Linear Regression

23 July 2020 05:43 PM

Starting @ 6:02 PM

Topic ⇁

→ Simple Linear Regression ←

→ Multiple Linear Regression ←

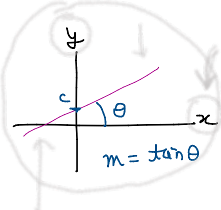
$ax + by + c = 0$ → 2D

Linear Algebra ⇁

Equation of a Line ⇁

$$\boxed{y = mx + c}$$

m ↳ slope, c ↳ intercept



Qu: How many dimensions?

Ans: 2 Dimension

General form (of 2D line) : $\boxed{ax + by + c = 0}$

→ $by = -ax - c$

→ $\boxed{y = -\frac{a}{b}x - \frac{c}{b}}$

Qu: What is the slope & intercept

Ans: $m = -\frac{a}{b}$ intercept $= -\frac{c}{b}$

equation of a line → 2D

$$\boxed{y = \left(-\frac{a}{b}\right)x \left(-\frac{c}{b}\right)}$$

$y = mx + c$

Slope $= -\frac{a}{b}$

intercept $= -\frac{c}{b}$

3D → PLANE

4D → Hyperplane ← 7D, ← nD

6D →, 5D →

Line → 2D

Plane → 3D

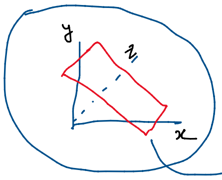
Hyperplane → nD

3-D Equation (Plane) ⇁

General Eqⁿ: $\boxed{ax + by + cz + d = 0}$

$\{by = -ax - cz - d\}$

Gen Eqⁿ : $\boxed{ax + by + cz + d}$



$$\{by = -ax - cz - d\}$$

$$\boxed{y = -\frac{a}{b}x - \frac{c}{b}z - \frac{d}{b}}$$

$-\frac{a}{b}$ → slope with x axis

$\frac{c}{b}$ → slope with z axis

$\frac{d}{b}$ → y-intercept

→ $$\boxed{y = -\frac{a}{b}x - \frac{c}{b}}$$

$-\frac{a}{b}$ → slope with (x)

$\frac{c}{b}$ → y-intercept

2D → General

2D → Line → No. of slopes → 1
No. of intercept = 1

3D → Plane → No. of slopes = 2
No. of intercept = 1

nD → Hyperplane → No. of slopes = n−1
No. of intercept = 1

4D → (Hyperplane) (w, x, y, z)

→ $ax + by + cz + dw + e = 0$

$$y = -\frac{a}{b}x - \frac{c}{b}z - \frac{d}{b}w - \frac{e}{b}$$

$-\frac{a}{b}$ → slope with 'x' axis

$-\frac{e}{b}$ → y intercept

2d dimensional Hyperplane is

$$\{ax + by + cz + dw + e \_ + f \_ + g \_ + \ldots + \_ = 0\}$$

$$\{a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots$$

$\rightarrow$ Coefficients $\rightarrow$ w_i

Axis $\rightarrow$ x_i

$$\cancel{a}\cancel{x} + \cancel{b}\cancel{y} + \cancel{c} = 0$$

2D $\rightarrow$

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

$$\boxed{\sum_{i=1}^{2} w_i x_i + w_0 = 0}$$

3D $\rightarrow$

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0 = 0$$

Plane $\rightarrow$ $\boxed{\sum_{i=1}^{3} w_i x_i + w_0 = 0}$

dot

$$\pi_3 : \boxed{w^T x + w_0 = 0}$$

n-D eq of Hyperplane

$$\rightarrow \boxed{\pi_n : w^T x + w_0 = 0}$$

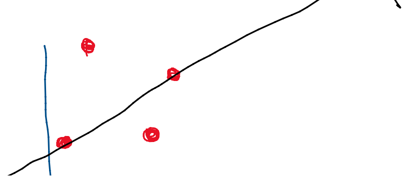
$[\quad\quad] \; [\quad]$

$$w^T x + w_0 = 0$$

Linear Regression Algorithm

Task : find a line that 'BEST FITS' the given data

h	w
—	—
—	—



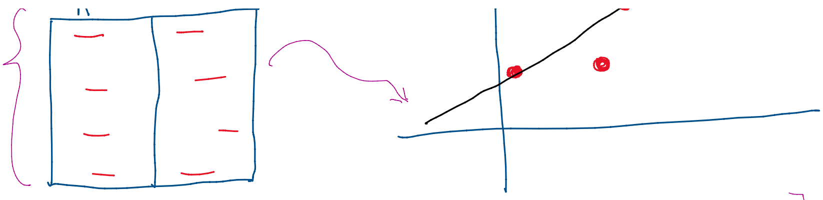
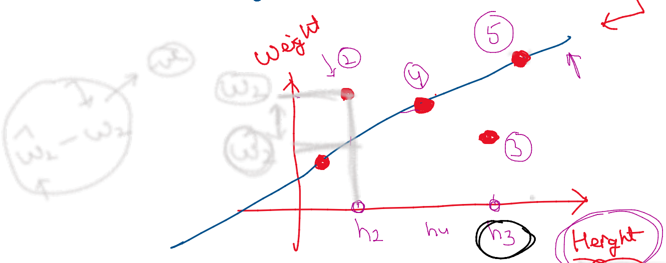
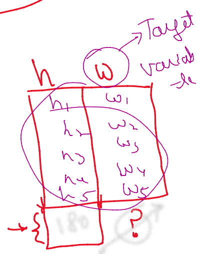


table → Linear Regression → Line → m, c

future x_i → $m x_i + c$ → $\hat{y}_i$ → Prediction

Task: Find slope & intercept (line) that Best fits the given data





Given Data

		input, output	
for 1	→	h_1, w_1	$\hat{w}_1$
for 2	→	h_2, w_2	$\hat{w}_2$
for 3	→	h_3, w_3	$\hat{w}_3$
for 4	→	h_4, w_4	$\hat{w}_4$
for 5	→	h_5, w_5	$\hat{w}_5$

$e_1 = \hat{w}_1 - w_1 = 0$

$e_2 = \hat{w}_2 - w_2$

$e_3 = \hat{w}_3 - w_3$

$e_4 = \hat{w}_4 - w_4 = 0$

$e_5 = \hat{w}_5 - w_5 = 0$

Total error

$\to \cancel{e_1} + \cancel{e_2} + \cancel{e_3} + \cancel{e_4} + \cancel{e_5}$

$= 0$

Total Squared error → continuous

OR

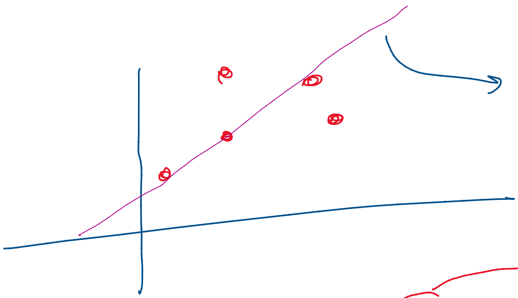
~~Total~~ absolute error (mod) → discontinuous

$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$

$|e_1| + |e_2| + |e_3| + |e_4| + |e_5|$

$\sum^{5} |e_i|$ ← discontinuous

$$\sum_{i=1}^{5} |e_i| \leftarrow \text{discontinuity}$$



$$\rightarrow \sum e^2$$

$$\min \sum_{i=1}^{n} e_i^2$$

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

$y_i \rightarrow$ actual

$\hat{y}_i \rightarrow$ Predicted

Simple Linear Reg →

$$\underbrace{m^*, c^*}_{\text{optimal values}} = \arg\min_{m,c} \sum_{i=1}^{n} \left(y_i - \{mx_i + c\}\right)^2$$

→ Multiple Linear Reg →

Given (h, w, age) → BP

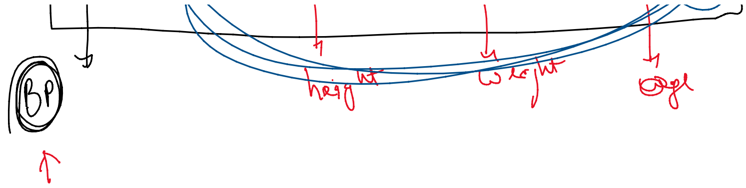
h	w	age	BP

BP → target variable

Hyperplane

$$\underset{h}{w_1 x_1} + \underset{BP}{w_2 x_2} + \underset{\text{weight}}{w_3 x_3} + \underset{\text{age}}{w_4 x_4} + w_0 = 0$$

$$\underbrace{x_2}_{BP} = -\frac{w_1}{w_2}\underbrace{x_1}_{\text{height}} - \frac{w_3}{w_2}\underbrace{x_3}_{\text{weight}} - \frac{w_4}{w_2}\underbrace{x_4}_{\text{age}} - \frac{w_0}{w_2}$$



Task : find slopes & intercept which BEST FITS the given data

Minimum sum of squared errors

$$\min \sum_{i=1}^{n} e_i^2$$

$$w_1^*, w_2^*, w_3^*, w_0, w_4 = \arg\min_{w_1, w_2, w_3, w_0} \underbrace{\sum_{i=1}^{n} \Big(\underbrace{y_i}_{\text{actual}} - \{ \overbrace{w^T x + w_0}^{\text{Pred}} \} \Big)^2}$$

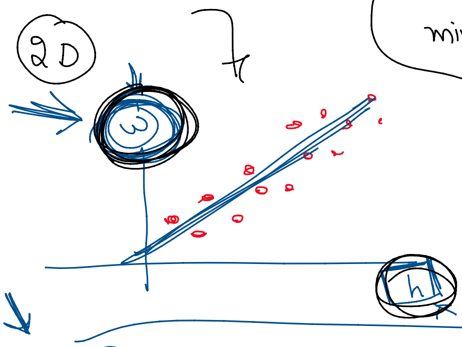
→ Multiple linear Regression

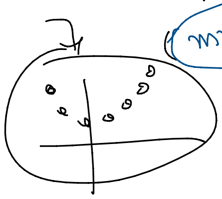
Linear Regression

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

$\hat{y}_i \to (m x_i + c)$

2D





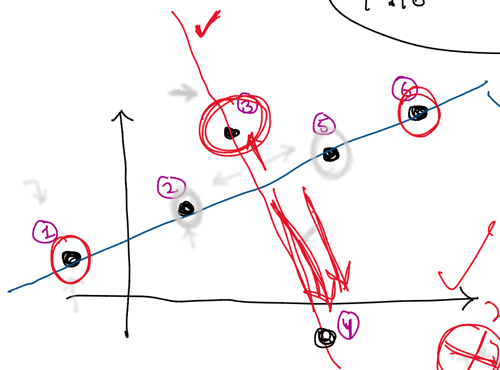
→ Linear Relationship b/w target & input variable

→ Error terms → with 0 mean Normally distributed

Given point → error

Code

7:16 — 7:22



Best fit line

$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 + e_6^2$

for Red line → $\sum e^2 \Rightarrow 0$

for Blue line → $e_1 + e_2 + e_3 + e_4 + e_5 + e_6$

→ $\sum e = 6$

$\min \sum e$ → Red line is best

Linear Reg → We are trying to find ~~best~~ w & w_0 which minimized Squared sum of errors

Task → find a line that BEST FITS the given data

$$w^*, w_0^* = \arg\min_{w, w_0} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\}$$

Prediction

Data → Machine Learning Algo → Prediction/ Function

Train

Test

future
↳ future datapoint → pred/funct → output

↑
client

1 → Read the data → read_csv

x	y
X_train	y_train
X_test	y_test

Training

Test

$\hat{y}$_test

actual

2 Split the data in train & test

3 Training phase

X_train, y_train

Train data → ML Algo → Predictor function

Test → X_test → Pred funct → $\hat{y}$_test

---

1 Read / Load the data — Read_csv

2 Split the data {X_train y_train X_test y_test}

X_train	y_train
X_test	y_test

X_train, y_train → ML algo → function

X_train, y_train → ML algo → {function}

X_test → {function} → $\hat{y}$_test

Evaluate → y_test − $\hat{y}$_test → error

→ Starting @ 4:02 PM

Topic ↓

→ Linear Regression (Handson)

→ Case Study on House Price prediction

Participation Certificate + 4* OR 5* certificate

4*	5*	
80%	90%	→ attendance
70% in each	80% in each	→ Marks
1 late submission is allowed	1 late submission is allowed	

Linear Reg →

Task : find a line that best fits the data.

Assumptions →

1. Linear Relationship b/w target & input variables
2. Error Terms should be normally distributed with mean 0.

X_train ← y_train

X_test ← y_test

Training X_train, y_train → Linear Regression → Model / function Predictor

Testing X_test → Predictor Model function → y_test_pred

Evaluation → Mean Squared error

$$\frac{\sum (\text{Actual} - \text{Pred})^2}{\text{No. of test data point}}$$

y_test y_test_pred

$$\frac{\sum_{i=1}^{n} (y\_test_i - y\_test\_pred_i)^2}{n}$$

where n → No. of test data points

MSE → Lesser is Better

Evaluation → Root Mean Squared error →

$$\sqrt{MSE}$$

Evaluation → Mean Absolute Error →

$$\frac{\sum |\text{Act} - \text{Pred}|}{n}$$

→ CRISP-DM Framework →

{Cross industry Standard process for Data Mining}

ML problem or DS problem or DA or DNS

Data Mining

ML problem OR DS problem or DA or DM

Iris

Medical

1-2 months

1. Bussiness Understanding

2. Data Understanding

3. Data Preparation

4. Train the Model

5. Predict & Evaluation

6. Development & Deployment

EDA

- Norm/Stand numerical columns
- Convert categorical cols to numerical colms
- Train test split

2 day

Data → Linear Reg → Prediction

Mathematics

h	w	gender
		Male
		Female
num	num	cat

cat → Object → Numerical

Numerical data

Ohe

h	w	BP	Age	Cancer
				Yes
				No.
cm	kg	mm	years	

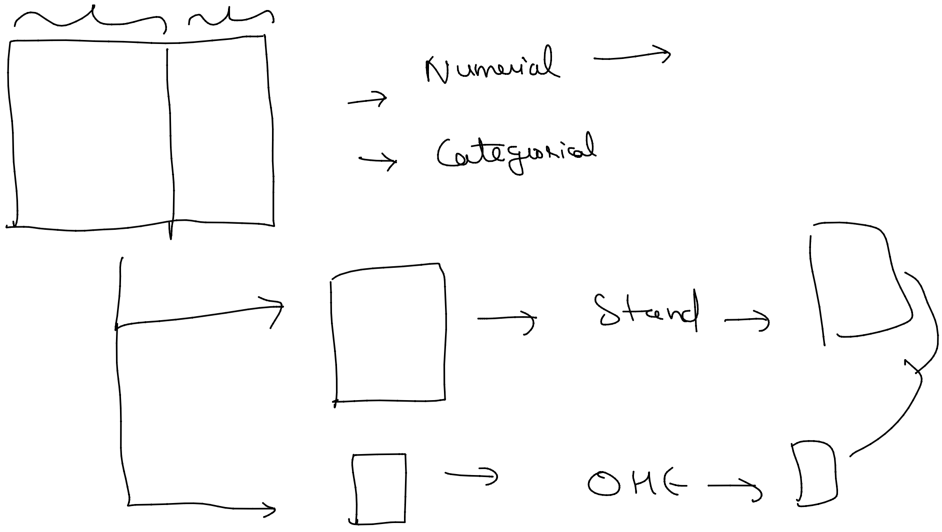
→ Remove the scale dependency of each numerical column

ML DL

→ ~~Normalization~~ ×

→ Standardization ✓

→ Numerical →



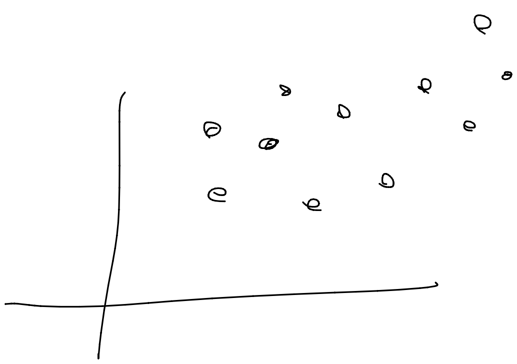
→ Train test Split

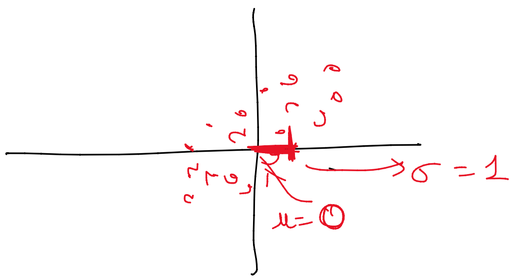
train data → Model

test → Pred + Eval

05:15 — 05:20

Standardization → $\mu = 0$, $\sigma = 1$





---

for

for Linear Regression → Assumption

1. Linear Relationship b/w target & input data
2. Error terms → Normally dist