Gradient desent

$w^{k+1} = w^k - \underbrace{\text{update}}_{\alpha \cdot \text{error}}$ → Gradient

Essential of statistics learning (maths)
by Hastie

Introduction of statistics learning (R codes)

2) C.M. Bishop - for Beginners

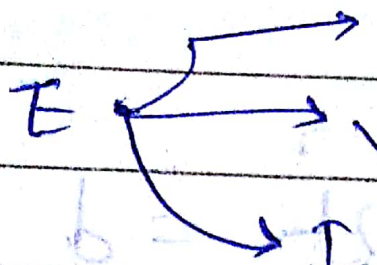
3) Mitchell - D.T., Gini, entropy.

function estimation / Approximation

linear regression

$p = mx + c$

$P = \underbrace{f(x) + \epsilon}_{y}$



Bias :-

The moment you decide that model will grow as i thought

error ~~due~~ is introduce due to Assumption or naive Assumption or false assumption called it Bias error

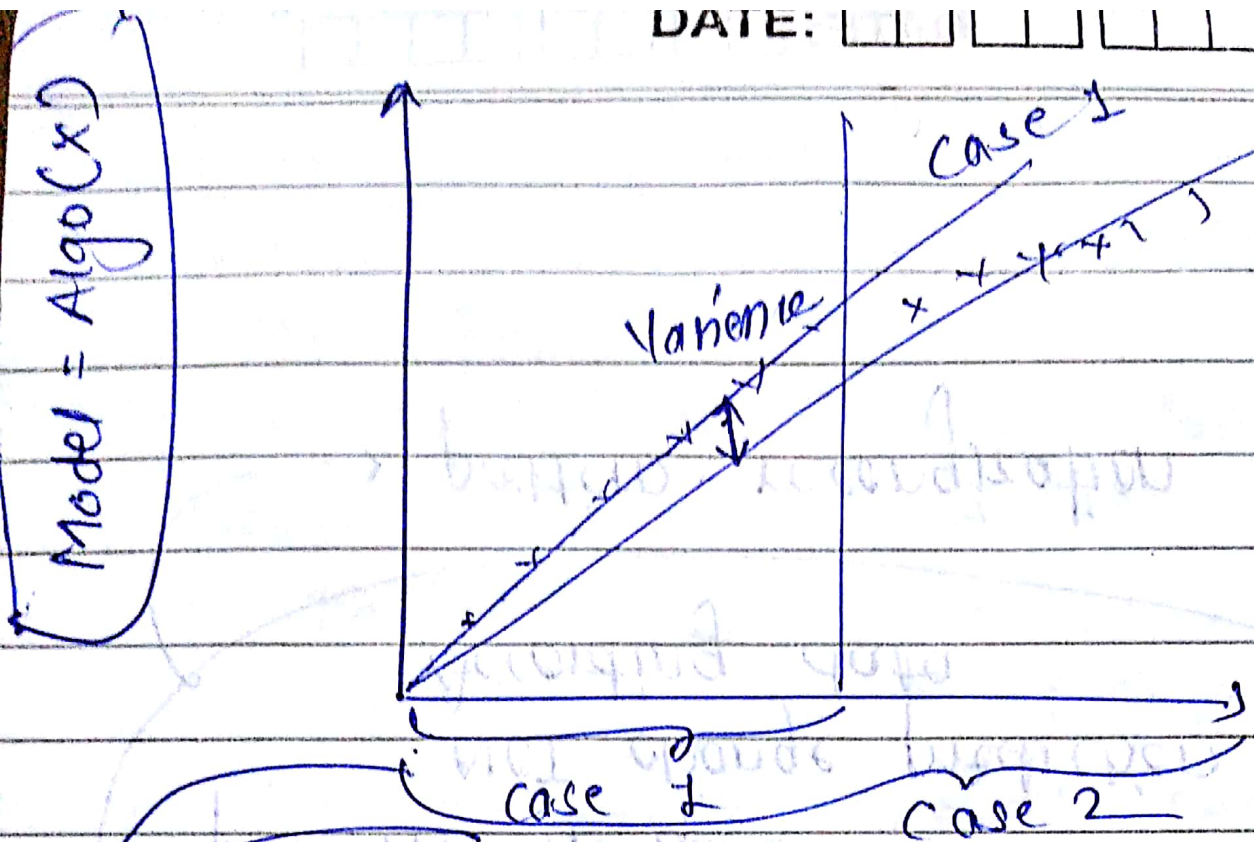
The moment when decide model then condition may be bias.

Variance :-

Who model is sensitive to data

The moment we introduce

DATE:



case 1 case 2

model $P = w_0 + w_1 x_1$ prediction is

say 1.8 3.2 ( )

case Now model is changed

say $P = w_0 + w_1 x_1$ prediction is change from above

$= 2.0 + 2.8$

- so value of w_0 & w_1 is change due to you re train the model when you take new data points.
- So value of w_0 & w_1 varies
- old data prediction is change when new data came in

high bias ⟶ low variance

highly Non-linear model ⟶ high varience

simple model ⟶ underfit
bcoz it can't learn complex thing

Complex / Non-linear model ⟶ Overfit
bcoz it taking any outlier, noise

Our model have to recongize pattern in data
NOT change prediction According data

⟶ pattern recongization.

Irreducible error:

- can't reduce error whatever you try
- electronic, EMF Noise, thermal noise, Eddy current
  Instrument Noise due to process data Acquistion
- Noise are occure ~~due to process~~ from process which getting us data.

Loss / cost function:-

Mean square loss.

$$L = \frac{1}{N} \sum (y_i - p_i)^2 \quad \text{M.S.E}$$

or $$L = \frac{1}{2N}$$

$$\text{RMSE} - L = \sqrt{\frac{1}{N} \sum (y_i - p_i)^2}$$

Why we take 2nd order ? in MSE

loss f^n $= \underset{w_i}{\operatorname{argmin}} \; L(x, y, w_i)$

optimation problem formulation

I want to find ~~to w~~ for which argument my weight is minimum

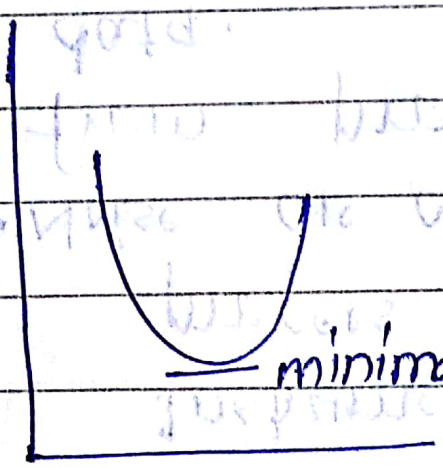
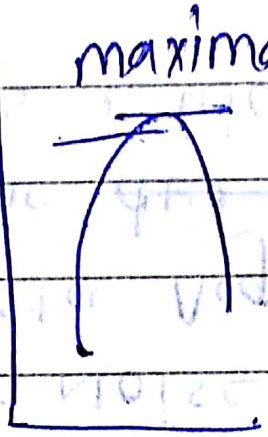
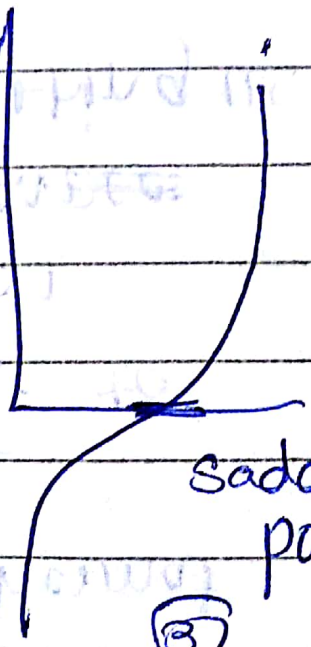


fig. (1)



(2)



(3)

$$L = \frac{1}{N} \sum (y_i - p_i)^2$$

never gives saddle point

if need to ~~point~~ that saddle need to go high order

$$\frac{dL}{dw_i} = 0$$

$$\frac{d^2L}{dw_i^2} = +ve \text{ (minima)}.$$

- using loss fn $\rightarrow \frac{1}{N}\sum(y_i - p_i)^2$ we eliminate (B)
- only need decide (A) & (B)

---

error = $y_i - p_i$ we let say me loss.

$$b_i = (y_i - p_i)^2 \rightarrow$$ this is for single instance

error = $y_i - \{w_0 + w_1 x_i\}$ for ith instant of x_1 feature.

$$\downarrow$$ say $a_{1,i} = x_i$

$$L_i = \left[y_i - \{w_0 + w_1 a_i\}\right]^2$$

$$= y_i^2 - 2y_i(w_0 + w_1 a_i) + (w_0 + w_1 a_i)^2$$

$$\downarrow$$

$$w_0^2 + 2w_0 w_1 a_i + (w_1 a_i)^2$$

$$\frac{dL}{dw_0} = -2y_i + 2w_0 + 2w_1 x_i$$

$$\frac{dL}{dw_0} = -2[y_i - p_i] * (1)$$

$$\frac{d^2L}{dw_0^2} = 2$$

$$\frac{dL}{dw_1} = -2y_i x_i + 2w_0 x_i + 2w_1 x_i^2$$

$$= -2[y_i - w_0 + w_1 x_i] x_i$$

$$= -2[y_i - p_i] x_i$$

$$\frac{d^2L}{dw_1^2} = -2x_i^2$$

ordinary least square (Analytics soln)

SGD.

Prasun goyal

Prasun

computer don't does OLS part.

The equation which have statics solution / deterministic solution

like $y = 2x^2 - 3x + 1$

Soln $y = 0$

So this is Not optimization

* we use optimization when we don't have clear solution about expression.

- for unconstraint optimization we use derivative & double derivate

→ Necessary condition
  ↓
  vanishing gradient
  ↓
  derivative is equal to zero

→ sufficient condition
  ↓
  Next derivate should be +ve or -ve.
  ↓
  if NOT +ve or -ve
  ↓
  if 2nd derivate then ↓ go for 3rd derivate
  ↓
  until you find +ve & -ve values.

for constraint optimization :- ∈ constraint

for equality constraint.

langrang's equation

$$L = f(x) + \lambda_1(g_1(a)) + \lambda_2(g_2(a))$$

→ Now you have $\frac{\partial L}{a}$ $\frac{\partial L}{\lambda_1}$ $\frac{\partial L}{\lambda_2}$

- So Constraint optimization is diffcult than unconstraint.

svm = Maximal Marginal classification

→ which requires optimization

So we are here

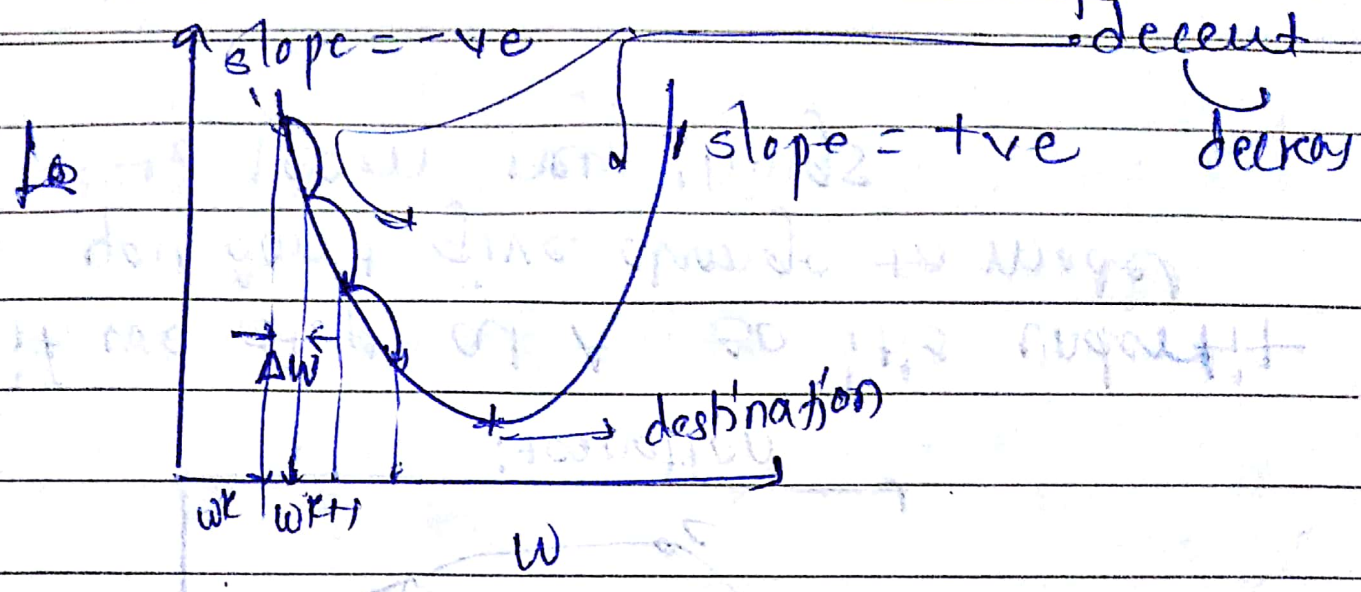
$$\frac{\partial L}{\partial w_1} = -2[y_i - p_i] * x_i$$ features

gradient descent

$$w_i^{k+1} = w_i^k - \Delta w$$

$$\Delta w = +\alpha \left(\frac{\partial L}{\partial w}\right)$$

DATE:



- how to Algorithm know do i have to go forward (-ve slope) or backword (+ve slope) is known by gradient. $\frac{\partial L}{\partial w}$
- and at what step / speed is (α)

$$\Delta W = +\alpha \left(\frac{\partial L}{\partial w}\right) \quad -2[y_i - p_i] x_i$$

step size ← α ; direction ← $\frac{\partial L}{\partial w}$

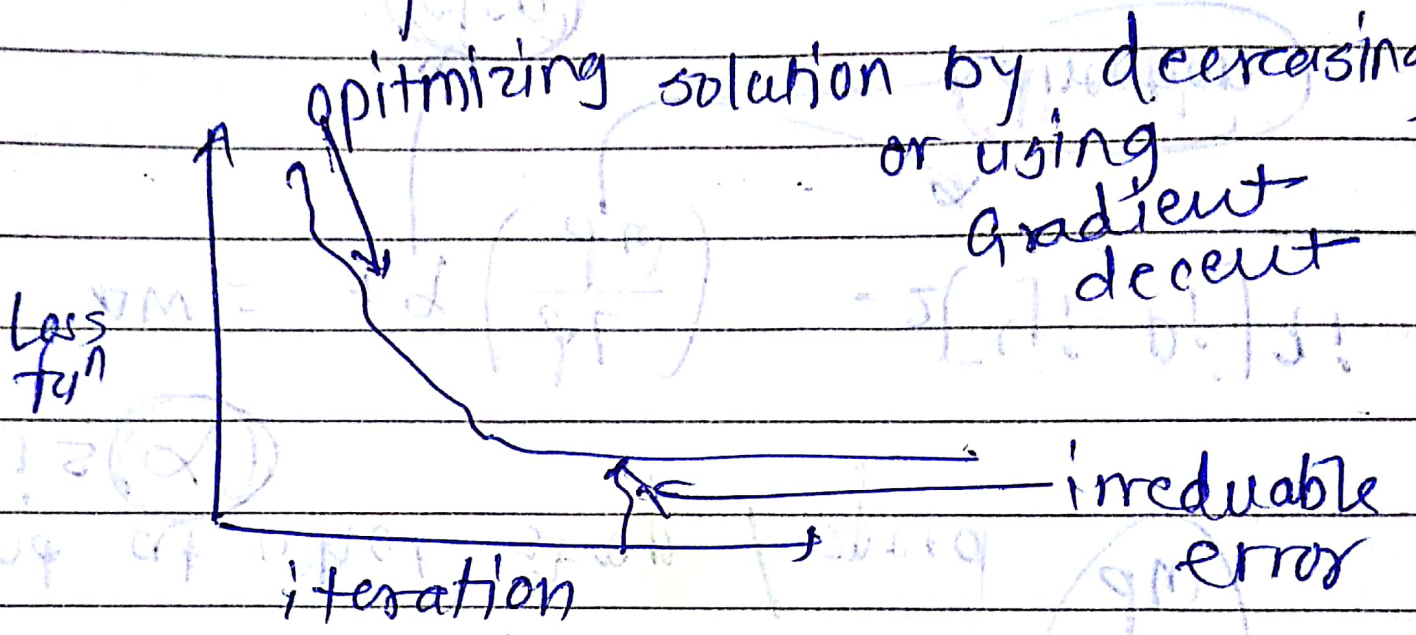
for minimization

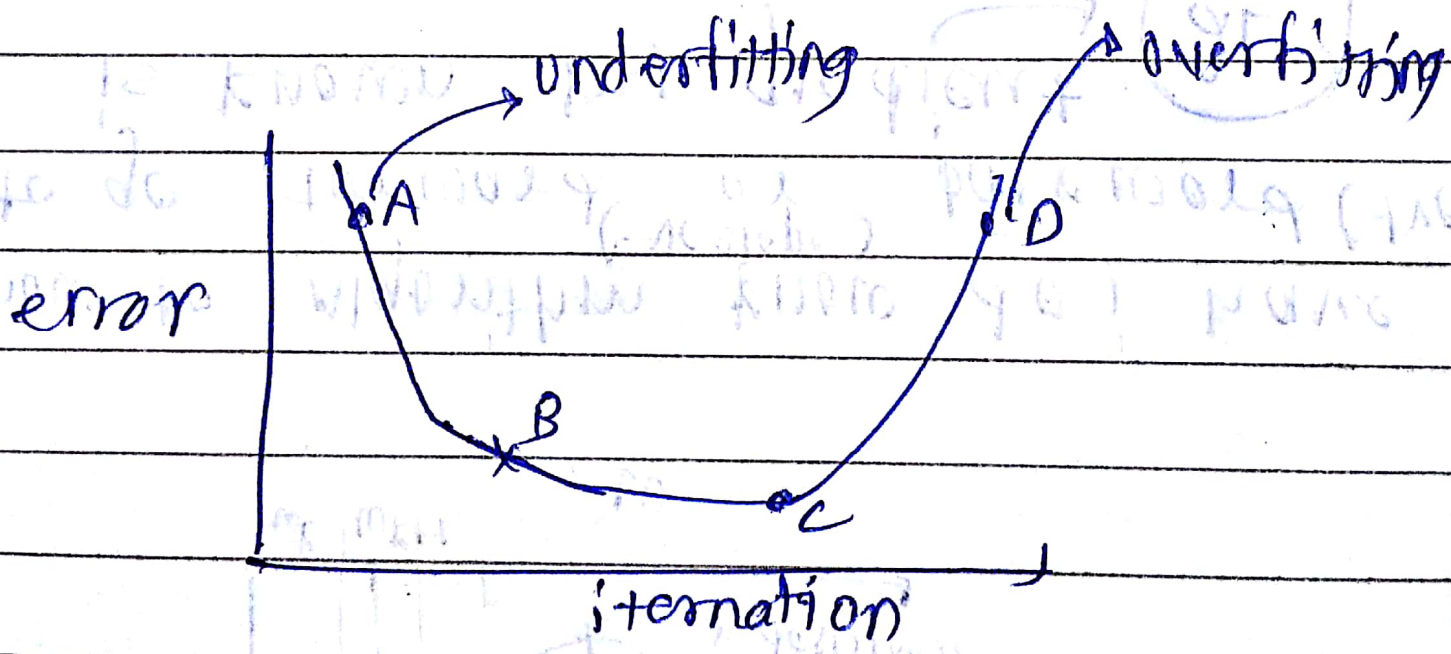
$$w_i^{k+1} = w_i^{k} - \Delta w$$

for Maximization

$$w_i^{k+1} = w_i^{k} + \Delta W$$

This proof the problem is optimization and this use gradient decent.

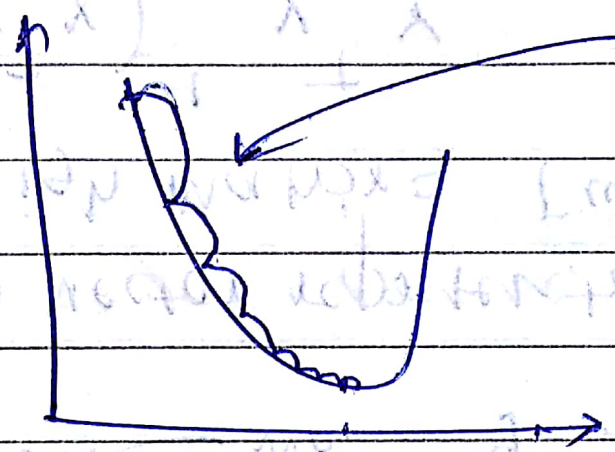




- if we stop at A so it's underfit you don't give change to model to learn new things

- we get at D due to overfitting, Complex model, ~~bata~~ more iteration

α = learning rate.



can we decrease α as we approach minima.

- but gradient descent already doing that ← can we make Adaptive.

↳ because of vanishing slope, as slope ↓es. with constant α we get to minima.

- Until Now we just consider only one instant. ~~to~~ in dataset.

↳ Now take for all dataset

Σ term

for $L = \frac{1}{N}\sum(y - p_i)^2$

$$\hat{w}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

1st variable eg. $\bar{x} = \frac{x_1 + x_2 + x_3}{3}$

2nd variable $\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$

In vector representation

W = weigh Matrix = $[w_1, w_2 \dots]$

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

$$Y = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ w_0 \end{bmatrix} = X^T W$$

$$X = \begin{matrix} w_0 & w_1 \\ \end{matrix} \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}$$

$X^T X = I^2$

$$L = \frac{1}{N} (y - p)^T (y - p)$$

→ similar to $L = \frac{1}{N} \sum (y_i - p_i)^2$

* **logistic regression**: gives linear fun^n

$$\log\left[\frac{p(y)}{1 - p(y)}\right] = w_0 + w_1 x$$

↓ even.

$$p(y) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

#why it called regression if it gives classification ?

Coz we working on number

- we working b/w 0 to 1
- cont'nuous value
- and we checking value is > 0.5 or < 0.5

Homework

prove that $\log\left[\frac{p(x)}{1-p(x)}\right] = w_0 + w_1 x$

given $p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$