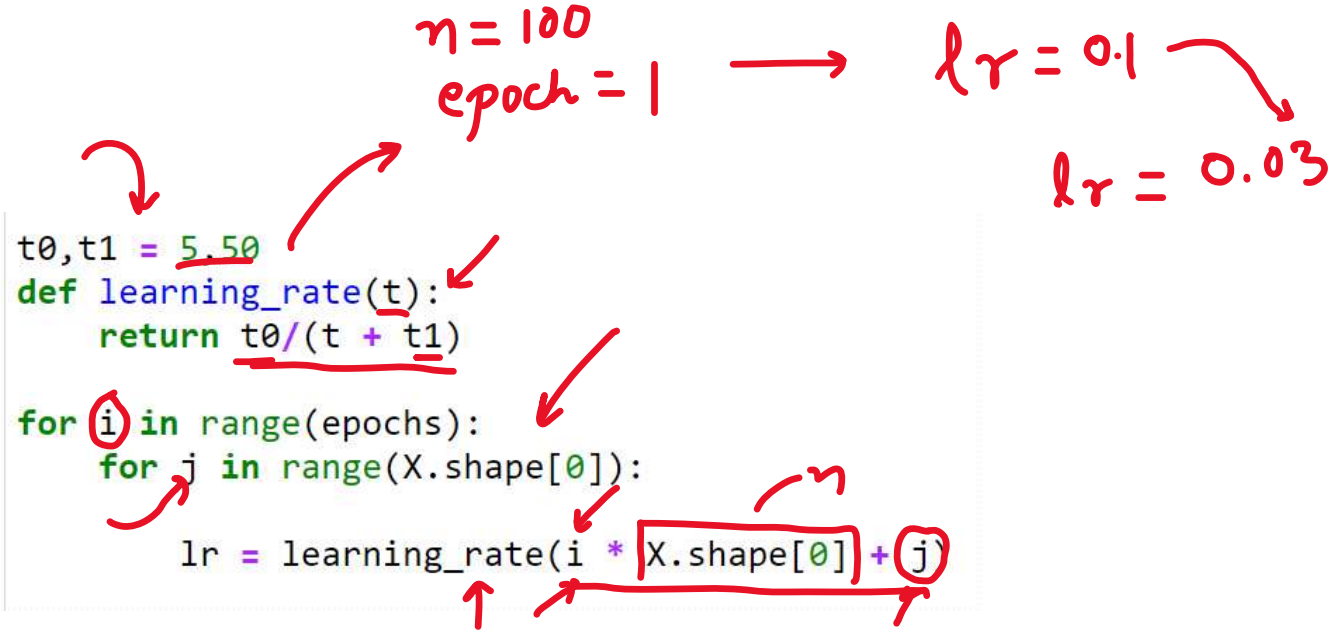




Learning Schedules

Tuesday, May 25, 2021 7:00 AM



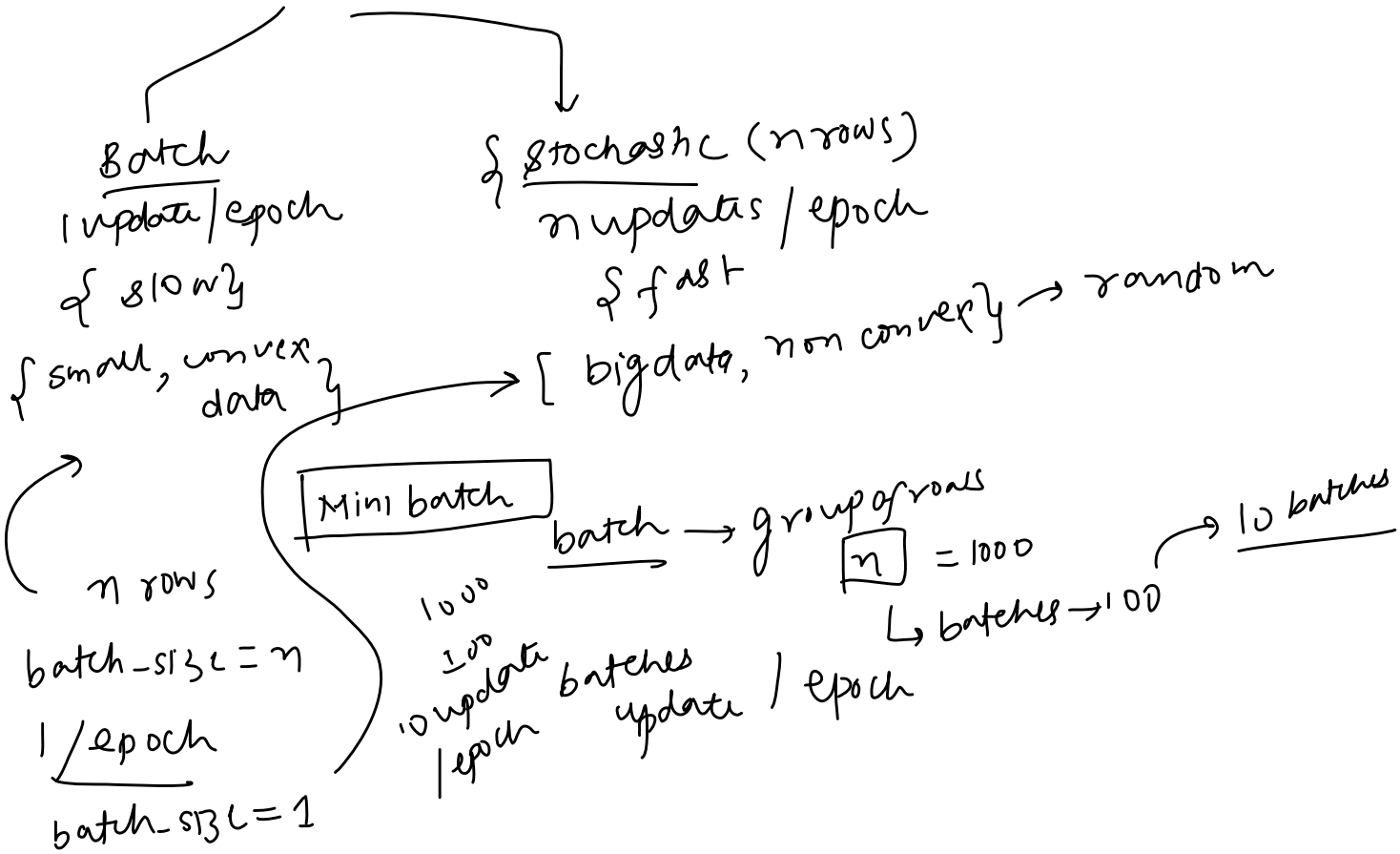


Sklearn Implementation

Tuesday, May 25, 2021 2:16 PM

Mini-Batch Gradient Descent

Wednesday, May 26, 2021 4:47 PM



Code

Wednesday, May 26, 2021 4:47 PM

Visualization

Wednesday, May 26, 2021 4:47 PM

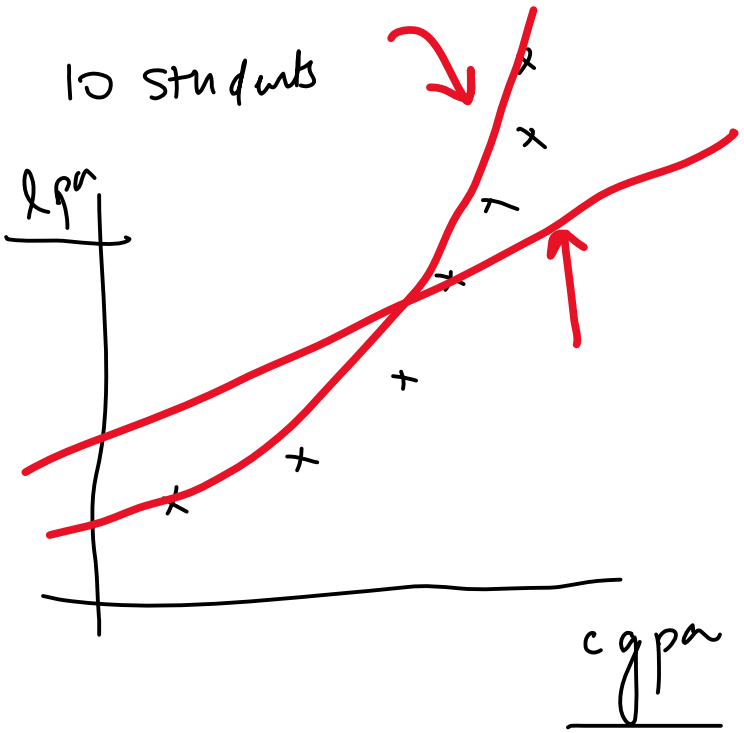
Sklearn Implmentation

Wednesday, May 26, 2021 4:47 PM

Introduction

Friday, May 28, 2021 1:47 PM

cgpa | lpa

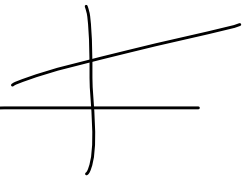


Intuition

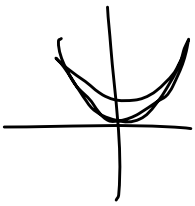
Friday, May 28, 2021 1:47 PM

polynomials

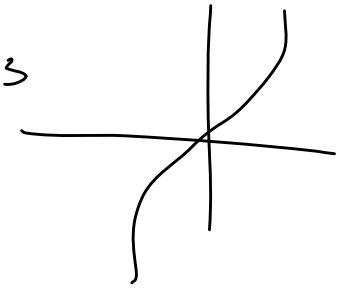
$y = mx + b$



$y = x^2$ →



$y = x^3$



$x^3 + x^2$

$y = ax^4 + bx^3 + cx^2 + dx + z$

degree

1

x	y
2	

3

transform → degree = 2 polynomial

$y = mx + b$ → $y = \beta_0 + \beta_1 x^0 + \beta_2 x^2$

3

x^0	x^1	x^2
1	2	4
1	3	9

Code

Friday, May 28, 2021 1:48 PM

Multiple Polynomial Regression

Friday, May 28, 2021 1:48 PM

Why is it linear

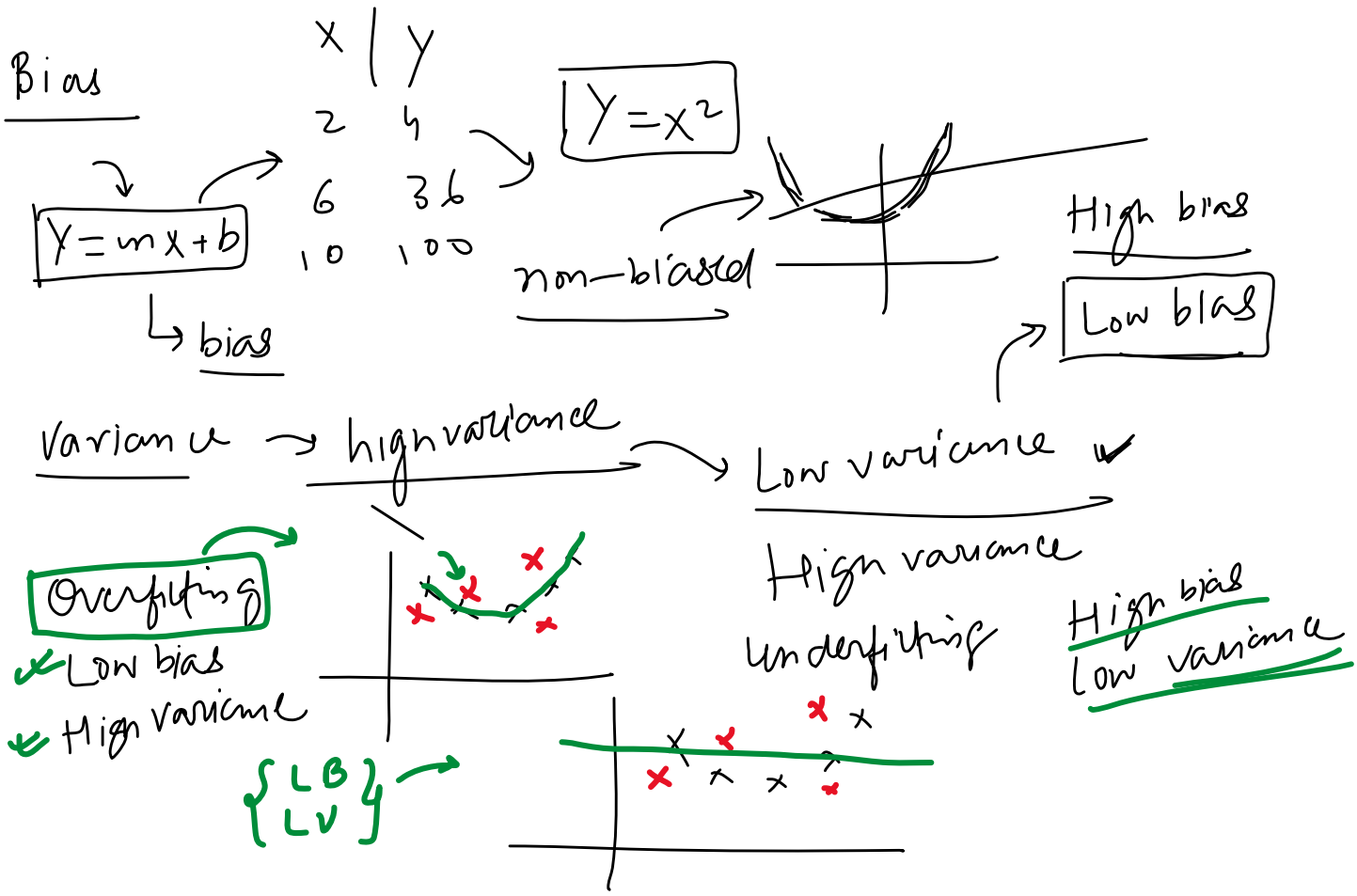
Friday, May 28, 2021 1:48 PM

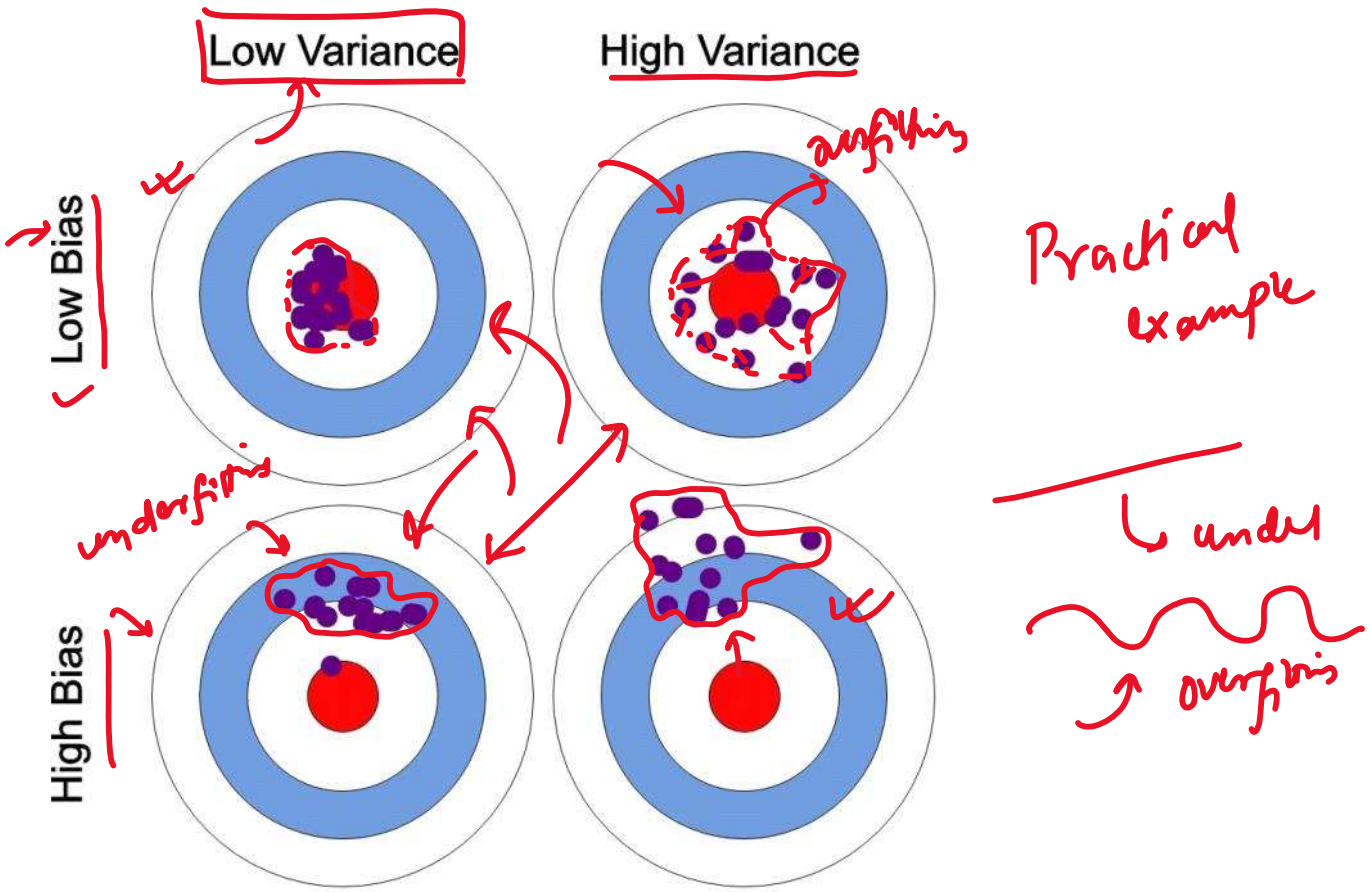
When to use Polynomial Regression

Friday, May 28, 2021 1:48 PM

Bias And Variance

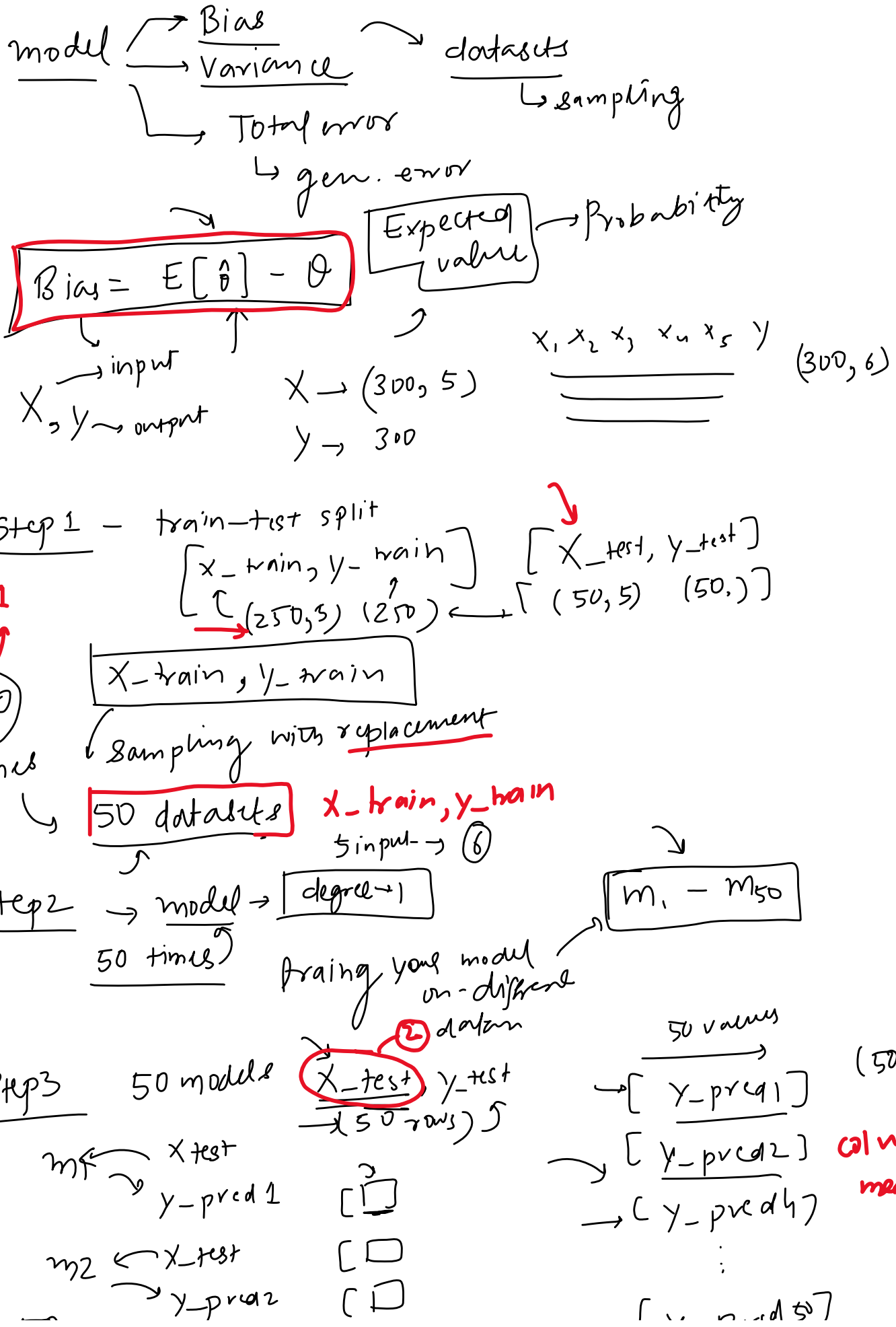
Tuesday, June 1, 2021 12:20 PM

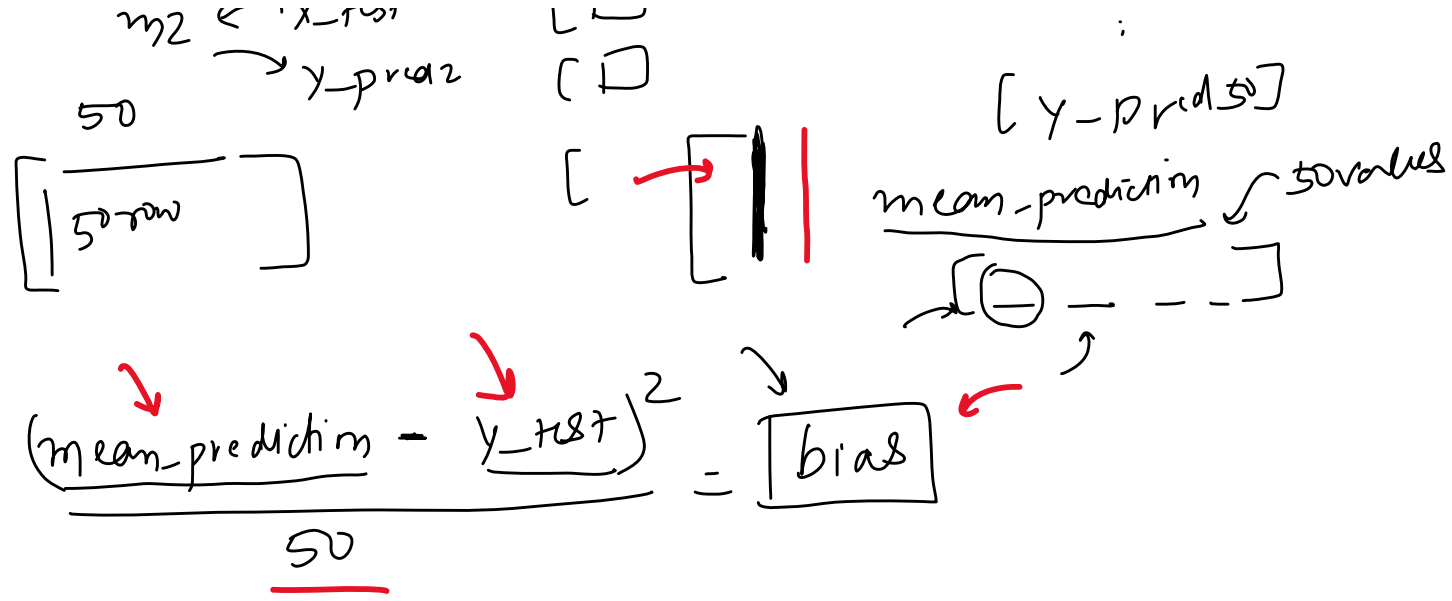




How to calculate Bias and Variance

Tuesday, June 1, 2021 1:44 PM




m2 < x_test
y_pred2
50
50 row
[y_pred 50]
mean_prediction
50 values
(mean_prediction - y_test)^2
50
= bias

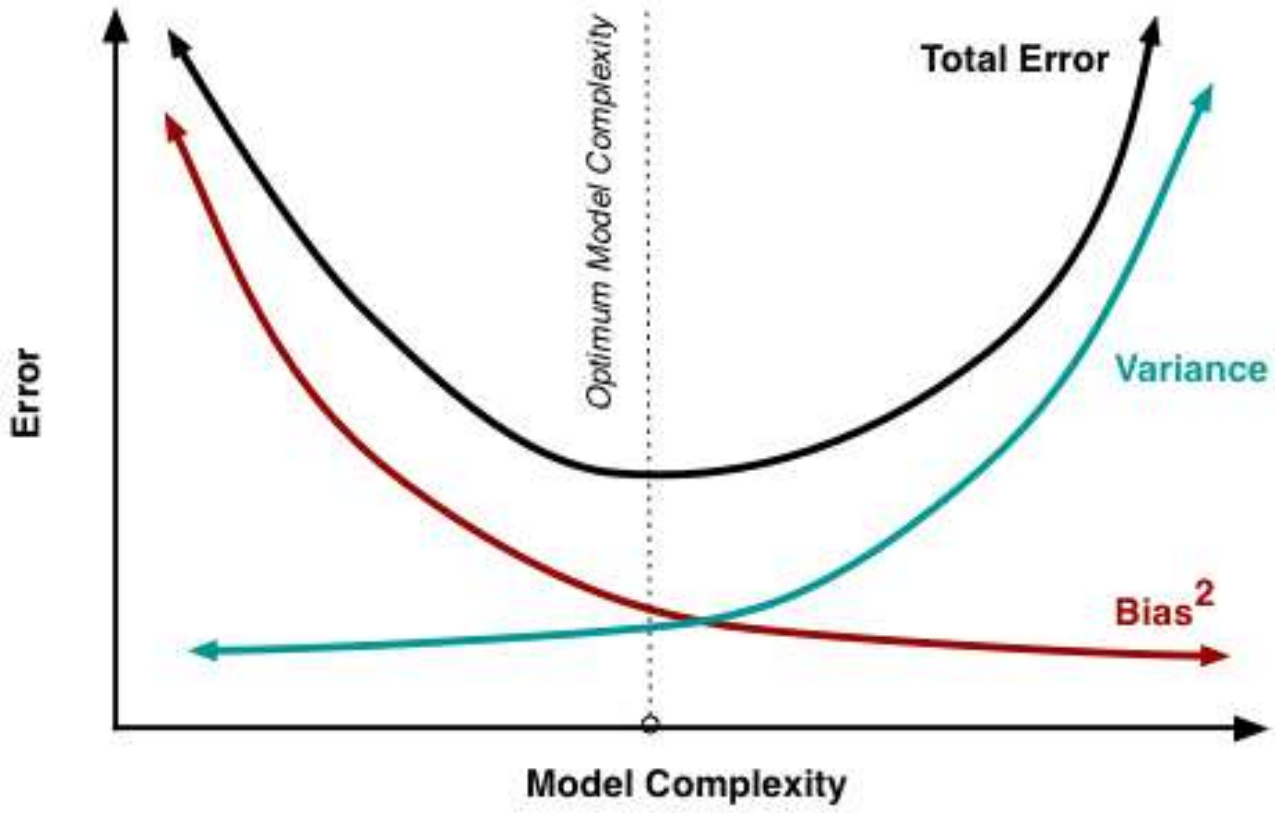
Variance

Code

Monday, May 31, 2021 9:26 AM

Bias Variance Decomposition for Squared Error

Monday, May 31, 2021 9:26 AM

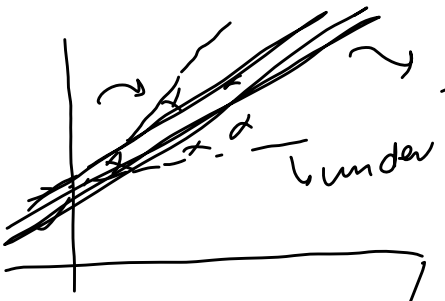


Relationship between Bias and Variance

Monday, May 31, 2021 9:26 AM

Ridge Regression

Wednesday, June 2, 2021 11:13 AM



$y = mx + b$ 2D line

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda m^2$$

m slope

$$L = \sum_{i=1}^{n} (y_i - m x_i - b)^2 + \lambda m^2$$

$\frac{\partial L}{\partial b} = 0 \rightarrow b$

$\frac{\partial L}{\partial m} = 0 \rightarrow m$

$$b = \bar{y} - m\bar{x}$$

$\bar{y} \rightarrow$ y-mean

$\bar{x} \rightarrow$ x-mean

m → slope

$$L = \sum_{i=1}^{n} (y_i - m x_i - \bar{y} + m\bar{x})^2 + \lambda m^2$$

$$\frac{\partial L}{\partial m} = 2 \sum_{i=1}^{n} (y_i - m x_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda m = 0$$

$$= -2 \sum_{i=1}^{n} (y_i - \bar{y} - m x_i + m\bar{x})(x_i - \bar{x}) + 2\lambda m = 0$$

$$= \lambda m - \sum_{i=1}^{n} [(y_i - \bar{y}) - m(x_i - \bar{x})](x_i - \bar{x}) = 0$$

n $(x_i - \bar{x})^2 - n$

$$= \lambda m - \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2 = 0$$

$$= \lambda m - \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + m \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$= \lambda m + m \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$= m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \lambda}$$

λ → hyperparam. alpha

$\lambda = 0 \quad = \lambda = 10,000$

$$m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$m \rightarrow b = \bar{y} - m\bar{x}$

Ridge Regression for 2D data

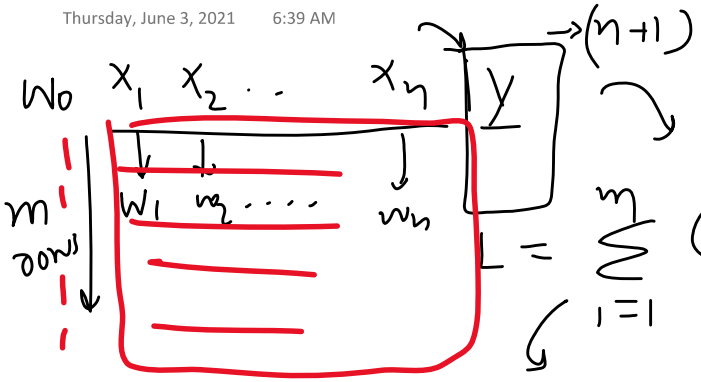
Thursday, June 3, 2021 6:38 AM

Code

Thursday, June 3, 2021 6:39 AM

Ridge Regression for nD data

Thursday, June 3, 2021 6:39 AM



Ridge

$$L = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$= (XW - y)^T (XW - y)$$

m rows

$$y = \begin{bmatrix} \\ - \\ - \\ - \\ - \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}_{(n \times 1)} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ 1 & . & . & & . \\ \vdots & . & . & & . \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Normal LR → Ridge

$$L = (XW - y)^T (XW - y) + \lambda \|W\|^2$$

$\lambda w_0^2 + \lambda w_1^2 + \lambda w_2^2 + \ldots + \lambda w_n^2$

$$[w_0\ w_1\ w_2 \ldots w_n] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad W^T W \rightsquigarrow \lambda(w_0^2 + w_1^2 + w_2^2 \ldots w_n^2)$$

$$(a-b)^T = a^T - b^T \qquad (ab)^T = b^T a^T$$

$$\frac{dL}{dW}$$

$$L = (XW - y)^T (XW - y) + \lambda W^T W$$

$$L = [(XW)^T - (y)^T](XW - y) + \lambda W^T W$$

$$= (W^T X^T - y^T)(XW - y) + \lambda W^T W$$

$$= W^T X^T X W - W^T X^T y - y^T X W + y^T y + \lambda W^T W$$

$$L = w^T X^T X w - 2 w^T X^T y + y^T y + \lambda w^T w$$

$$\frac{dL}{dw} = 2 X^T X w - 2 X^T y + 0 + 2 \lambda w = 0$$

$$X^T X w + \lambda w = X^T y$$

$$(X^T X + \lambda I) w = X^T y$$

3 (4×4)

$(n \times 1, n \times 1)$

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

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

Code

Thursday, June 3, 2021 6:39 AM

Ridge Regression using Gradient Descent

Friday, June 4, 2021 2:17 PM

Vector form loss

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow L = (XW - Y)^T (XW - Y) + \lambda \|W\|^2$$

$$\boxed{L = (XW - Y)^T (XW - Y) + \lambda W^T W}$$

$x_1 x_2 \dots x_n$ (y)

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ 1 & x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \rightarrow (n+1)$$

X: m rows

$w_0, w_1, \dots, w_n$ (parameters)

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}; \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1} \dots \quad w_n = w_n - \eta \frac{\partial L}{\partial w_n}$$

$$w_{new} = w_{old} - \eta \left[\frac{\Delta L}{\Delta W}\right] \rightarrow \text{gradient} \begin{bmatrix} \partial L/\partial w_0 \\ \partial L/\partial w_1 \\ \vdots \\ \partial L/\partial w_n \end{bmatrix}$$

(L, W)

$$L = \frac{1}{2}(XW - Y)^T (XW - Y) + \frac{1}{2}\lambda W^T W$$

$$= \frac{1}{2}(W^T X^T - Y^T)(XW - Y) + \frac{1}{2}\lambda W^T W$$

$$= \frac{1}{2}\left[W^T X^T XW - W^T X^T Y - Y^T XW + Y^T Y \right] + \frac{1}{2}\lambda W^T W$$

$= \frac{1}{2} [$ [illegible] $]$

$= \frac{1}{2} [W^T X^T X W - 2 y^T X W + y^T y] + \frac{1}{2} \lambda W^T W$

$2 W^T X^T y$

$\frac{dL}{dW} = \frac{1}{2} [2 X^T X W - 2 y^T X] + \frac{1}{2} 2 \lambda W$

$X^T y$

$= \boxed{X^T X W - y^T X + \lambda W = \frac{dL}{dW}} \quad \left(\frac{\Delta L}{\Delta W} \right)$

$W = [w_0, w_1, \ldots, w_n]$ starting

$W = [0, 1, \ldots, 1]_{n+1}$

epochs

$W = W - \eta \frac{dL}{dW}$

$W \rightarrow$ final answer

epoch times

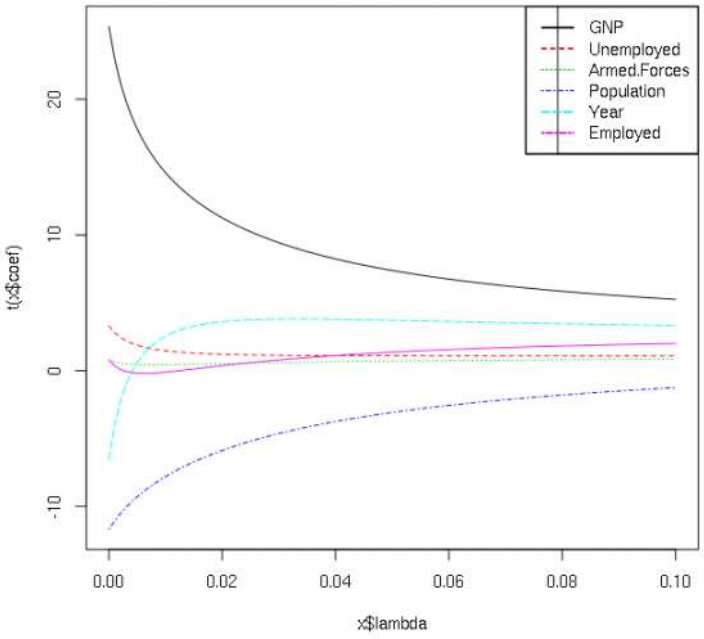
↳ $W = W - \eta \frac{dL}{dW}$

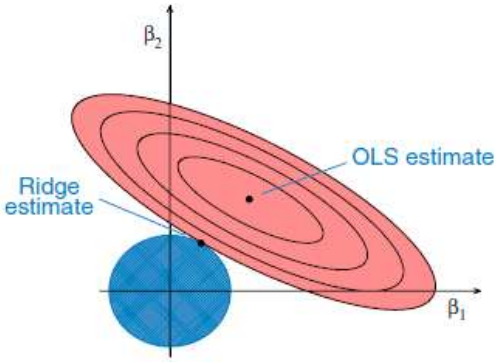
$$\boxed{\frac{dL}{dW} = X^T X W - X^T y + \lambda W}$$

Notes

Friday, June 4, 2021 4:47 PM

Why is it called ridge





5 Key Understandings

Saturday, June 5, 2021 4:20 PM

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \boxed{\lambda \|w\|^2}$$

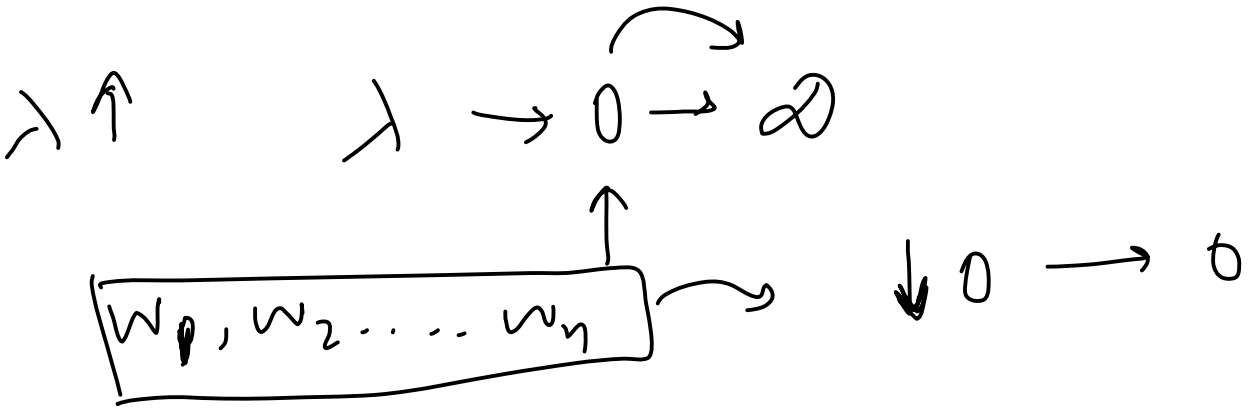
$\lambda(w_1^2 + w_2^2 + \ldots + w_n^2)$

Shrinkage coef

Overfitting ↓

1. How the coefficients get affected?

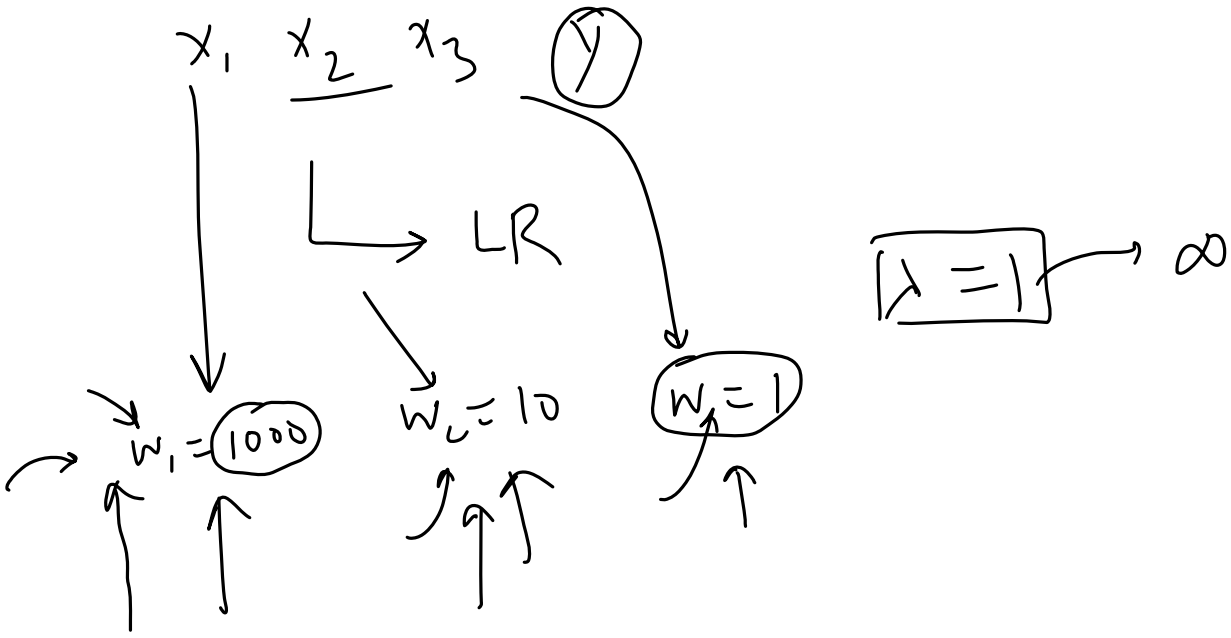
Saturday, June 5, 2021 4:20 PM



2. Higher Values are impacted more

Saturday, June 5, 2021 4:21 PM

Never reaches 0



3. Bias Variance Tradeoff

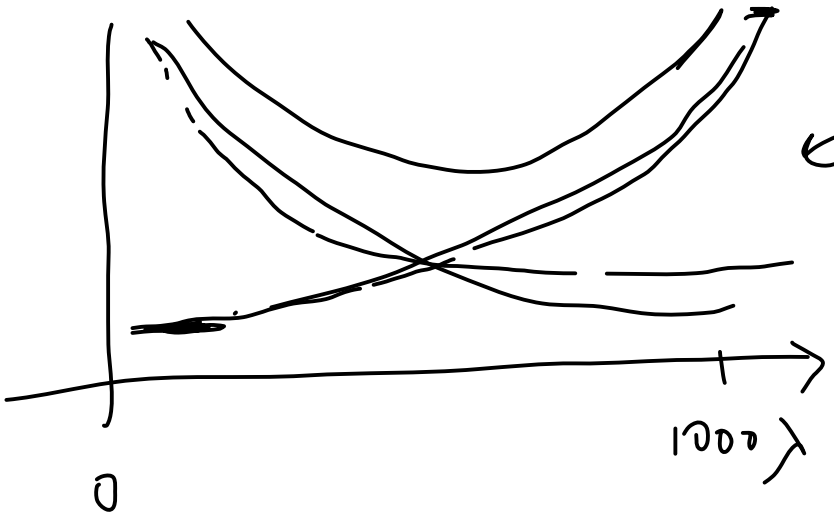
Saturday, June 5, 2021 4:21 PM

Bias Variance

λ ↓ ↑

Bias ↓ overfit Variance ↑

Bias ↑ underfitting Variance ↓



← graph

4. Impact on the Loss Function

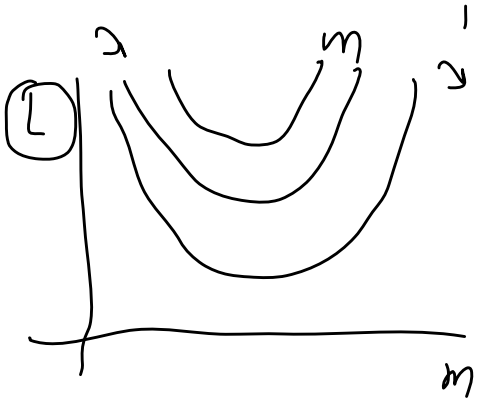
Saturday, June 5, 2021 4:21 PM

$\lambda \rightarrow$ $\rightarrow L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|W\|^2$

$X_1, y \rightarrow [m, b]$ b is constant $b = -2.29$

m $\quad b = 0$

$L = \sum_{i=1}^{n} (y_i - m x_1)^2 + \lambda m^2$

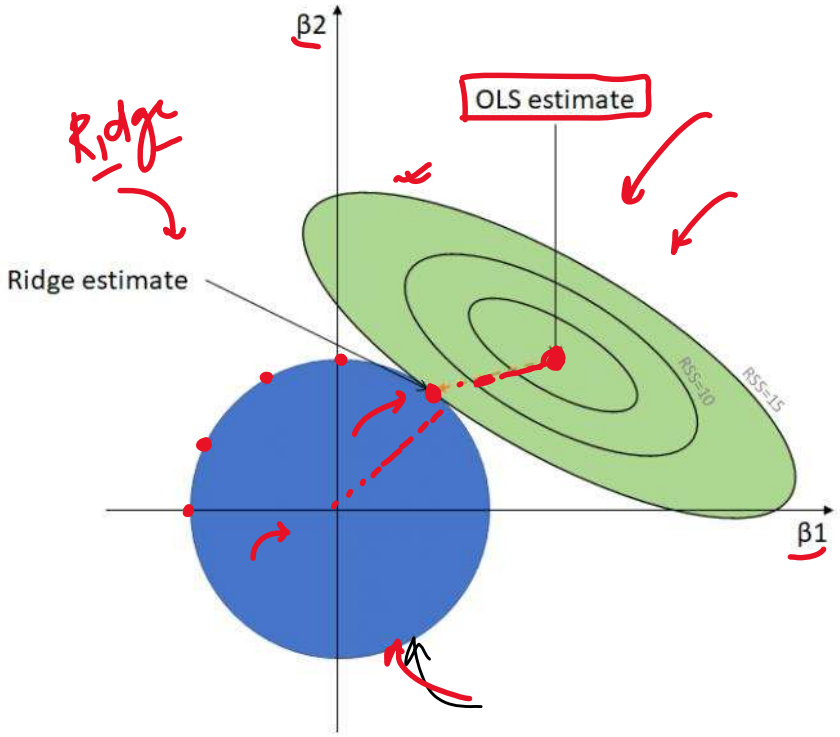


$\lambda = 1$

$\lambda =$

5. Why called Ridge

Saturday, June 5, 2021 4:22 PM



Hard constraint
Ridge constrain

2 coefi β_1 β_2 β_0

$L = MSE + \lambda \|w\|^2$

Contour

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

$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$

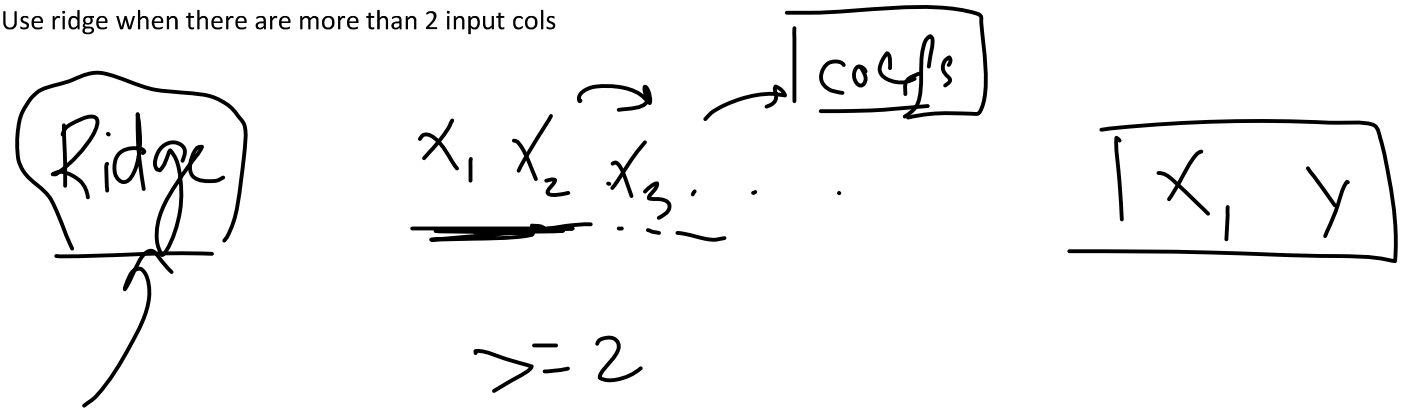
MSE

$\lambda(\beta_1^2 + \beta_2^2)$

Practical Tip

Monday, June 7, 2021 1:20 PM

Use ridge when there are more than 2 input cols



Lasso Regression

Thursday, June 10, 2021 6:42 AM

$y = mx + b$ → m ↓ λ ↑ → $\hat{y} = b$ underfitting

L1 Regularization | overfitting → L2 reg

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$$

$\lambda \geq 0$

$\lambda (w_1^2 + w_2^2 + \dots + w_n^2)$

→ 0 $w_1 \to w_n$ → coeff

overfitting | underfitting

alpha → Lasso → L1 norm $[|w_1| + |w_2| + |w_3| + \dots + |w_n|]$

$L = MSE + \lambda \|w\|$

→ code implement (sklearn)

→ key point →

→ difference Ridge Vs Lasso

underfitting ↑ λ ↑

$\lambda = 0 \to$ Linear

X_1 X_2 X_3 ... X_n Y ← overfitting ← degree

0.5 (0.2)

→ ridge ←

→ lasso → $\lambda \to \uparrow$

dim ↓

0 ← feature selection → lasso

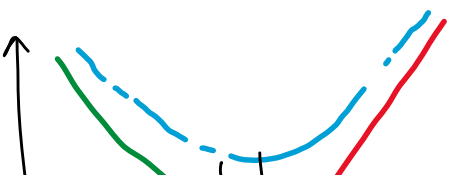
↓ decrease

$\lambda \uparrow$

→ overfitting ↓ Bias ↑

Variance ↑ Bias ↓ Variance ↓

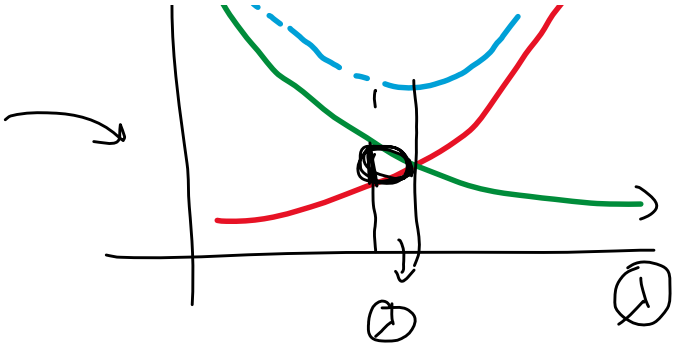
value

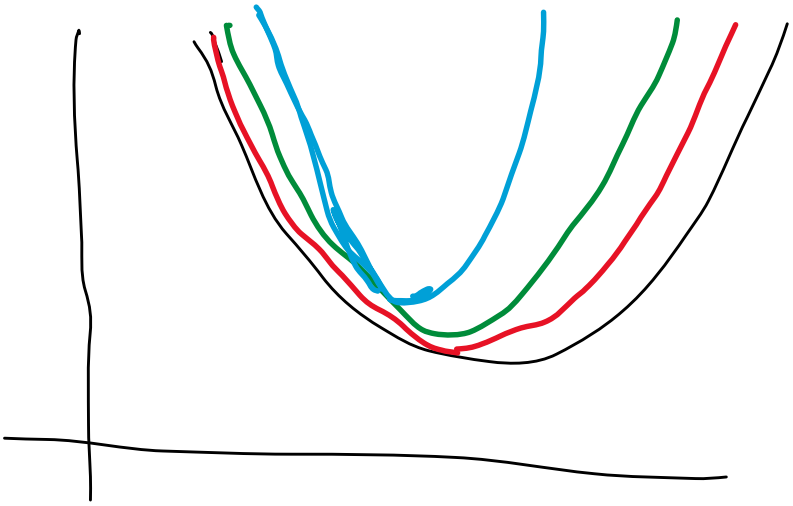


Red → bias

green – variance

blue – total error





Understanding Sparsity

Friday, June 11, 2021 11:21 AM

Lasso sparsity

$\lambda \uparrow \quad w \to 0$

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543609	561.214533	153.884786	126.734316	861.121400	52.419828
0.0001	-9.118336	-205.337133	516.880570	340.556792	-883.415291	551.553259	148.578680	125.355917	856.480254	52.467627
0.0010	-8.763583	-204.321125	518.371729	339.975385	-787.690766	475.274718	106.786540	114.632063	819.739542	52.872100
0.0100	-6.401088	-198.669767	522.048548	336.348363	-383.709187	152.663678	-66.060583	75.611090	659.869402	55.828128
0.1000	6.642753	-172.242166	485.523872	314.682122	-72.939323	-80.590053	-174.466515	83.616653	484.363285	73.584154
1.0000	42.242217	-57.305508	282.170831	198.061386	14.363544	-22.551274	-136.930053	102.023193	260.104308	98.552274
10.0000	21.174004	1.659796	63.659772	48.493240	18.421492	12.875448	-38.915435	38.842464	61.612405	35.505355
100.0000	2.858979	0.629452	7.540604	5.849997	2.710879	2.142134	-4.834047	5.108223	7.448466	4.576129
1000.0000	0.295726	0.069290	0.769004	0.597829	0.282900	0.225936	-0.495607	0.527031	0.761497	0.471029
10000.0000	0.029674	0.006995	0.077054	0.059915	0.028412	0.022715	-0.049686	0.052870	0.076321	0.047241

Ridge

single $x|y \to$

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543596	561.214523	153.884780	126.734314	861.121395	52.419828
0.0001	-9.071288	-205.337332	516.780313	340.539730	-888.652320	555.952271	150.585260	125.453044	858.639860	52.379002
0.0010	-8.264924	-204.213177	517.641106	339.751339	-826.653342	508.609613	120.899583	113.924518	836.314382	52.011583
0.0100	-1.361404	-192.944226	526.348511	332.649058	-430.205495	191.277876	-44.048113	68.990747	688.384976	47.939528
0.1000	0.000000	-113.976046	526.737112	292.635423	-82.691928	-0.000000	-152.691332	0.000000	551.077200	7.169852
1.0000	0.000000	0.000000	363.882636	27.278420	0.000000	0.000000	-0.000000	0.000000	336.135971	0.000000
10.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
100.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
1000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
10000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000

→ feature selection

$$m = \frac{\sum_{i=1}^{n} (y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n} (x_i-\bar{x})^2 + \lambda}$$

simple $x|y \to y = mx+b$

$b = \bar{y} - m\bar{x}$

$\bar{y} \to \text{mean}(y)$

$\bar{x} \to \text{mean}(x)$

$$m = \frac{\sum_{i=1}^{n} (y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n} (x_i-\bar{x})^2}$$

$b = \bar{y} - m\bar{x}$ $\quad m = ?$

$|m| = m$ $\quad m > 0$

$|m| = -m$ $\quad m < 0$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \pm \lambda |m|$$

$$\frac{d}{dm} \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda m$$

$$\frac{dL}{dm} = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda|m| = 2\sum (y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda = 0$$

$$-2\sum [(y_i - \bar{y}) - m(x_i - \bar{x})](x_i - \bar{x}) + 2\lambda = 0$$

$$m\sum (x_i-\bar{x})^2 = \sum (y_i-\bar{y})(x_i-\bar{x}) - \lambda$$

$$m = \sum (y_i-\bar{y})(x_i-\bar{x}) - \lambda$$

$$-\sum [(y_i-\bar{y})(x_i-\bar{x}) - m(x_i-\bar{x})^2] + \lambda = 0$$

$$-\sum (y_i-\bar{y})(x_i-\bar{x}) + m\sum (x_i-\bar{x})^2 + \lambda = 0$$

$$\boxed{m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) - \lambda}{\sum (x_i-\bar{x})^2}}$$

Lasso — coeff sparsity

for $m>0$

$\lambda > 0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) - \lambda}{\sum (x_i-\bar{x})^2}$$

for $m=0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2}$$

for $m<0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) + \lambda}{\sum (x_i-\bar{x})^2}$$

$m \rightarrow$ 2, $\frac{9}{5}$, 1, 0, -1, -5, 0

$$m = \frac{yx - \lambda}{x^2} \quad \begin{cases} yx = 100 \\ x^2 = 50 \end{cases}$$

$\lambda = 100$

$m = 0$

$$m = \frac{yx + \lambda}{x_2} = \frac{100+\lambda}{50}$$

$$= \frac{100+150}{50} =$$

$m = -5$

$$m = \frac{100 - \lambda}{50}$$

$\lambda = 0 \quad m = 2$

$\lambda = 10 \quad m = \frac{9}{5}$

$\lambda = 50 \quad m = 1$

$\lambda > 100 \quad m = -1$

150

150

$m < 0$

$\lambda > 0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) + \lambda}{\sum (x_i-\bar{x})^2}$$

(+ve)

$$m = \frac{-100 - \lambda}{50}$$

$$= \frac{-100-150}{50} = -5$$

$$m = \frac{-100 + \lambda}{50}$$

$\lambda = 0 \quad m = -2$

$\lambda = 50 \quad m = -1$

$\lambda = 100 \quad m = 0 \rightarrow 1$

$\lambda = 150$

$\lambda = 150, \; m = -5$

1) 0 →

2) stop

$\boxed{m = 1}$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2 + \lambda}$$

0.1 → 0 ✗

$\lambda \rightarrow$ Ridge $<<0$ $==0$

Lasso → $\lambda \rightarrow$ numerator

$\lambda = 10000000$ ← Denominator

ElasticNet Regression

Saturday, June 12, 2021 1:04 PM

Ridge $\lambda(w_1^2+w_2^2+..+w_n^2)$ (L2)

$$L = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \lambda\|w\|^2$$

mse → overfitting

$\lambda \uparrow$ $w \downarrow 0$

↳ 100 cols

Lasso $\lambda(|w_1|+|w_2|+..+|w_n|)$ (L1)

$$L = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \lambda\|w\|$$

mse

↳ $\lambda \uparrow$ $w \to 0$ → feature selection

EN Reg

$$L = \sum(y_i-\hat{y}_i)^2 + a\|w\|^2 + b\|w\|$$

{ λ, l1_ratio }

$\lambda = a+b$

l1_ratio $= \frac{a}{a+b}$

$l1 = \frac{a}{\lambda}$

$a = l_1 \times \lambda$

$b = \lambda - a$

$\lambda = 1$ l1_ratio = 0.5

$a = 0.5$ $b = 0.5$

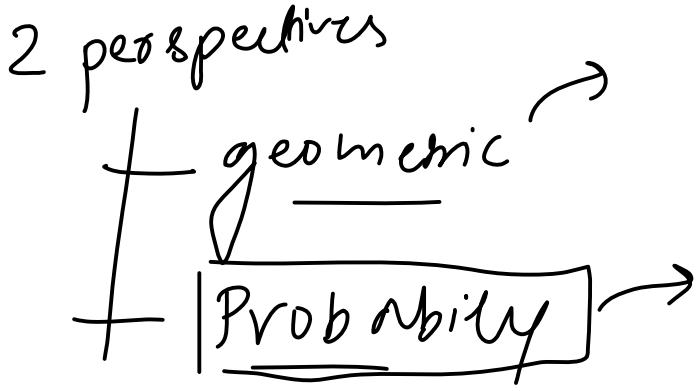
l1_ratio > 0.9 ↙ 90% ridge 10% lasso

Input cols → multicollinearity (ElasticNet)

x_1 | x_2

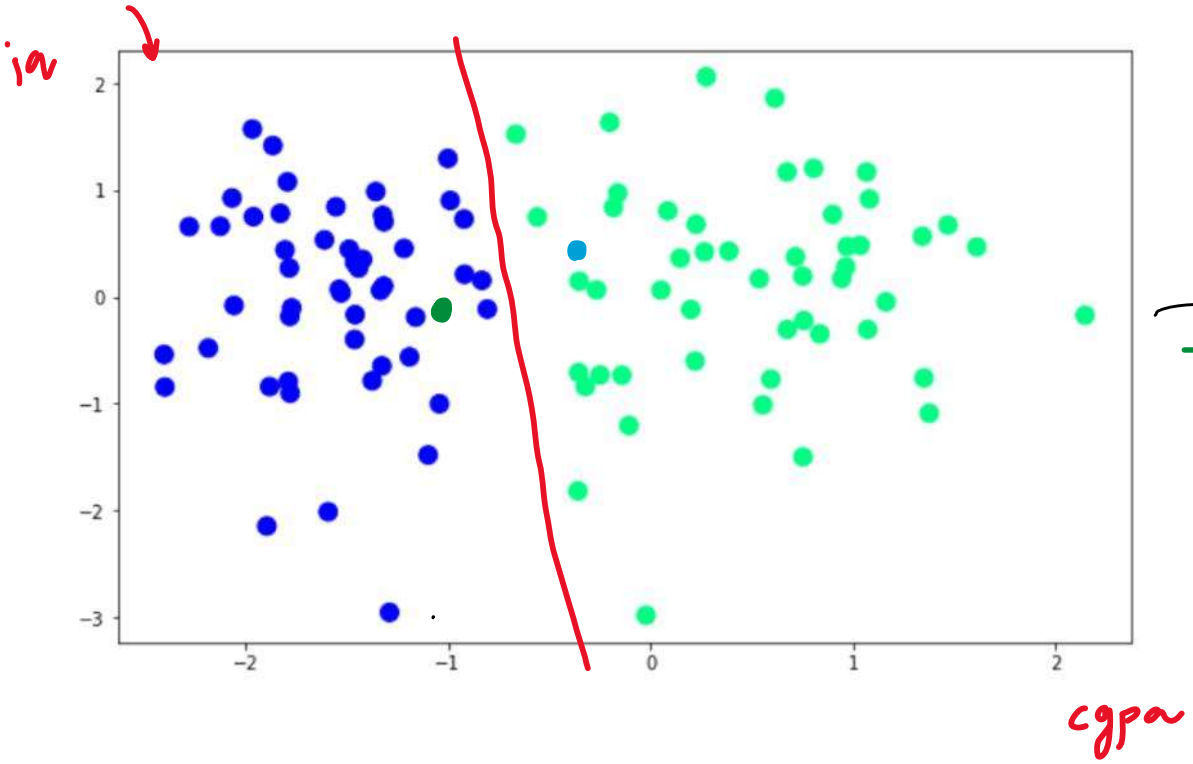
Introduction

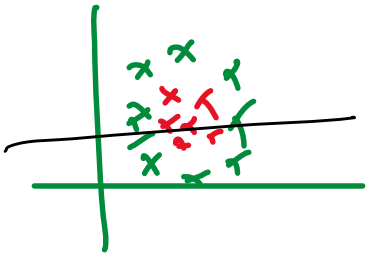
Tuesday, June 15, 2021 12:07 PM



Requirement

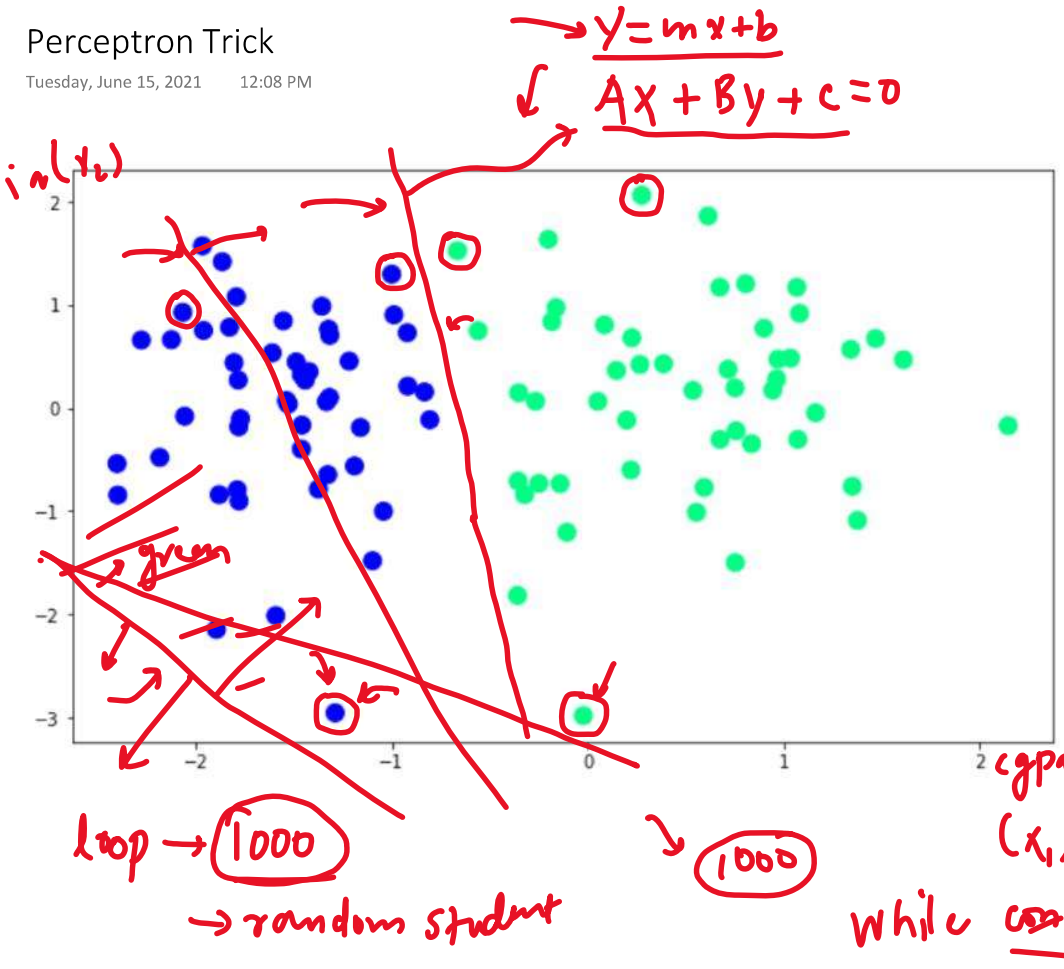
Tuesday, June 15, 2021 12:07 PM





Perceptron Trick

Tuesday, June 15, 2021 12:08 PM



$Ax_1 + Bx_2 + c = 0$

$Ax_1 + Bx_2 + Cx_3 + d = 0$

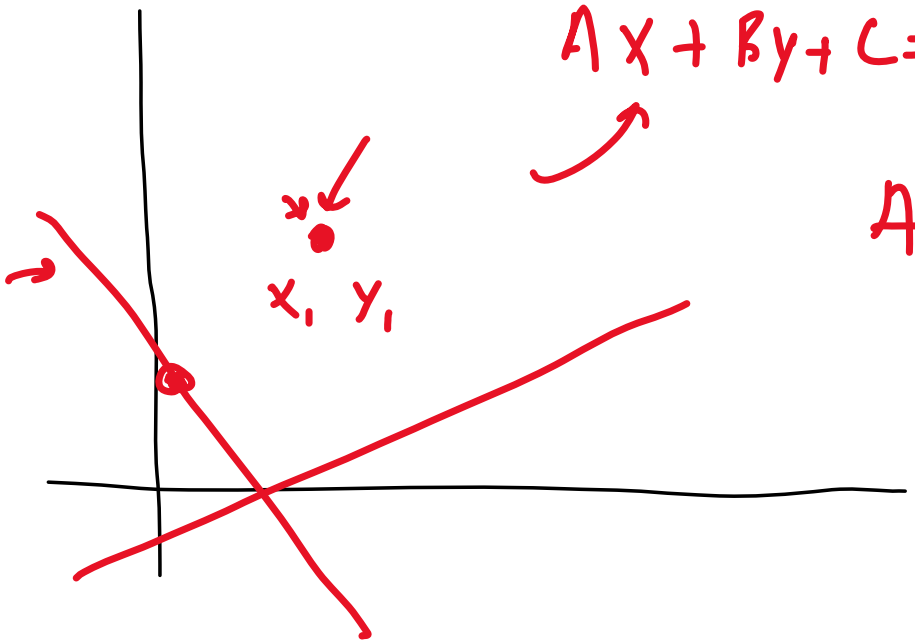
A, B, C

A = 1, B = 1

C = 0

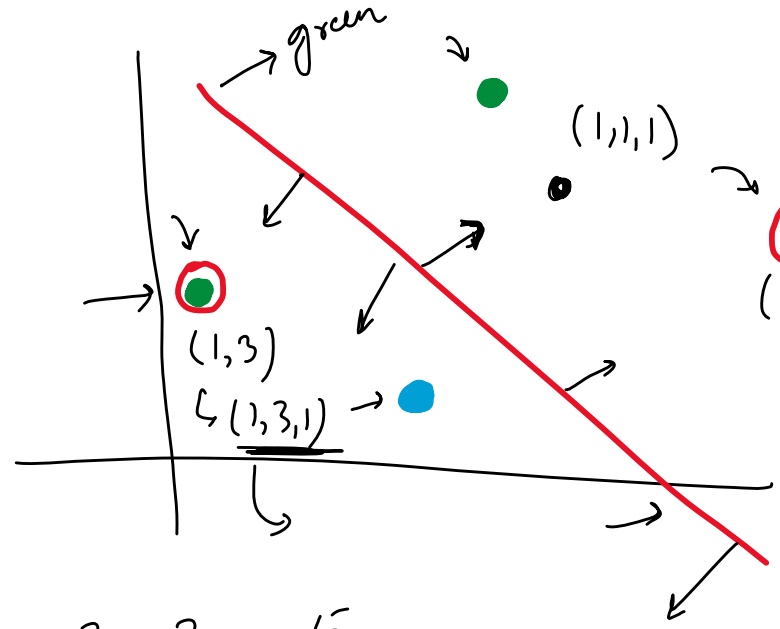
How to label regions?

Tuesday, June 15, 2021 1:12 PM



Transformations

Tuesday, June 15, 2021 1:31 PM



2 3 5

0.01 0.03 0.01

$3x + 6y + 6 = 0$

2 3 5

4 5 1

$-2x - 2y + 4 = 0$

learning rate → 0.01

coef = coef − η coordinate

Algorithm

Tuesday, June 15, 2021 2:31 PM

x_0	x_1 cgpa	x_2 iq	y placed
1	7.5	61	1
1	8.9	109	1
1	7.0	81	0

$Ax + By + C = 0$

$w_0 + w_1 x_1 + w_2 x_2 = 0$

$\sum_{i=0}^{2} w_i x_i = 0$

$w_0 = C, \; w_1 = A, \; w_2 = B$

$w_0 \; w_1 \; w_2$

$w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$

$w_0 \times 1 + w_1 \times 7.5 + w_2 \times 81 \rightarrow \sum_{i=0}^{2} w_i x_i = 0$

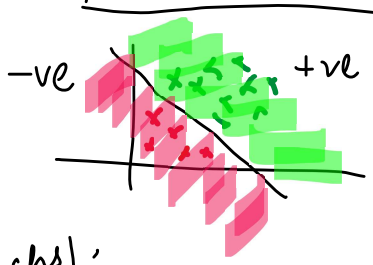
$[w_0 \; w_1 \; w_2] \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$

$\geq 0 \rightarrow 1$

$< 0 \rightarrow 0$

$w_0 x_0 + w_1 x_1 + w_2 x_2 \rightarrow 1$

↳ 0



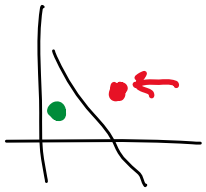
epoch → 1000 ; $\eta = 0.01$

$x_i \in P$

$x_i \in N$

for i in range (epochs):

randomly select a student



$[w_0 \; w_1 \; w_2]$

- if $x_i \in N$ and $\sum_{i=0}^{2} w_i x_i \geq 0$ { (+ve)

  $w_{new} = w_{old} - \eta x_i$

- if $x_i \in P$ and $\sum_{i=1}^{2} w_i x_i < 0$

  $w_{new} = w_{old} + \eta x_i$

$\begin{cases} x_i \in N & w_i x_i < 0 \\ x_i \in P & w_i x_i \geq 0 \end{cases}$

Simplified Algo

Tuesday, June 15, 2021 2:44 PM

if $x_i \in N$ and $\sum w_i x_i \geq 0$

$$w_n = w_0 - \eta x_i$$

if $x_i \in P$ and $\sum w_i x_i < 0$

$$w_n = w_0 + \eta x_i$$

for i in 1000

random student

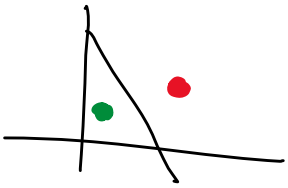
$$w_n = w_0 + \eta (y_i - \hat{y}_i) x_i$$

$$w_n = w_0 \quad (0)$$

$$w_n = w_0 + \eta x_i \quad (1)$$

$$w_n = w_0 - \eta x_i \quad (-1)$$

x_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
1	0	1
0	1	-1



```
for i in range(epochs):
    select a random student (i)
    w_n = w_0 + η (y_i - ŷ_i) x_i
```

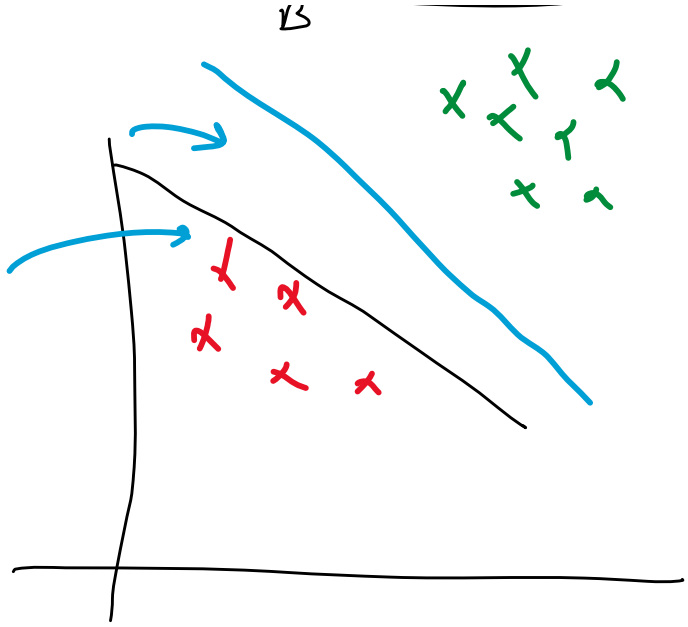
$Ax + By + C = 0$ → A, B, C

$y = mx + b$

$m = -\frac{A}{B}$

$c = -\frac{C}{B}$

coef [_ _]

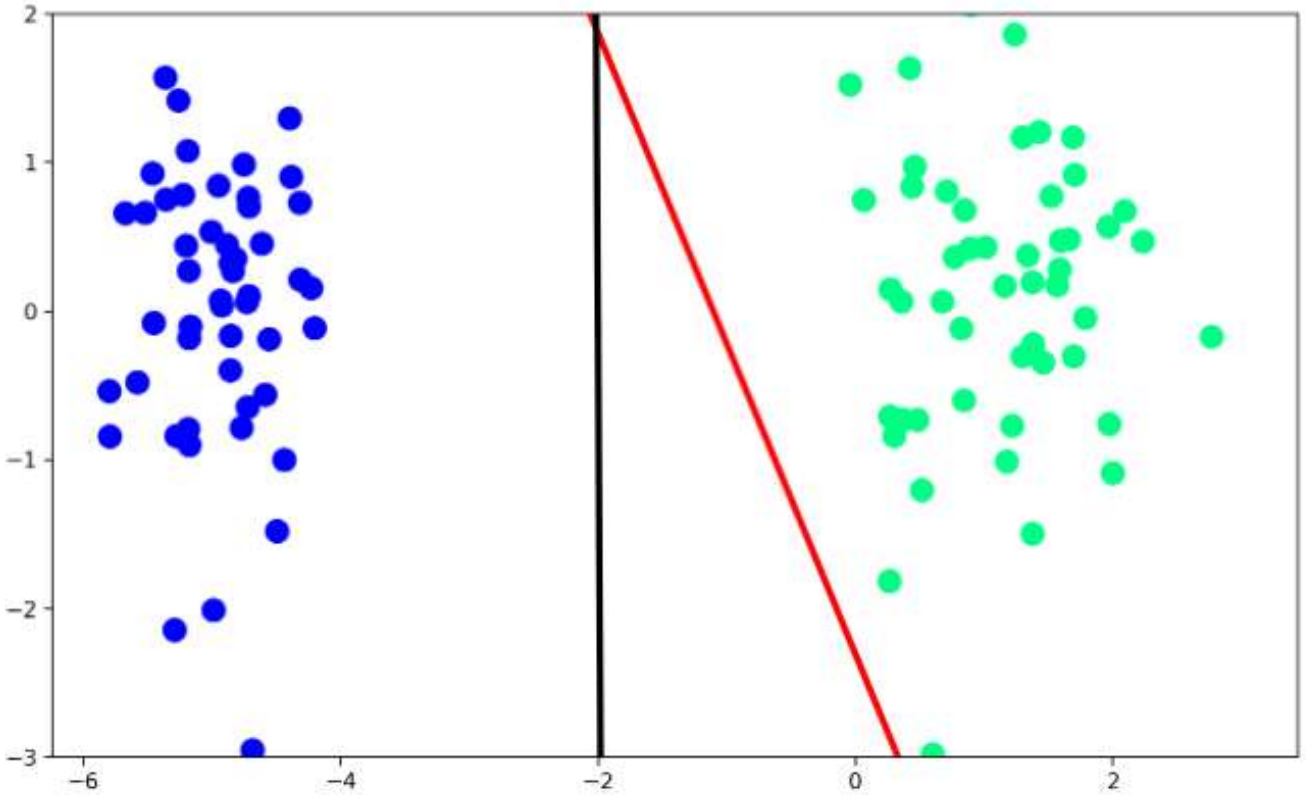


training error

test

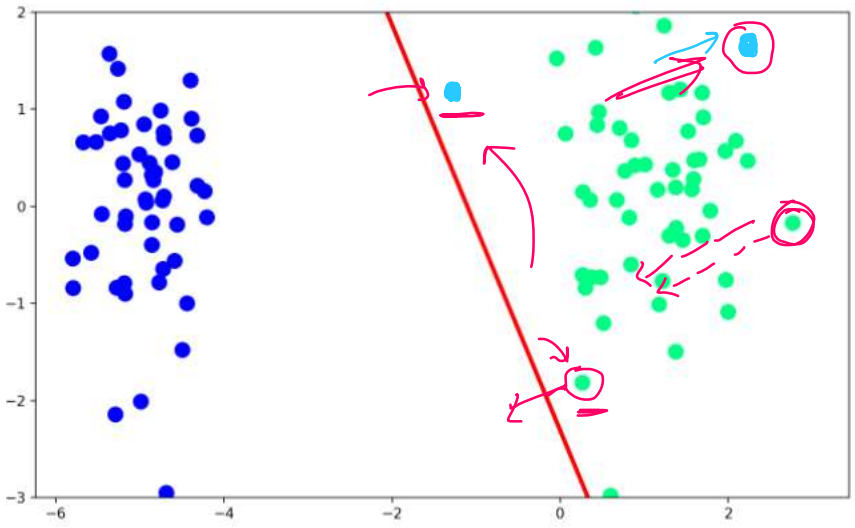
Problem with Perceptron

Thursday, June 17, 2021 12:18 PM



Possible Solution?

Thursday, June 17, 2021 12:18 PM



$$w_{new} = w_0 + (y_i - \hat{y}_i)x_i$$

→ misclassified
line – pull

→ correctly
↳ line push

push pull

$$w_{new} = w_0 + \eta(y_i - \hat{y}_i)x_i$$

w_0 : $(y_i - \hat{y}_i)$: 0

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
1	0	1
0	1	-1

$(y_i - \hat{y}_i) \neq 0$ model prediction

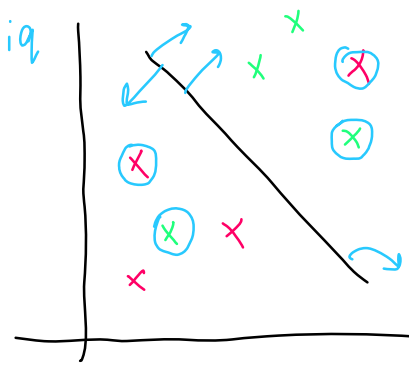
$\Leftarrow w_i x_i = \boxed{0, 1}$

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

$w_1 \times 8.1 + w_2 \times 81 + w_0 =$

≥ 0 (1) < 0 (0)

$\{8.1, 81\}$



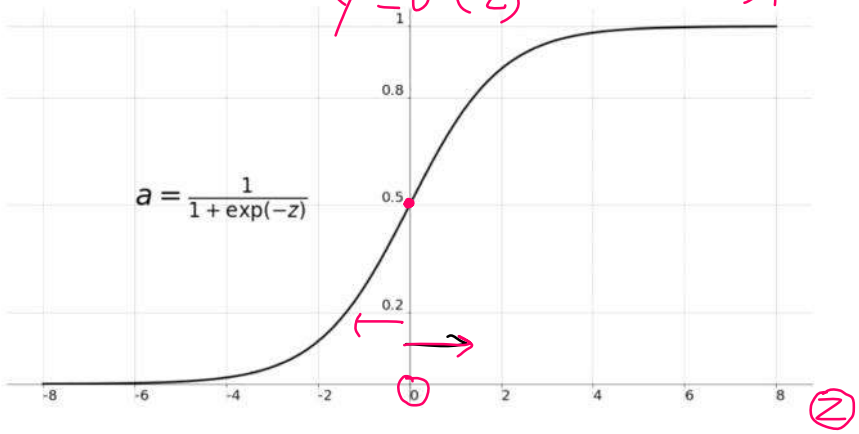
$w_1 x_1 + w_2 x_2 + w_0 = 0$

cgpa	iq	placed (y_i)
9	91	0
8.8	78	1
8.1	102	1
7.9	98	1

The Sigmoid Function

Thursday, June 17, 2021 12:19 PM

Sigmoid Function



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

$y = \sigma(z) \rightarrow 1$

$-\infty < z < \infty$

$0 < y < 1$

$\{7.5, 81\} \rightarrow w_1\ w_2\ w_0$

$\hat{y}_i = w_1 \times 7.5 + w_2 \times 81 + w_0 = z$

$z \geq 0 \rightarrow 1$

$z < 0 \rightarrow 0$

$\sigma > 0.5 \leftarrow z = \sum w_i x_i = \sigma(z)$

$\sigma(z) < 0.5$

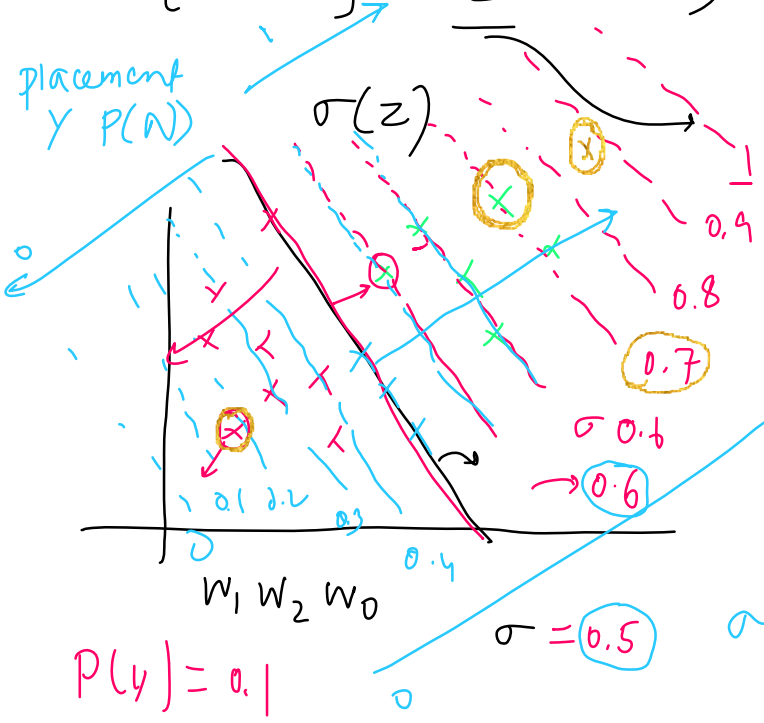
$\{7.5, 81\} \rightarrow z \rightarrow \sigma(z) = 0.5 \geq$

$0.5 <$

if 1 $P(N) = 1 - P(y)$

else 0

$P(y) = 0.7$



$\sum w_i x_i = 0 \quad \sigma = 0.5 \quad P(y) = 0.6$

$\sum w_i x_i > 0 \quad \sigma > 0.5$

$\sum w_i x_i \quad \sigma > 0.6$

$\sigma(\ ) = P(\ )$

$P(y)\ P(N)$

$\sigma = 0.5$

$P(0.5) \rightarrow P(y) = 0.5$

$P(y) = 0.1$

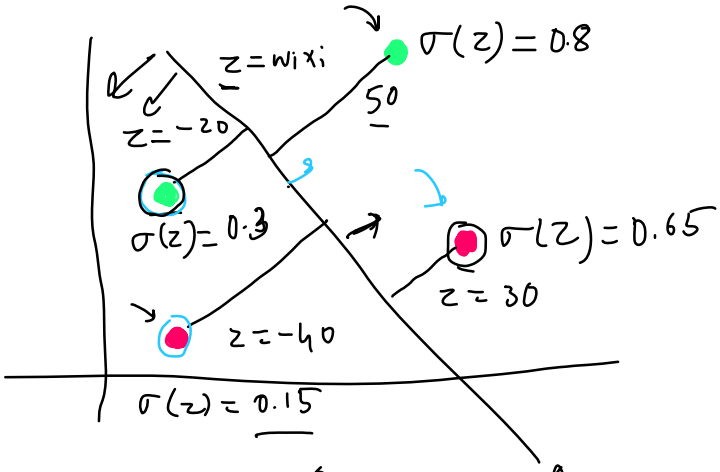
Impact of Sigmoid

Thursday, June 17, 2021 3:27 PM

$$w_\eta = w_0 + \eta (y_i - \hat{y}_i) x_i$$

$$\hat{y}_i = \sigma(z)$$

where $z = \sum w_i x_i$



$$w_\eta = w_0 + \eta \times 0.2 \times x_i$$

$$w_\eta = w_0 - \eta * 0.65 * x_i$$

$$w_\eta = w_0 + \eta * 0.7 * x_i$$

$$w_\eta = w_0 - \eta * 0.15 * x_i$$

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.8	0.2
0	0.65	−0.65
1	0.3	0.7
0	0.15	−0.15

$$w_\eta = w_0 + \eta (y_i - \hat{y}_i) x_i$$

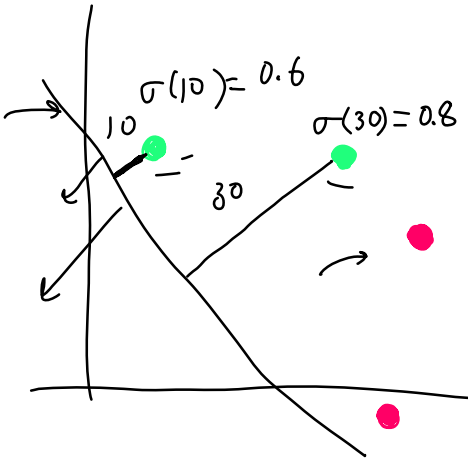
$$\hat{y}_i = \sigma(z)$$

where $z = \sum w_i x_i$

1st

$$w_\eta = w_0 + \boxed{\eta * 0.4 * x_i} = x_1$$

$$w_\eta = w_0 + \boxed{\eta * 0.2 * x_i} = x_2$$

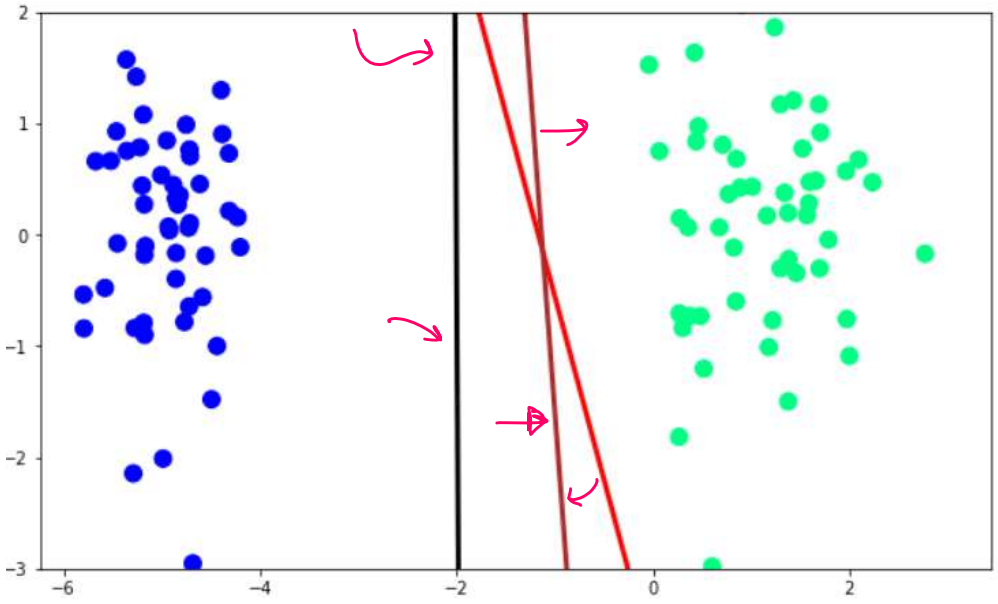


y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.6	0.4
1	0.8	0.2
0		
0		

$x_1 > x_2$ → code implement

The Problem

Friday, June 18, 2021 2:20 PM



red → step

brown → sigmoid

black → sklearn

for i in 1000

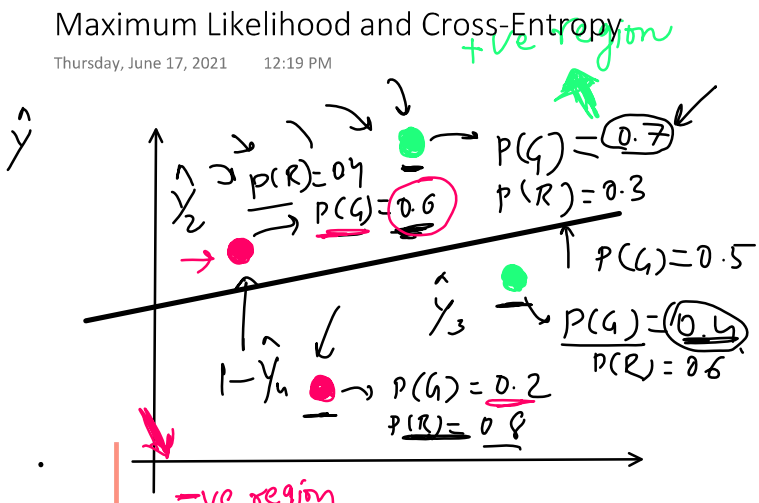
random

Loss function

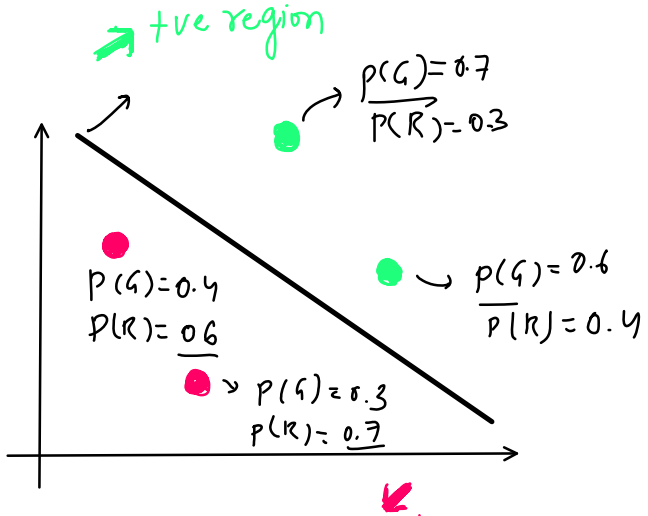
error function

Maximum Likelihood and Cross-Entropy

Thursday, June 17, 2021 12:19 PM



model 1 $\hat{y}_i$



model 2

$\hat{y} = \sigma(z)$

$z = \sum w_i x_i$

model 1 $\rightarrow$ $0.7 \times 0.4 \times 0.4 \times 0.8 = 0.089$

model 2 $\rightarrow$ $0.7 \times 0.6 \times 0.6 \times 0.7$

$= 0.176$

$\log(ab) = \log a + \log b$

$\log(\max) = -\log(0.7) - \log(0.4) - \log(0.4) - \log(0.8)$

0-1 = -ve

cross entropy minimize

$\log(0.1) > \log(0.01)$

$\downarrow$ $\downarrow$

1 0.01

NO

$-\log(\hat{y}_1) - \log(\hat{y}_2) - \log(\hat{y}_3) - \log(\hat{y}_4)$

$(1-\hat{y})$

$y_i = 1$

$-y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)$

$-y_i \log(\hat{y}_i) = -\log(\hat{y}_1)$

$= -\log(0.7)$

$y_2 = 0 \quad y_3 = 1$

$= -\log(0.7)$

$y_2 = 0 \quad y_3 = 1$

$-y_2 \log(\hat{y}_2) - (1-y_2)\log(1-\hat{y}_2)$

$-\log(1-\hat{y}_2) = -\log 0.4$

$-y_3 \log(\hat{y}_3)$

$-\log(\hat{y}_3) = -\log(0.9)$

$y_4 = 0$

$-(1-y_4)\log(1-\hat{y}_4)$

$-\log(1-\hat{y}_4)$

$-\log(0.8)$

$$L = \sum_{i=1}^{n} -y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)$$

MSE

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

closed form

gradient descent

log-loss error

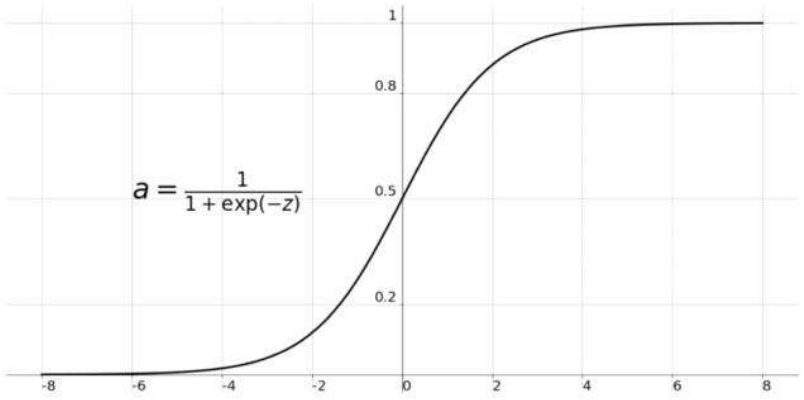
Binary cross entropy

min $w_1 w_2 w_0$

Derivative of Sigmoid

Thursday, June 17, 2021 12:20 PM

Sigmoid Function



$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$\frac{d}{dx}\left(\frac{1}{x}\right) = \frac{d}{dx}(x)^{-1}$$

$$= -x^{-2} = -\frac{1}{x^2}$$

$$\frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right] = -\frac{1}{(1+e^{-x})^2}\frac{d}{dx}(1+e^{-x})$$

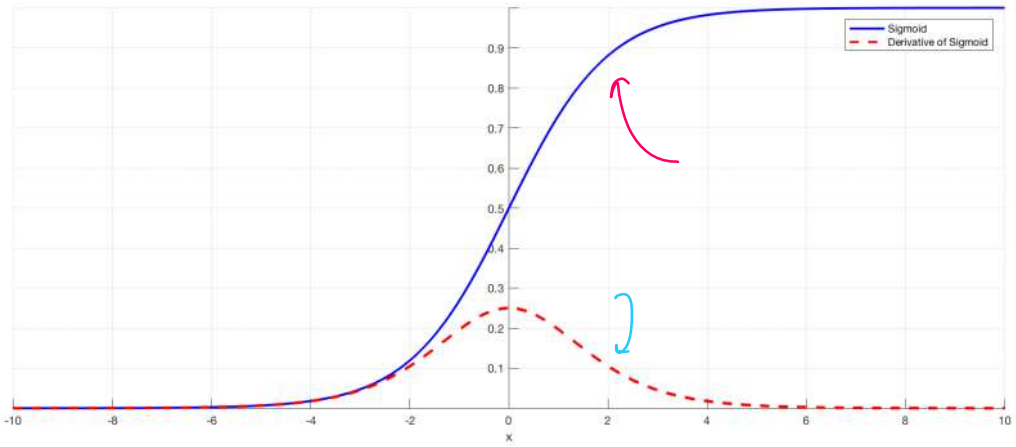
$e^{-x} = e^{-x}$; $-x = -1$

$$= -\frac{1}{(1+e^{-x})^2}\frac{d}{dx}(e^{-x}) = -\frac{e^{-x}}{(1+e^{-x})^2}\frac{d}{dx}(-x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\frac{1\cdot e^{-x}}{(1+e^{-x})(1+e^{-x})} = \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\left[\frac{e^{-x}}{1+e^{-x}}\right]$$

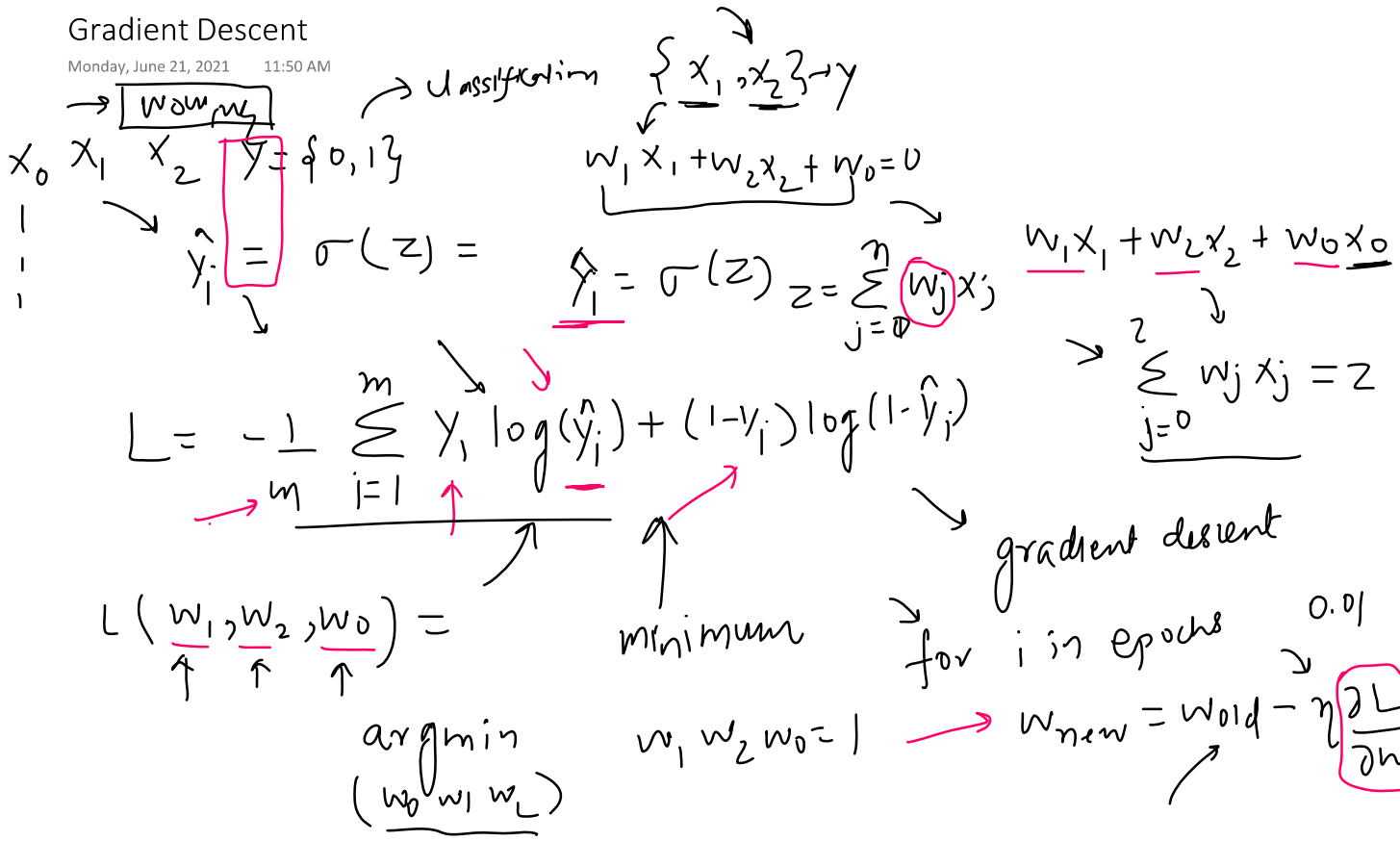
$$= \sigma(x)\left[\frac{1+e^{-x}-1}{1+e^{-x}}\right] = \sigma(x)\left[\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right]$$

$$\sigma(x)\left[1-\sigma(x)\right] \Rightarrow \sigma'(x) = \sigma(x)\left[1-\sigma(x)\right]$$


Sigmoid
Derivative of Sigmoid
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-10
-8
-6
-4
-2
0
2
4
6
8
10
x

Gradient Descent

Monday, June 21, 2021 11:50 AM



$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}, \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1}, \quad w_2 = w_2 - \eta \frac{\partial L}{\partial w_2}$$

n cols → n+1 derivatives

$$w_n = w_n - \eta \frac{\partial L}{\partial w_n} \qquad j = 0, 1, 2, \ldots, n$$

$$L = -\frac{1}{m} \sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{\partial L}{\partial w_j} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right) \right]$$

$\hat{y}_i = \sigma(z)$ $z = \sum_{j=0}^{n} w_j x_j$

$$\frac{\partial L}{\partial w_j} y_i \log(\hat{y}_i) = \frac{\partial L}{\partial w_j} y_i \log(\sigma(z))$$

$$- y_i \frac{\partial L}{\partial} \sigma\left(\sum^{n} w_j x_j\right)$$

$$= \frac{y_i}{\hat{y}_i} \frac{\partial L}{\partial w_j} \sigma\left(\sum_{j=0}^{n} w_j x_j\right)$$

$$\frac{\partial L}{\partial w_j} \; y_i \log\left(\sigma\left(\sum_{j=0}^{n} w_j x_j\right)\right) \qquad \hat{y}_i$$

$$= y_i \frac{\partial L}{\partial w_j} \log\left(\sigma\left(\sum_{j=0}^{n} w_j x_j\right)\right) = \frac{y_i}{\hat{y}_i} \hat{y}_i(1-\hat{y}_i) \frac{\partial L}{\partial w_j} \sum_{j=0}^{n} w_j x_j$$

$$y_i(1-\hat{y}_i) \sum_{j=0}^{n} \frac{\partial L}{\partial w_j} w_j x_j = \boxed{y_i(1-\hat{y}_i) \sum_{j=0}^{n} x_j}$$

$$\frac{\partial L}{\partial w_j} \; (1-y_i)\log(1-\hat{y}_i) \Rightarrow (1-y_i) \frac{\partial L}{\partial w_j} \log(1-\hat{y}_i) \qquad \hat{y}_i = \sigma(z) \quad \sigma\left(\sum_{j=0}^{n} w_j x_j\right)$$

$$\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} (1-\hat{y}_i) \Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \hat{y}_i \Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \sigma(z)$$

$$\Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \sigma(z)(1-\sigma(z)) \frac{\partial L}{\partial w_j} \sigma(z) = -\frac{(1-y_i)}{(1-\hat{y}_i)} \hat{y}_i(1-\hat{y}_i) \frac{\partial L}{\partial w_j} \sigma(z)$$

$$\Rightarrow -\hat{y}_i(1-y_i) \frac{\partial L}{\partial w_j} \sum_{j=0}^{n} w_j x_j \Rightarrow -\hat{y}_i(1-y_i) \sum_{j=0}^{n} \frac{\partial L}{\partial w_j} (w_j x_j)$$

$$\Rightarrow \boxed{-\hat{y}_i(1-y_i) \sum_{j=0}^{n} x_j}$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i(1-\hat{y}_i) \sum_{j=0}^{n} x_j - \hat{y}_i(1-y_i) \sum_{j=0}^{n} x_j \right]$$

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{m} \left[\left(y_i - \hat{y}_i\right) \sum_{j=0}^{n} x_j \right]$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i(1-\hat{y}_i) - \hat{y}_i(1-y_i) \right]\sum_{j=0}^{n} x_j$$

$$= -\frac{1}{m}\sum_{j=1}^{m}\left[y_i - y_i\hat{y}_i - \hat{y}_i + y_i\hat{y}_i \right]\sum_{j=0}^{n} x_j$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)\sum_{j=0}^{n} x_j$$

x_1	x_2	y_i	$\hat{y}_i$
1	2	1	0
3	4	0	0

no. of rows $= m = 2$

no. of cols $= n = 2$

$$\frac{\partial L}{\partial w_0} = -\frac{1}{m}\left[1+0\right]\left[1+1\right]$$

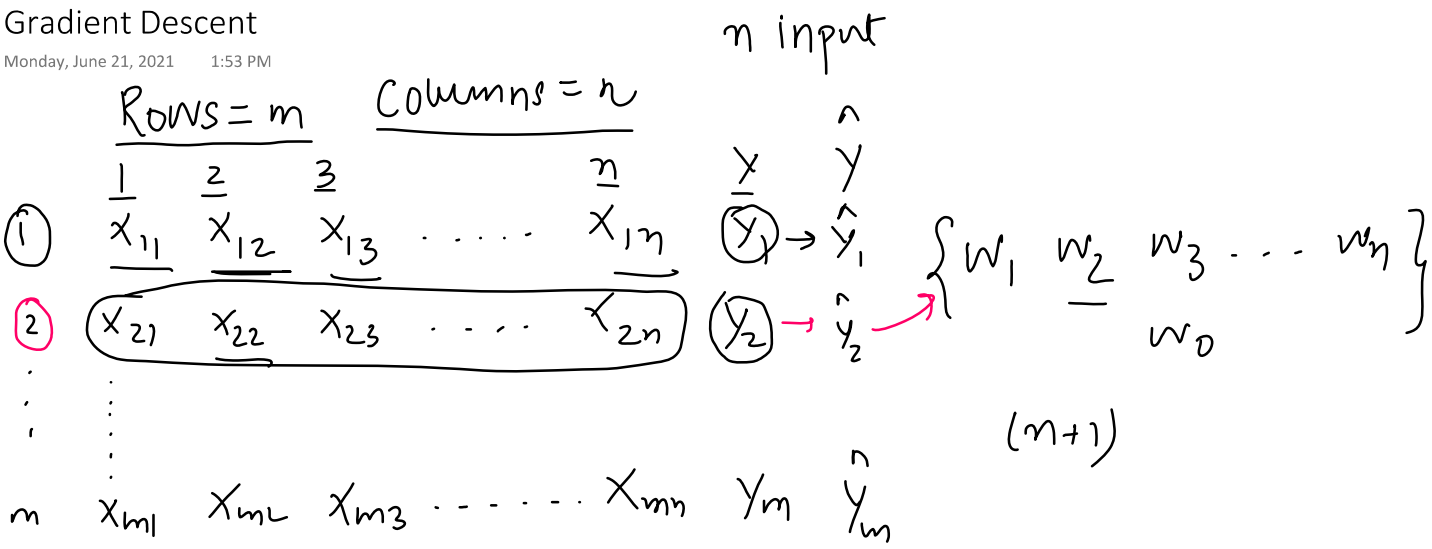
$$= -\frac{1}{2}\left[1\right]\left[2\right] = -1$$

$$= -\frac{1}{2}\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}$$

$$= -\frac{1}{2}\left[1\right] = -\frac{1}{2}$$

Gradient Descent

Monday, June 21, 2021 1:53 PM



$$\sigma(w_1 x_{11} + w_2 x_{12} + w_3 x_{13} + \ldots + w_n x_{1n} + w_0) = \hat{y}_1$$

$$\sigma(w_1 x_{21} + w_2 x_{22} + w_3 x_{23} + \ldots + w_n x_{2n} + w_0) = \hat{y}_2$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \sigma(w_0 + w_1 x_{11} + w_2 x_{12} + \ldots + w_n x_{1n}) \\ \sigma(w_0 + w_1 x_{21} + w_2 x_{22} + \ldots + w_n x_{2n}) \\ \vdots \\ \sigma(w_0 + w_1 x_{m1} + w_2 x_{m2} + \ldots + w_n x_{mn}) \end{bmatrix}$$

$$\hat{Y} = \sigma\left(\begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} + \ldots + w_n x_{1n} \\ w_0 + w_1 x_{21} + w_2 x_{22} + \ldots + w_n x_{2n} \\ \vdots \\ w_0 + w_1 x_{m1} + w_2 x_{m2} + \ldots + w_n x_{mn} \end{bmatrix} \right)$$

$$\hat{Y} = \sigma\left(\begin{bmatrix} 1 \\ 1 \\ \vdots \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \end{bmatrix} \right)$$

w_0 · A · X · B · W

$$\hat{y} = \sigma\left(\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} x_{21} & x_{22} & \cdots & x_{2n} \\ & & & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} \vdots \\ w_n \end{bmatrix}\right)$$

$$\hat{y} = \sigma(XW)$$

$$L = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

$$L = -\frac{1}{m}\left[\sum_{i=1}^{m} y_i \log(\hat{y}_i) + \sum_{i=1}^{m} (1-y_i)\log(1-\hat{y}_i)\right]$$

$(1-y)\log\sigma(1-XW)$

$$\sum_{i=1}^{m} y_i \log(\hat{y}_i) = y_1 \log\hat{y}_1 + y_2 \log\hat{y}_2 + y_3 \log\hat{y}_3 + \cdots + y_m \log\hat{y}_m$$

$$\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix} \begin{bmatrix} \log\hat{y}_1 \\ \log\hat{y}_2 \\ \log\hat{y}_m \end{bmatrix}$$

$$\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix} \log\left(\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix}\right)$$

$$y \log\hat{y} = y \log(\sigma(XW))$$

$$L = -\frac{1}{m}\left[y\log\hat{y} + (1-y)\log(1-\hat{y})\right] \quad \text{(min)}$$

where $\hat{y} = \sigma(XW)$ ← (GD) [W] find

where $\hat{y} = \sigma(XW)$ — GD

Loss function in Matrix form

$$L = -\frac{1}{m}\left[y \log(\sigma(wx)) + (1-y)\log(1-\sigma(wx)) \right]$$

minimum

gradient descent

$W = [\quad]$ $w_0 \rightarrow w_n$

for i in epochs:

$$W = W - \eta \frac{\Delta L}{\Delta W}$$

$(\eta+1)$

$$\frac{\Delta L}{\Delta W} = \left[\frac{\partial L}{\partial w_0}, \frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \cdots \frac{\partial L}{\partial w_n} \right]$$

$\frac{\Delta L}{\Delta W}$

$$L = -\frac{1}{m}\left[y \log \hat{y} + (1-y)\log(1-\hat{y}) \right]$$

$\frac{\Delta L}{\Delta W}$

$$\rightarrow \frac{dL}{dw} =$$

$$\frac{d}{dw} y \log \hat{y} \Rightarrow y \frac{d}{dw} \log \hat{y} \Rightarrow \frac{y}{\hat{y}} \frac{d}{dL}(\hat{y})$$

$\Rightarrow \hat{y}$

$$\Rightarrow \frac{y}{\hat{y}} \frac{d}{dL} \sigma(wx) \Rightarrow \frac{y}{\hat{y}} \sigma(wx)[1-\sigma(wx)] \frac{d}{dw}(wx)$$

$$= \frac{y}{\hat{y}} \hat{y}(1-\hat{y}) x = \boxed{y(1-\hat{y})x} \qquad \hat{y} = \sigma(wx)$$

$= \frac{1}{\hat{y}}$

$\frac{d}{dw}(1-y)\log(1-y) \Rightarrow (1-y)\frac{d}{dw}\log(1-\hat{y}) \Rightarrow \frac{(1-y)}{(1-\hat{y})}\frac{d}{dw}[1-\hat{y}]$

$= -\frac{(1-y)}{(1-\hat{y})}\frac{d}{dw}\sigma(wx) \Rightarrow -\frac{(1-y)}{(1-\hat{y})}[\sigma(wx)[1-\sigma(wx)]\frac{d}{dL}(wx)$

$\Rightarrow -\frac{(1-y)}{(1-\hat{y})}\hat{y}(1-\hat{y})x = \boxed{-\hat{y}(1-y)x}$

$\frac{dL}{dw} = -\frac{1}{m}[y(1-\hat{y})x - \hat{y}(1-y)x]$

$= -\frac{1}{m}[y(1-\hat{y}) - \hat{y}(1-y)]x$

$= -\frac{1}{m}[y - y\hat{y} - \hat{y} + y\hat{y}]x$

$\boxed{\frac{\Delta L}{\Delta w} = -\frac{1}{m}(y-\hat{y})x}$

gd

$\boxed{w = w + \eta\frac{1}{m}(y-\hat{y})x}$

$$\underline{W} = \underline{W} + \eta \frac{1}{m} (y - \hat{y}) \underline{X}$$

$(n+1, 1)$

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

$m, n+1$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \rightarrow m, 1$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} \rightarrow (m, 1)$$

$$\underline{W} = \underline{W} + \boxed{\frac{\eta}{m}} (\underline{y - \hat{y}}) \underline{X}$$

$(n+1, 1) = (n+1, 1) \quad (1, m) \; m, (n+1) \rightarrow (1, n+1)$

$(n+1, 1)$

Accuracy

Tuesday, June 22, 2021 1:12 PM

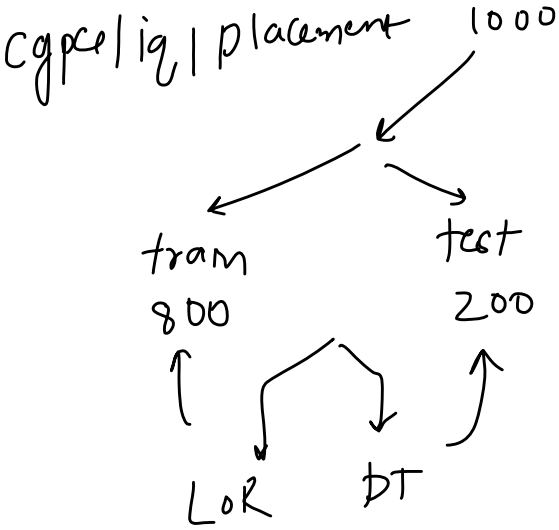
binary

Actual Label	Logistic Regression Prediction	Decision Tree Prediction
1	✓ 1	✓ 1
0	✗ 1	✗ 1
0	✓ 0	✓ 0
0	✓ 0	✓ 0
1	✓ 1	✓ 1
1	✓ 1	✓ 1
0	✗ 1	✓ 0
0	✓ 0	✓ 0
0	✓ 0	✓ 0
1	✓ 1	✓ 1

Accuracy = $\frac{\text{no. of } \checkmark}{\text{total predictions}}$

$\frac{9}{10} = 90\%$

$\frac{8}{10} = 0.8 \rightarrow 80\%$



Accuracy of multi-classification problem

Wednesday, June 23, 2021 8:44 AM

↘

Actual Label	Logistic Regression Prediction	Decision Tree Prediction
0	0	0
0	0	0
0	0	0
2	2	2
0	0	0
2	2	2
0	0	0
2	2	2
1	1	1
1	1	1

iris

setosa, virginica, versicolor

0 1 2

accuracy = # correct prediction / # total

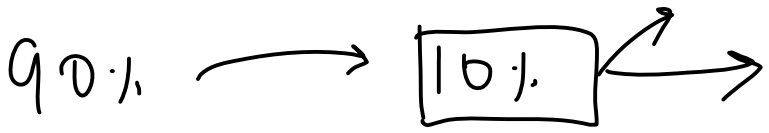
= 10/10 = 1 = 100%

How much accuracy is good?

Wednesday, June 23, 2021 11:14 AM

The Problem with Accuracy

Wednesday, June 23, 2021 11:14 AM



binary

y | N

→ y ✓ → x

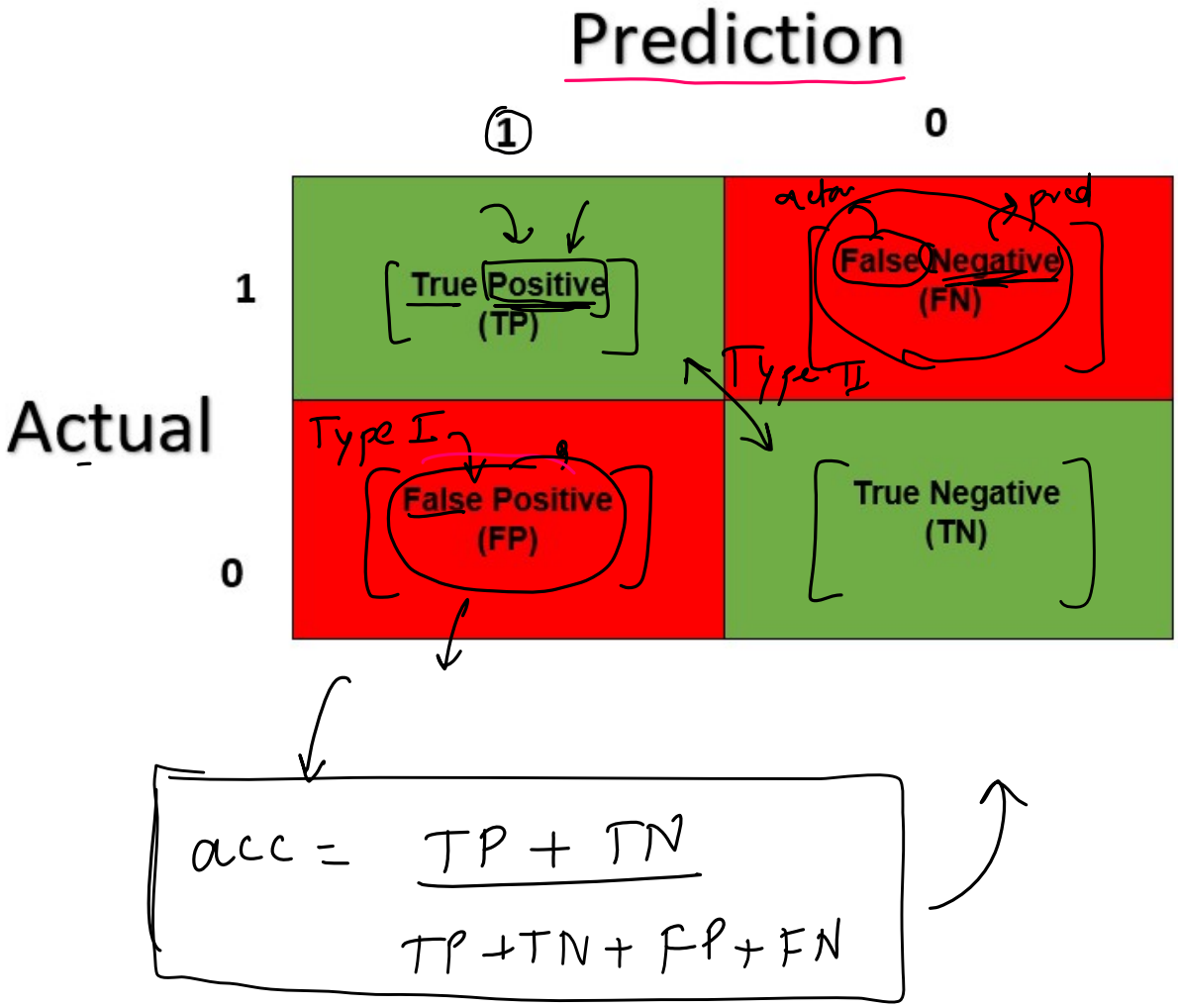
→ N x → ✓

y ✓ → x

N x → ✓

Confusion Matrix

Tuesday, June 22, 2021 1:12 PM



$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Extented

↳ { echo or not }

duplicate

↳ echo

1+2+3=6 → outcome

x	y
2	y

Type 1 Error

Wednesday, June 23, 2021 8:45 AM

Type 2 Error

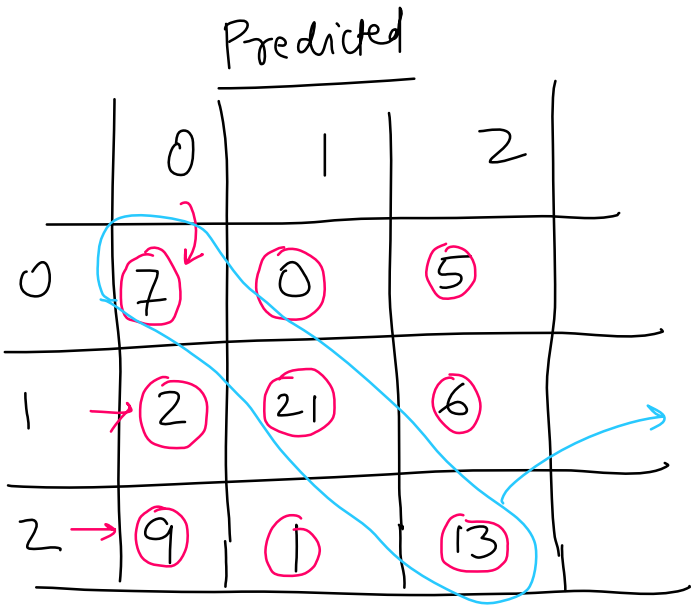
Wednesday, June 23, 2021 8:45 AM

Confusion Matrix for Multi-classification Problem

Wednesday, June 23, 2021 8:45 AM

0, 1, 2 →

10 x 10



3 x 3

binary

2 x 2

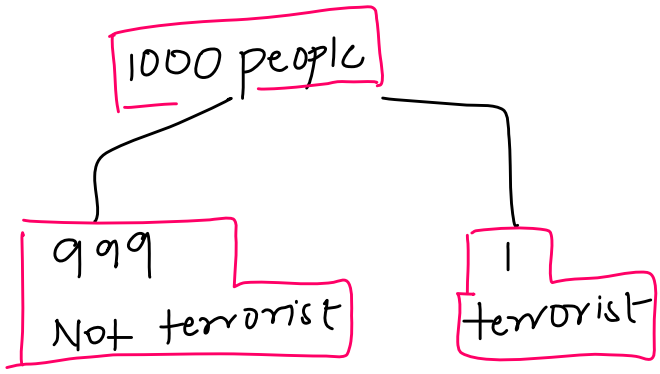
True predict / total = accuracy

3 x 3

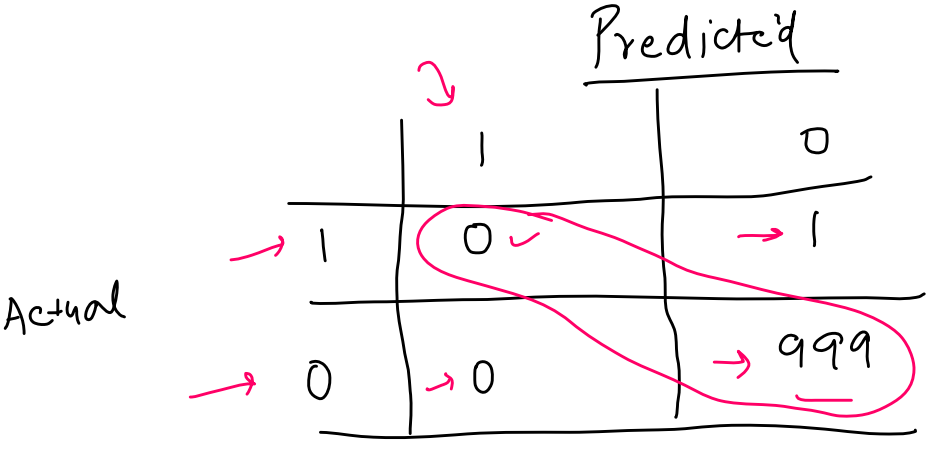
When accuracy is misleading?

Wednesday, June 23, 2021 8:45 AM

Imbalanced Dataset



model → No one is terrorist



Accuracy = $\frac{999}{999+1}$

= 99.9%

Precision

Wednesday, June 23, 2021 8:46 AM

{ spam : 1 , not spam : 0 }

Predicted

Actual

	Sent to Spam	Not sent to spam
Spam	100	170 FN
Not Spam	30 FP	700

	Sent to Spam	Not sent to spam
Spam	100	190
Not Spam	10	700

$P_A = \frac{100}{100+30}$ $P_B = \frac{100}{100+10}$

$P_A < P_B$

80: FP → Not spam → spam

FN → spam → Not spam

What proportion of predicted Positives is truly Positive?

1 0

Binary +ve

	Sent to Spam	Not sent to spam
Spam	True Positive	False Negative
Not Spam	False Positive	True Negative

$$\text{Precision} = TP/(TP+FP)$$

Recall

Wednesday, June 23, 2021 8:46 AM

has cancer : 1 , no cancer : 0

Predicted

Actual

	Detected Cancer	Not Detected
Has Cancer	1000	200 (FN)
No Cancer	800 (FP)	8000

A 90%

Recall = $\frac{1000}{1200}$

	Detected Cancer	Not Detected
Has Cancer	1000	500 ← B
No Cancer	500	8000

B 90%

$R_B = \frac{1000}{1500}$ ← (FP)

$R_A > R_B$

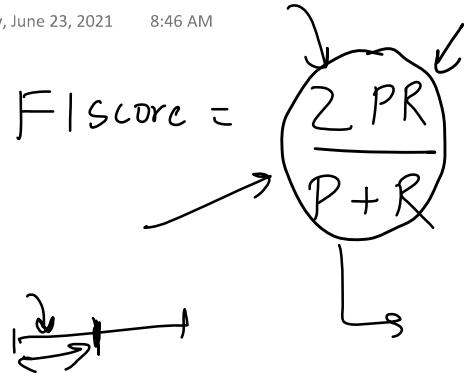
What proportion of actual Positives is correctly classified?

	Detected Cancer	Not detected Cancer
Has Cancer	True Positive	False Negative
No Cancer	False Positive	True Negative

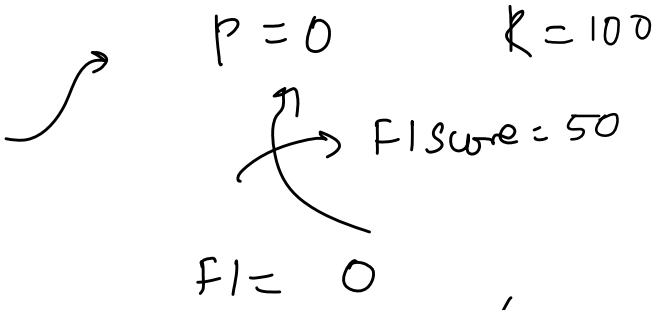
$$\text{Recall} = TP/(TP+FN)$$

F1 Score

Wednesday, June 23, 2021 8:46 AM



$$F1score = \frac{P+R}{2}$$



(A)

$P = R = 80$

$\frac{80+80}{2} = 80$ $\quad F1 = 80$

$\frac{2 \times 80 \times 80}{160} = 80$

(B) (75)

$P = 60 \quad R = 100$

$\frac{100+60}{2} = 80$

$= \frac{2 \times 60 \times 100}{160}$

$= 75$

Multi-class Precision and Recall

Wednesday, June 23, 2021 8:46 AM

Positive (1) ↙ Yes / NO

Binary (2) class

Multi (3)

Dog Cat Rabbit

$\frac{2 \times 0.86 \times 0.66}{0.86 + 0.66}$

R_D, R_C, R_R

$F1_D = \frac{2 P_D R_D}{P_D + R_D}$

$F1_C = \frac{2 P_C R_C}{P_C + R_C}$

$F1_R$

Predicted

Actual	Dog	Cat	Rabbit	Total	Recall
Dog	25	5	10	40	0.62
Cat	0	30	4	34	0.88
Rabbit	4	10	20	34	0.58
Total	29	45	34		
Precision	0.86	0.66	0.58		

$R_D = \frac{25}{40} = 0.62$, $R_C = \frac{30}{34} = 0.88$, $R_R = \frac{20}{34} = 0.58$

macro recall ←

weighet recall

$F1_D, F1_C, F1_R$ → macro f1, weighted f1

Multi-class F1 Score

Thursday, June 24, 2021 11:14 AM

Softmax Regression

Monday, June 28, 2021 3:35 PM

cgpa	iq	placed?
7.1	71	0
8.5	85	1
9.5	95	2

{ Yes } 1
{ No } 0
Optout 2

↘ classified

{6.5, 65}
↳ Yes / No / Optout

{ Softmax
Multinomial
LR } → Deep Learning

Binary classification

Softmax → general
↑ for 2class

$0 < \sigma < 1$

Softmax function

K = no. of class {3}

Yes → 1
no → 2
opt → 3

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1} e^{z_j}}$$

$$\sigma(z)_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

(Yes)

$$\sigma(z)_1 = \left\{ \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \right\}$$

No

$$\sigma(z)_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

Training Intuition

Monday, June 28, 2021 3:35 PM

{Y, N, D} → 0 1 2

cgpa | iq | placed?

–	–	0
–	–	1
–	–	2

→ transform → OHE → 5

cgpa ✓	iq ✓	=0	=1	=2
		1	0	0
		0	1	0
		0	0	1

3 coeff

$w_1\ w_2\ w_0$

D → D_1, D_2, D_3

D_1: cgpa | iq | = 0 → m_1 → $w_1^{(0)}, w_2^{(0)}, w_0^{(0)}$

D_2: cgpa | iq | = 1 → m_2 → $w_1^{(1)}, w_2^{(1)}, w_0^{(1)}$

D_3: cgpa | iq | = 2 → m_3 → $w_1^{(2)}, w_2^{(2)}, w_0^{(2)}$

{ loss function, gradient descent }

Loss Function

Monday, June 28, 2021 3:47 PM

$$L = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

$K = \{1, 2, 3\}$ $i = 1$

$$L = -\frac{1}{m}\sum_{i=1}^{m}\sum_{K=1}^{K} y_K^{(i)} \log(\hat{y}_K^{(i)})$$

$w_1^{(1)}\ w_2^{(1)}\ w_0^{(1)}$

	x_1	x_2	y	$y_{K=1}$	$y_{K=2}$	$y_{K=3}$
1 →	x_{11}	x_{12}	1	1	0	0
2 →	x_{21}	x_{22}	2	0	1	0
3	x_{31}	x_{32}	3	0	0	1

$$y_1^{(1)} \log(\hat{y}_1^{(1)}) + y_2^{(1)} \log(\hat{y}_2^{(1)}) + y_3^{(1)} \log(\hat{y}_3^{(1)}) +$$

$$+ y_1^{(2)} \log(\hat{y}_1^{(2)}) + y_2^{(2)} \log(\hat{y}_2^{(2)}) + y_3^{(2)} \log(\hat{y}_3^{(2)}) +$$

$$+ y_1^{(3)} \log(\hat{y}_1^{(3)}) + y_2^{(3)} \log(\hat{y}_2^{(3)}) + y_3^{(3)} \log(\hat{y}_3^{(3)})$$

$$L = y_1^{(1)} \log(\hat{y}_1^{(1)}) + y_2^{(2)} \log(\hat{y}_2^{(2)}) + y_3^{(3)} \log(\hat{y}_3^{(3)})$$

$\hat{y}_1^{(1)}$, $\hat{y}_2^{(2)}$, $\hat{y}_3^{(3)}$

softmax

$$\hat{y}_1^{(1)} = \sigma(w_1^{(1)} x_{11} + w_2^{(1)} x_{12} + w_0^{(1)})$$

$$y_2^{(2)} = \sigma(w_1^{(2)} x_{21} + w_2^{(2)} x_{22} + w_0^{(2)})$$

$$y_3^{(3)} = \sigma(w_1^{(3)} x_{31} + w_2^{(3)} x_{32} + w_0^{(3)})$$

$$\begin{bmatrix} w_1^{(1)} & w_2^{(1)} & w_0^{(1)} \\ w_1^{(2)} & w_2^{(2)} & w_0^{(2)} \\ w_1^{(3)} & w_2^{(3)} & w_0^{(3)} \end{bmatrix}$$

(L)

$\frac{\partial L}{\partial w_1^{(1)}}, \frac{\partial L}{\partial w_2^{(1)}}, \frac{\partial L}{\partial w_0^{(1)}} \ldots$ 9 du

gradient

9 - values init = 1

$\begin{bmatrix} \\ \\ \\ \end{bmatrix} =$

loop → 1000 epochs

$$w_1^{(1)} = w_1^{(1)} - \eta \frac{\partial L}{\partial w_1^{(1)}}$$

$$w_2^{(1)} = w_2^{(1)} - \eta \frac{\partial L}{\partial w_2^{(1)}}$$

⋮

Prediction

Monday, June 28, 2021 3:35 PM

$S_x = \{7, 70\} \Rightarrow$ Y, N, Opt

m_1 Yes

m_2 No

m_3 Opt out

$w_1^{(1)}$ $w_2^{(1)}$ $w_0^{(1)}$

$w_1^{(2)}$ $w_2^{(2)}$ $w_0^{(2)}$

$w_1^{(3)}$ $w_2^{(3)}$ $w_0^{(3)}$

$z_1 = 7 \times w_1^{(1)} + 70 \times w_2^{(2)} + w_0^{(1)}$

$z_2 = 7 \times w_1^{(2)} + 70 \times w_2^{(1)} + w_0^{(2)}$

$z_3 = 7 \times w_1^{(3)} + 70 \times w_2^{(3)} + w_0^{(3)}$

$$\sigma(y) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$= 0.40$

$$\sigma(N) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

0.35

$$e(O) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

0.25

Sigmoid Vs Softmax

Monday, June 28, 2021 3:47 PM

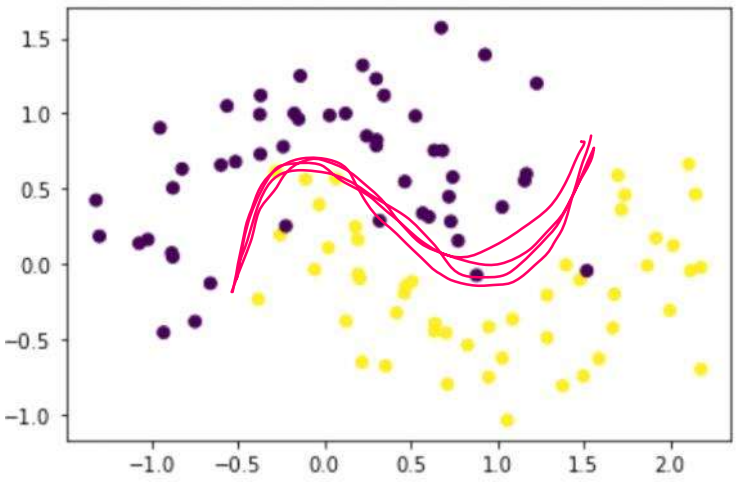
Code Sample

Monday, June 28, 2021 3:36 PM

Polynomial Logistic Regression

Monday, June 28, 2021 6:43 PM

x_1 x_2 y {0,1}



Linear polynomial

degree

x^0 x^1 x^2 x^3

degree = 2

x_1 x_2 → 2 cols

6 cols

x_1^0 x_1^1 x_1^2 x_2^0 x_2^1 x_2^2 y

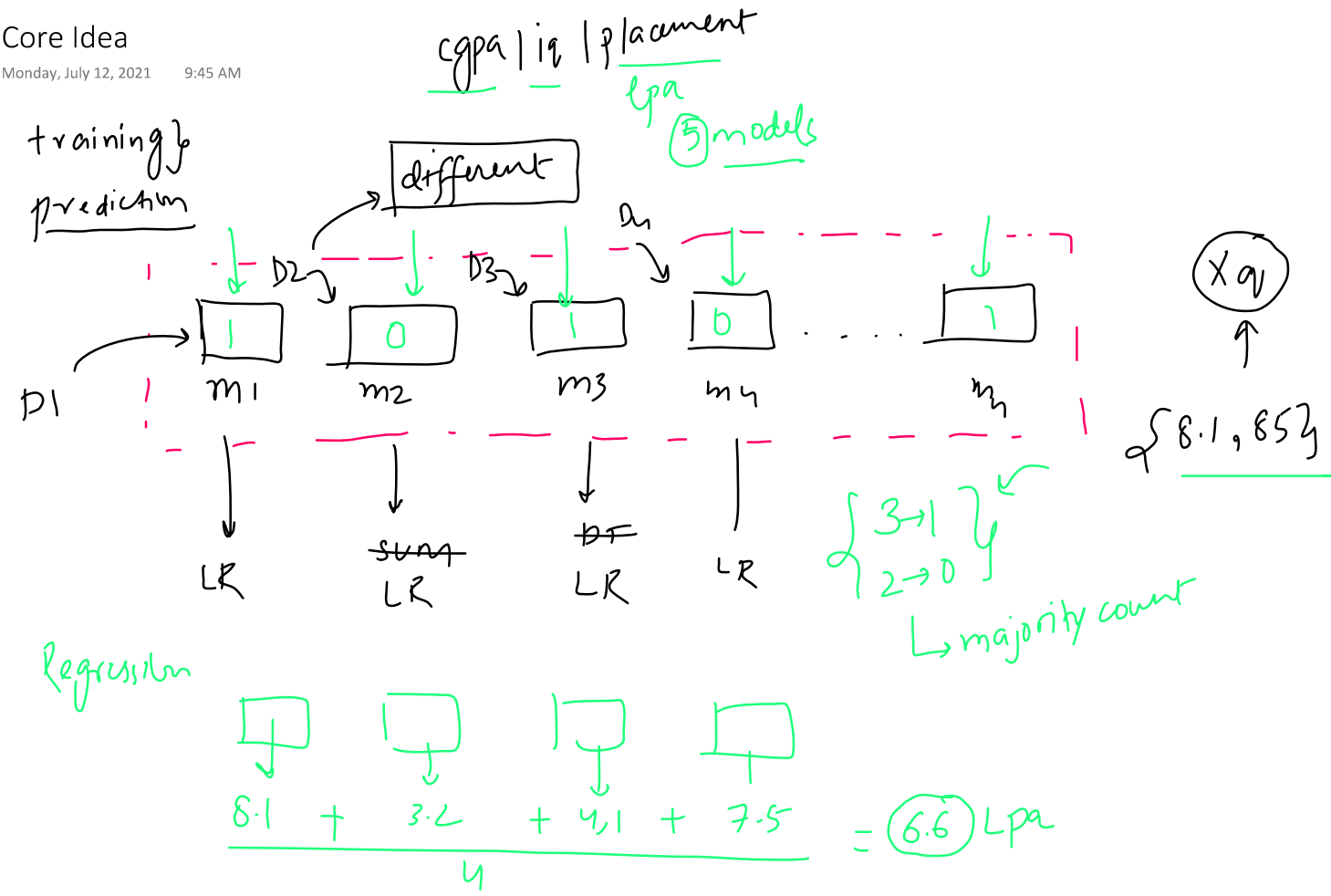
w_0 w_1 w_2 w_3 w_5 w_8

Wisdom of the Crowd

Monday, July 12, 2021 9:56 AM

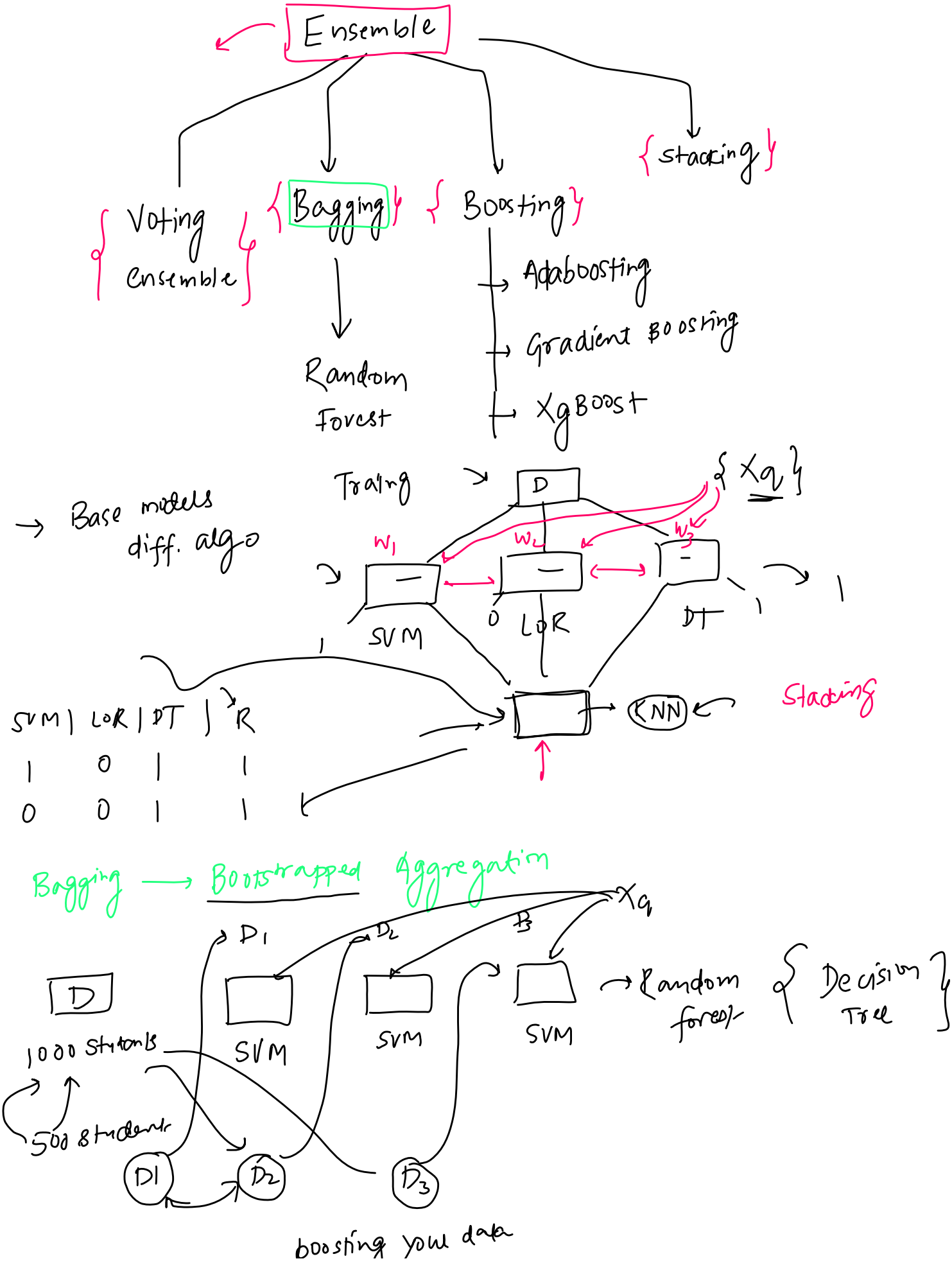
Core Idea

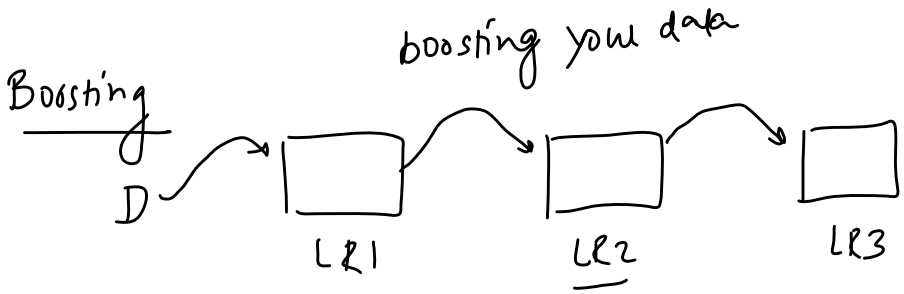
Monday, July 12, 2021 9:45 AM



Type of Ensemble Learning

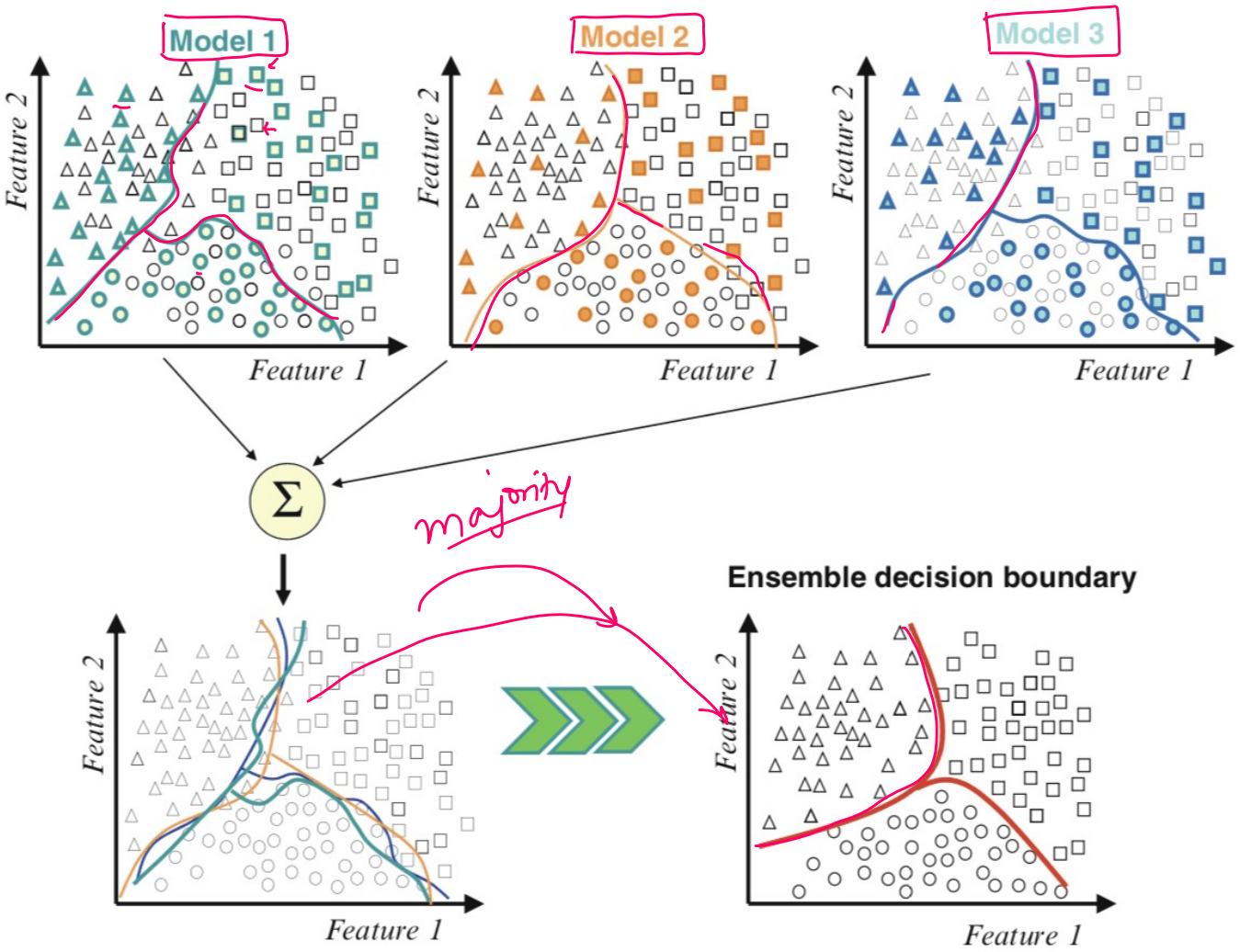
Monday, July 12, 2021 9:45 AM

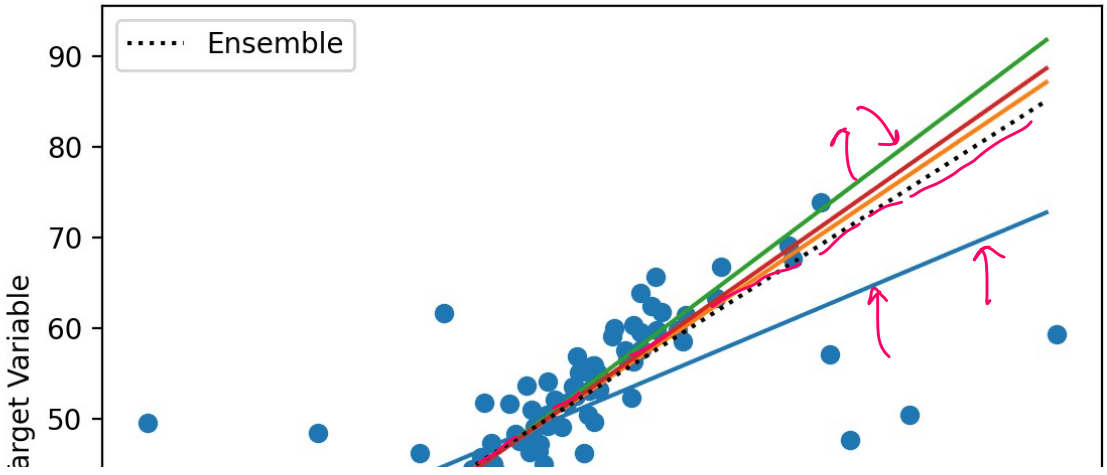


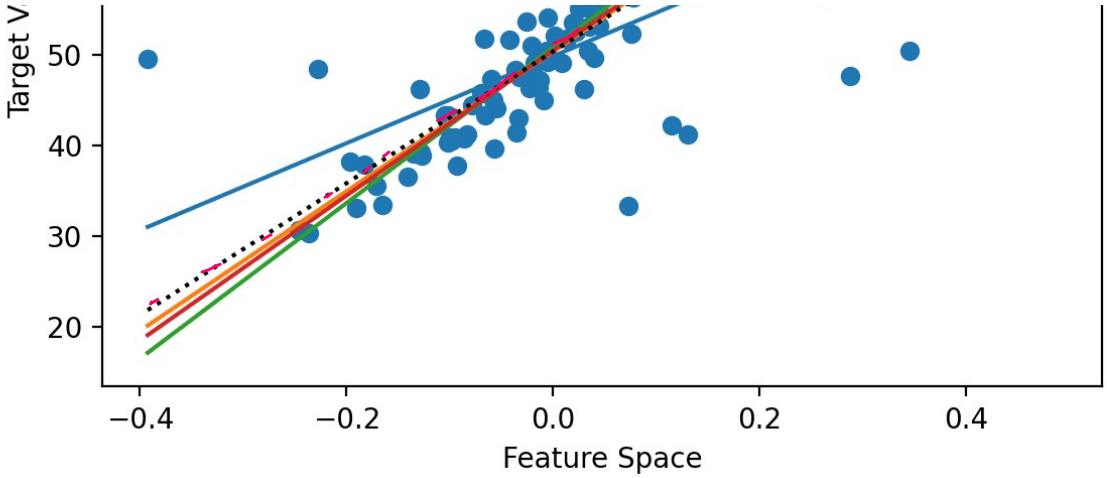

boosting your data
Boosting
D
LR1
LR2
LR3

Why it works?

Monday, July 12, 2021 9:45 AM





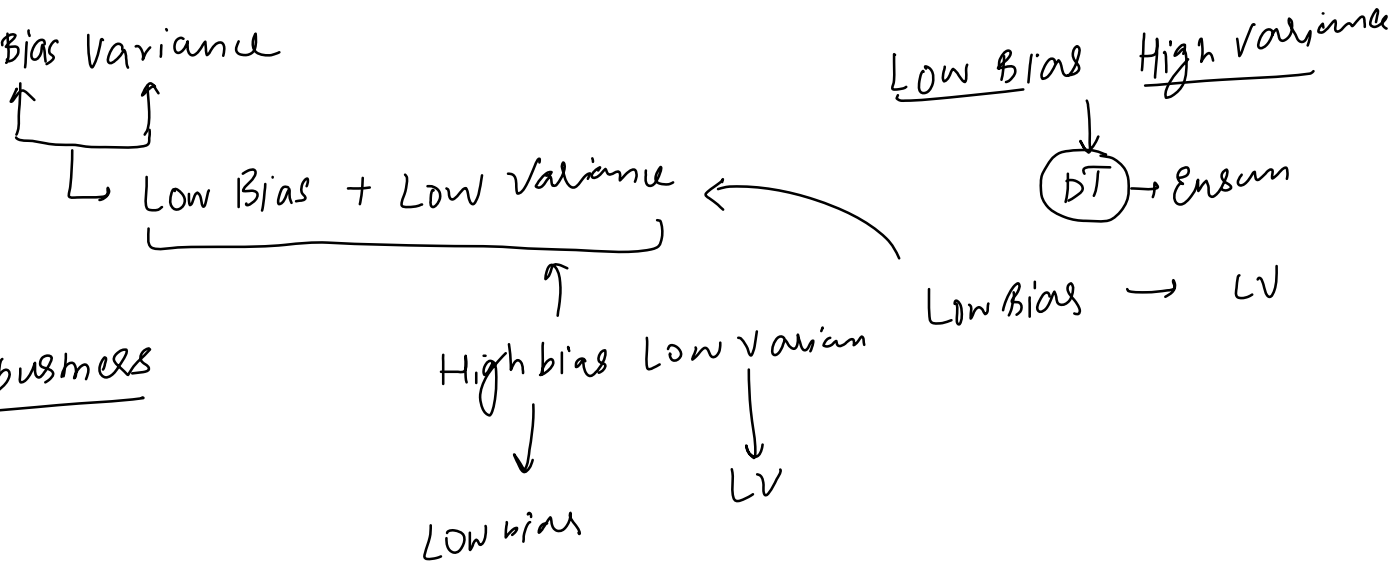

Target V
50
40
30
20
−0.4
−0.2
0.0
0.2
0.4
Feature Space

Benefits

Monday, July 12, 2021 9:45 AM

1) Improvement in performance

2) Bias Variance



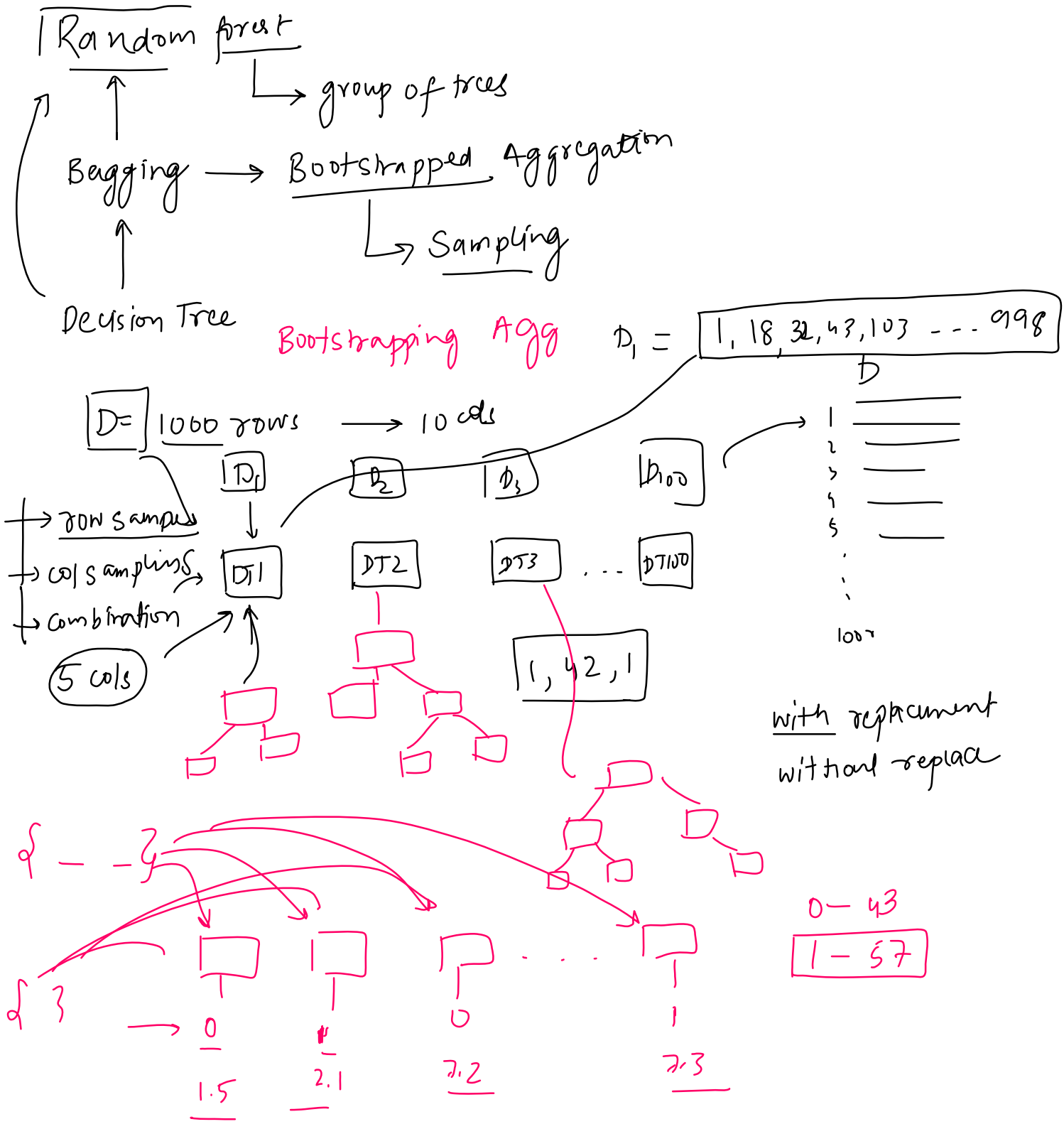
3) Robustness

When to use?

Monday, July 12, 2021 9:46 AM

Intuition

Monday, July 19, 2021 4:53 PM



Demo

Monday, July 19, 2021 5:05 PM

Bias Variance and Random Forest

Tuesday, July 20, 2021 10:52 AM

LB LV

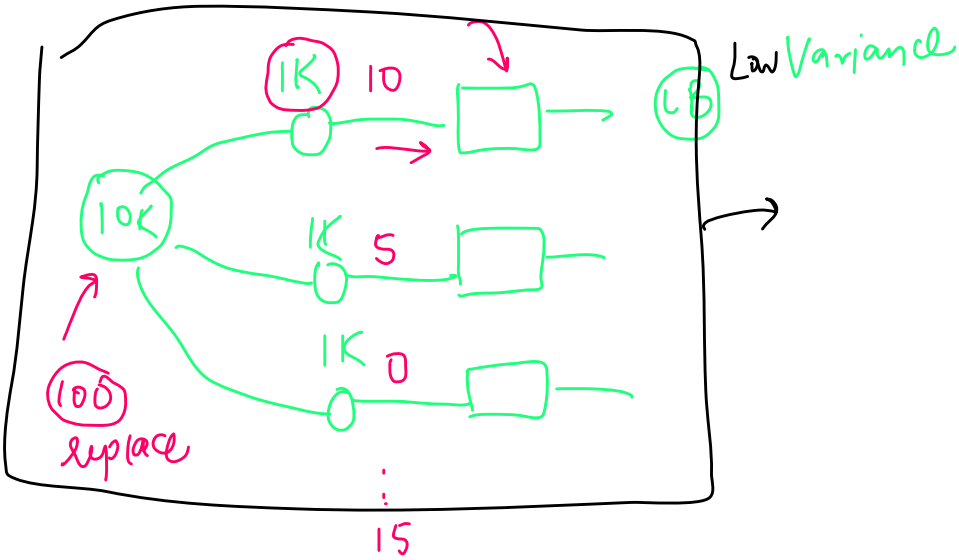
$B \propto \frac{1}{V}$

LB HV	HB LV
→ Fully grown DT	→ LR
→ SVM	
→ KNN	

Random Forest

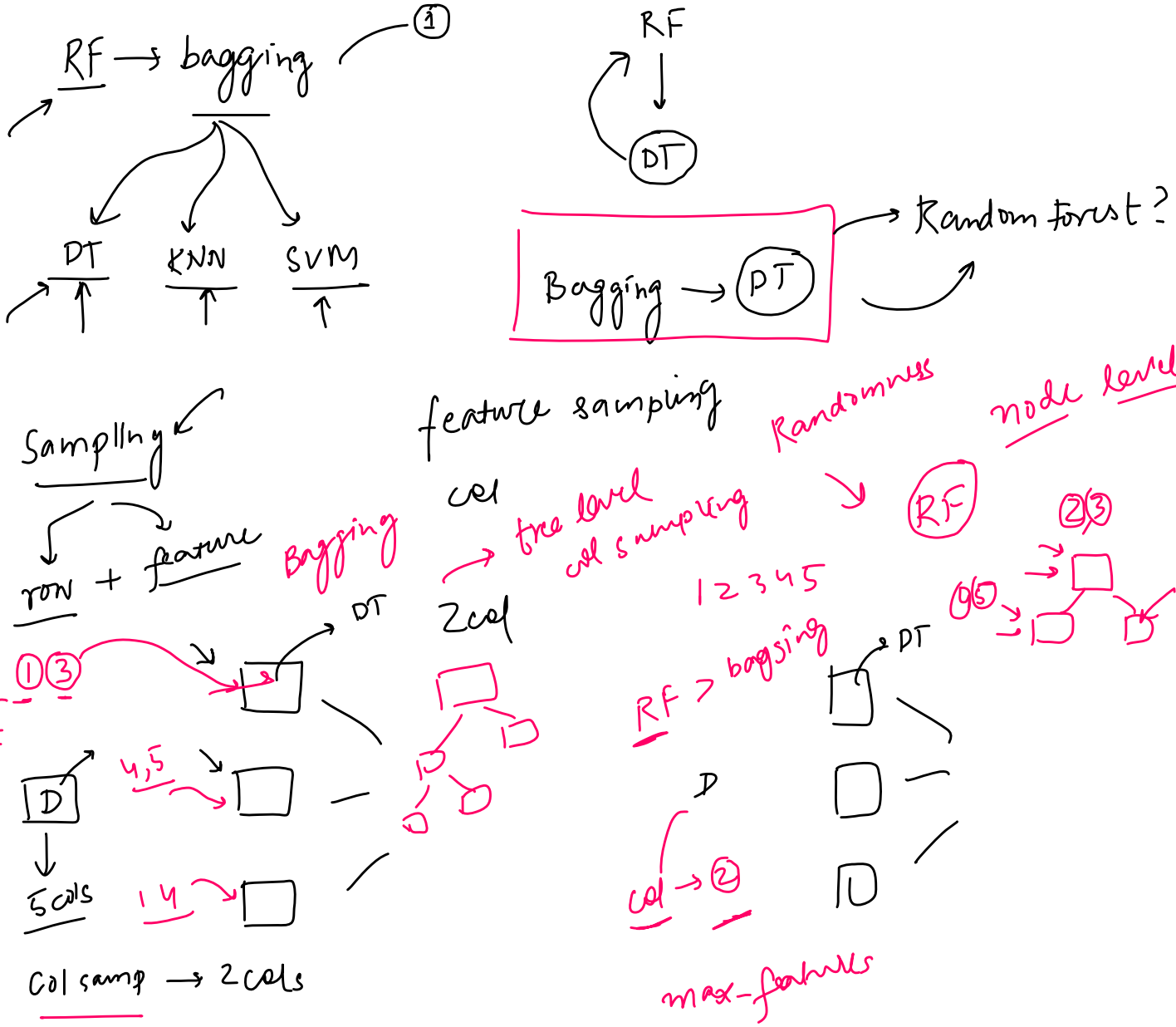
→ LB HV

LB LV



Bagging Vs Random Forest

Monday, July 19, 2021 4:53 PM



Hyperparameters

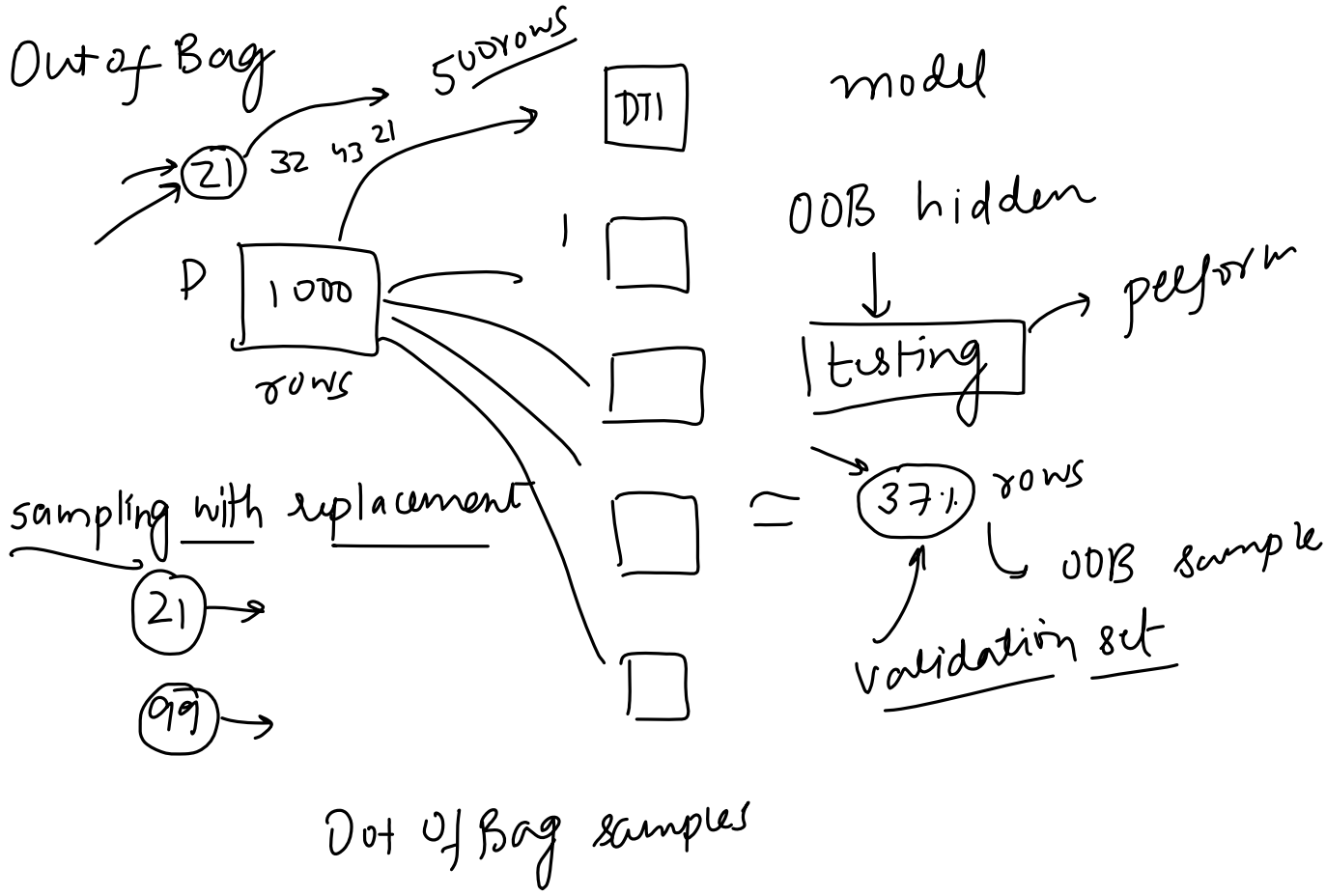
Monday, July 19, 2021 4:56 PM

Code Demo

Monday, July 19, 2021 5:05 PM

OOB Evaluation

Monday, July 19, 2021 5:14 PM



Feature Importance

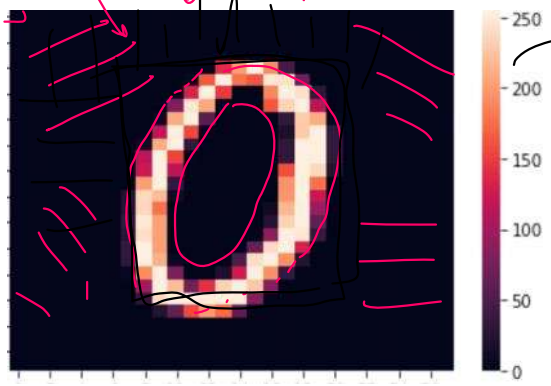
Monday, July 19, 2021 4:56 PM

MNIST

	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel774	pixel775	pixel776	pixel777	pixel778
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

42000 images

28×28 → 784



feature import

Formula

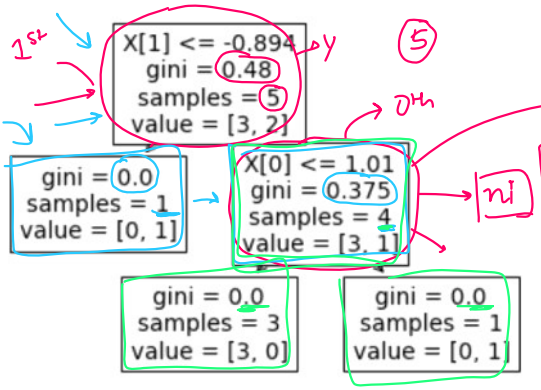
$$ni = \frac{N\_t}{N}\left[\text{impurity} - \left(\frac{N\_t\_r}{N\_t} \times \text{right\_impurity}\right) - \left(\frac{N\_t\_l}{N\_t} \times \text{left\_impurity}\right)\right]$$

Ntr → 4 5

NTL-1 5

[0, 1] → fi

$$fi_K = \frac{\sum_{j \in \text{node split on feature } K} ni}{\sum_{j \in \text{all nodes}} ni}$$



ni

0th = X / x+y

1st = Y / x+y → 0.18

$$= \frac{5}{5}\left[0.48 - \frac{4}{5} \times 0.375 - \frac{1}{5} \times 0.\right] = y \Rightarrow \left[0.48 - (0.8 \times 0.375)\right]$$

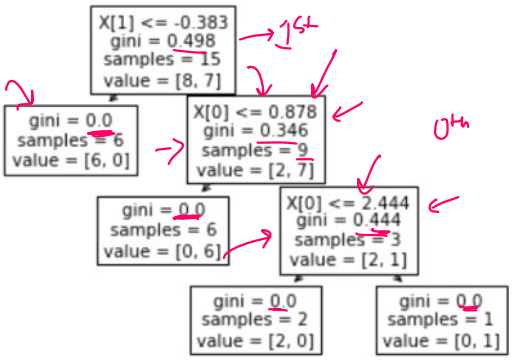
$$\Rightarrow 0.48 - 0.30 = 0.18$$

$$\frac{4}{5}\left[0.375\right] = x = 0.8 \times 0.375 = 0.30 = x$$

$$0^{th} = \frac{0.3}{0.3 + 0.18} = 0.625$$

$0^{th} = \frac{0.3}{0.3 + 0.18} = 0.625$

$1^{st} = \frac{0.18}{0.3 + 0.18} = 0.375$



0th Node

$\frac{15}{15}\left[0.49 - \frac{9}{15} \times 0.34\right]$

$= 0.290$

$\frac{9}{15}\left[0.34 - \frac{3}{15} \times 0.44\right]$

$= 0.118$ ←

$\frac{3}{15}\left[0.44\right]$ ←

$= 0.088$

$fi[0] = \frac{0.11 + 0.08}{0.29 + 0.11 + 0.08} = \frac{0.19}{0.48}$

$= 0.39$

$fi[1] = \frac{0.29}{0.48} = 0.60$

Extra Trees

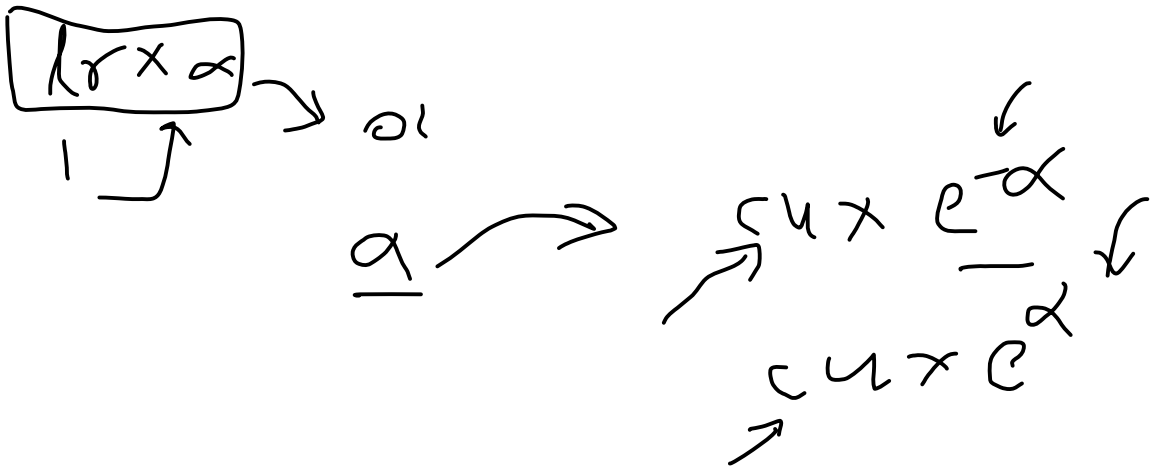
Monday, July 19, 2021 4:57 PM

Before Starting

Monday, August 9, 2021 10:42 AM

1) Weak classifiers (> 50%)

$$\alpha = \frac{1}{2}\log\left(\frac{1-e}{e}\right)$$



The Big Idea

Monday, August 9, 2021 10:43 AM

Step by Step Breakdown

Monday, August 9, 2021 10:43 AM

Points to remember

Monday, August 9, 2021 10:43 AM

Code Walkthrough

Monday, August 9, 2021 10:43 AM

Example and Hyperparameters

Monday, August 9, 2021 10:43 AM

Algorithm

Monday, September 20, 2021 8:23 AM

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.

2. For $m = 1$ to M:

   (a) For $i = 1, 2, \ldots, N$ compute

   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$

   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.

   (c) For $j = 1, 2, \ldots, J_m$ compute

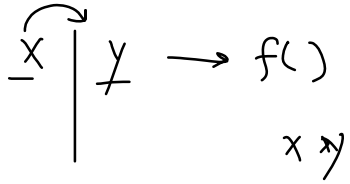
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$

   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.

Additive Modelling

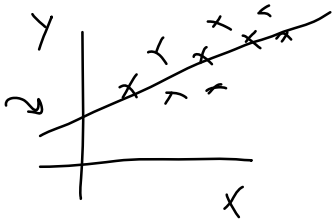
Monday, September 20, 2021 8:28 AM

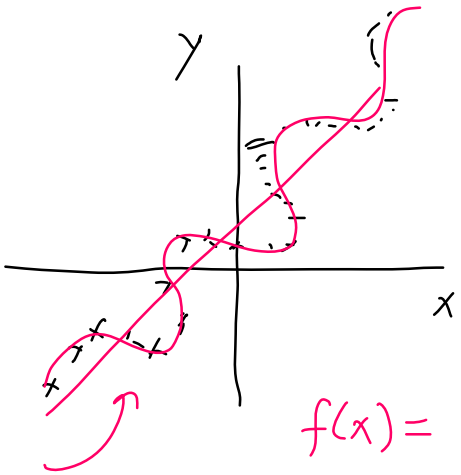


$\Rightarrow y = f(x)$

$y = f(x_1, x_2, x_3)$

$x_1\ x_2\ x_3 \mid y$





$f(x) = x + \sin x$

additive

$F(x) = f_0(x) + f_1(x) + f_2(x) + \dots$

DT PT

$y = x \quad y = \sin(x)$

Explanation

Monday, September 20, 2021 8:25 AM

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

y_i → actual, $\hat{y}_i$ → pred

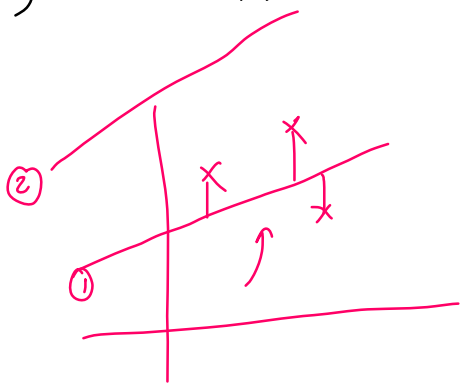
ls

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

$\frac{1}{2}$

$L(y, F(x)) \rightarrow L(y, \hat{y})$

$\hat{y}$



① = 10 = 5 W

② = 20 = 10

$y = f(x)$

$$f(x) = f_0(x) + \underline{f_1(x) + f_2(x) + \dots + f_n(x)}$$

DT

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

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \gamma)^2$$

γ

$$\frac{d f_0(x)}{d\gamma} = \frac{d}{d\gamma} \frac{1}{2} \sum_{i=1}^{n} (y_i - \gamma)^2 = \frac{1}{2} \sum_{i=1}^{n} \frac{d}{d\gamma} (y_i - \gamma)^2$$

$$\sum_{i=1}^{n} (y_i - \gamma) \frac{d}{d\gamma} (y_i - \gamma) = -\sum_{i=1}^{n} (y_i - \gamma) = 0$$

$$\sum_{i=1}^{n} (\gamma - y_i) = 0$$

$$\sum_{i=1}^{3} (\gamma - y_i) = 0 \Rightarrow (\gamma - 192) + (\gamma - 144) + (\gamma - 91) = 0$$

$$\sum_{i=1}(\gamma - y_i) = 0 \Rightarrow (\gamma - 192) + (\gamma - 144) + (\gamma \quad \cdots$$

$$3\gamma = 192 + 144 + 91$$

mean

Formula

$$\gamma = \frac{192 + 144 + 91}{3}$$

mean of output

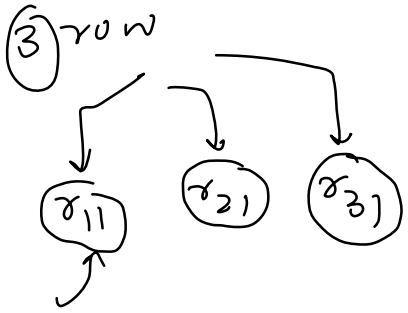
$$F(x) = \underbrace{f_0(x)}_{\text{mean (leaf)}} + \underbrace{f_1(x)}_{\uparrow} + \underbrace{f_2(x)}_{\uparrow} + \cdots \rightarrow \underbrace{f_m(x)}_{\uparrow}$$

(DT)

$$\boxed{m = 1}$$

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f = f_{m-1}}$$

$$r_{i1} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f = f_0}$$



$$\hat{y}_i = f(x_i)$$

$$r_{i1} = -\left[\frac{\partial}{\partial \hat{y}_i} L(y_i, \hat{y}_i)\right]_{f = f_0}$$

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

$$r_{i1} = -\left[\frac{\partial}{\partial \hat{y}_i} \frac{1}{2}(y_i - \hat{y}_i)^2\right]_{f = f_0}$$

$$= \left[(y_i - \hat{y}_i)\right]_{f = f_0} = \left[(y_i - f(x_i)\right]_{f = f_0}$$

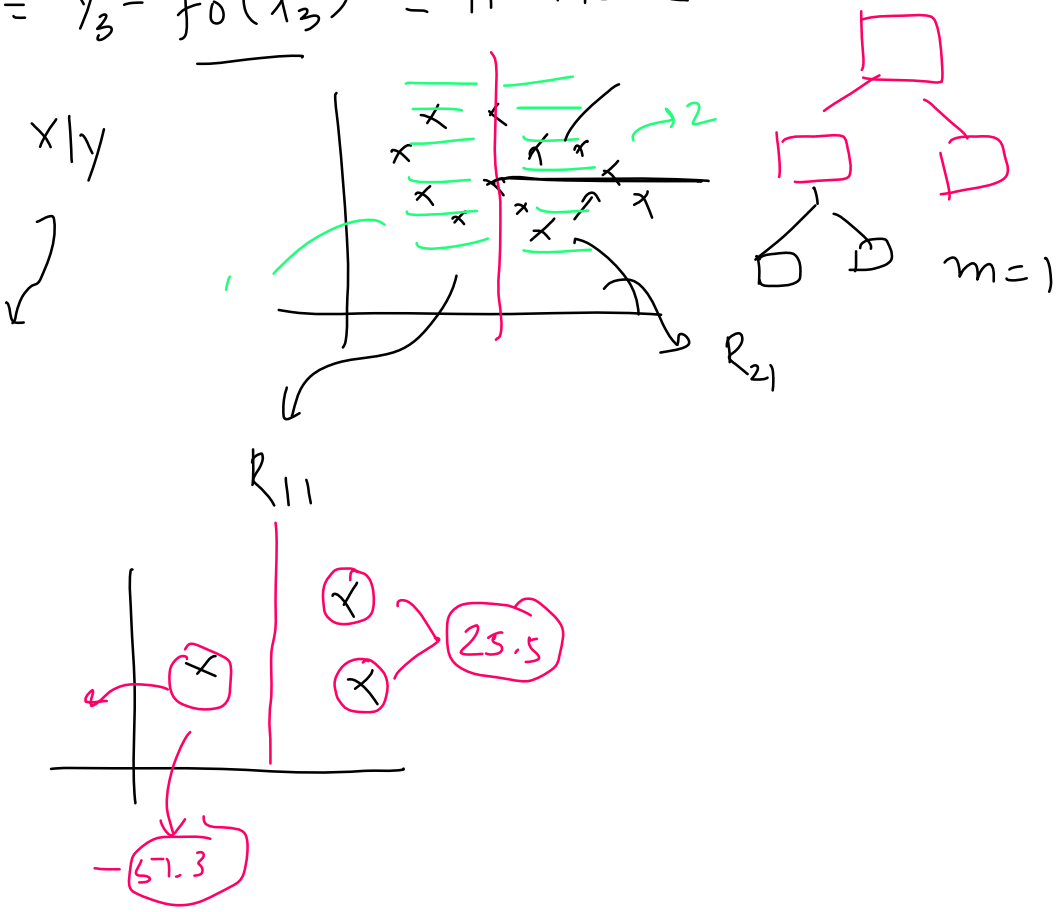
$$= \left[\frac{1}{2}(y_i - \hat{y}_i) \right]_{f=f_0} = \left[\frac{1}{2}(y_i - \hat{y}^{(n-1)}) \right]_{f=f_0}$$

$$r_{i1} = (y_i - f_0(x_i))$$

$$r_{11} = y_1 - f_0(x_1) = 192 - 142 =$$

$$r_{21} = y_2 - f_0(x_2) = 144 - 142 =$$

$$r_{31} = y_3 - f_0(x_3) = 91 - 142 =$$



$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

$$\gamma_{j1} \Rightarrow \gamma_{11} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{11}} L(y_i, f_{m-1}(x_i) + \gamma)$$

γ_{11} γ_{21}

γ_{11} γ_{21}

$$\gamma_{11} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(y_i - (f_0(x_i) + \gamma))^2$$

$$\frac{dL}{d\gamma} = \frac{1}{2} \times 2 (y_i - (f_0(x) + \gamma)) \frac{d}{d\gamma} (y_i - f_0(x) - \gamma) = 0$$

$$= -(y_i - f_0(x) - \gamma) = 0$$

$$= y_i - f_0(x) - \gamma = 0$$

$$\gamma_{11} = 91 - 142 - \gamma = 0$$

$$\boxed{\gamma = 91 - 142 = -51}$$

$$\gamma_{21} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{21}} L(y_i, f_0(x_i) + \gamma)$$

$$= \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{2} (y_i - (f_0(x_i) + \gamma))^2$$

$$= -\sum_{i=1}^{2} (y_i - f_0(x_i) - \gamma) = 0$$

$$= \sum_{i=1}^{2} (y_i - f_0(x_i) - \gamma) = 0$$

$$= y_1 - f_0(x_1) - \gamma + y_2 - f_0(x_2) - \gamma = 0$$

336
284

$$= 192 - 142 - \gamma + 144 - 142 - \gamma = 0$$

$336 - 284$

$52 - 2\gamma = 0$

$\gamma = \frac{52}{2} = 26$