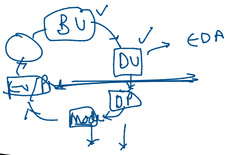
Logistic Regression

27 July 2020 03:58 PM



Starting @ 4:02 PM

→ Logistic Regression ←

Linear Reg

→ Project / Case Study on Car Price Prediction (EDA)

T → Regression → Price → Real / Continuous variable

→ Classification

Logistic Regression Algorithm

→ Classification ←

→ Supervised Learning Technique ←

Species y_i

$$D_n = \{ (x_i, y_i) \mid x_i \in \mathbb{R}^d, \; y_i \in \{+ve, -ve\} \}$$

y_i → target variable

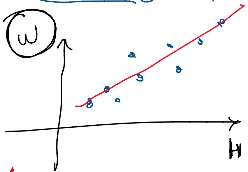
Binary classification Problem

y_i → More than 2 classes (eg: Iris Dataset → species)

Multi class Classification

Linear Reg ⟩

Target Var: Weight

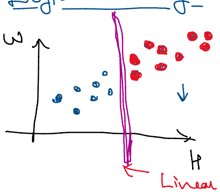


Task → Find a line that BEST FITS the data

↳ Min Total error

Logistic Reg ⟩

Target V: Gender

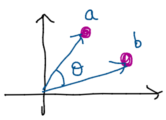


Task → Find a line that BEST SEPERATE the +ve from -ve data point

→ Min no. of misclassified pts

OR Max no. of correctly classified pts

Linear Algebra ⟩



Dot Prod

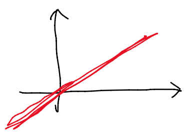
$$a \cdot b = a^T b$$

$$= \|a\| \, \|b\| \cos\theta_{a,b}$$

$$= \sum_{i=1}^{d} a_i b_i$$

$$y = mx + c \rightarrow y = mx$$

Passes through the origin



General form ⟩

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

2D → $$x_2 = \frac{-w_1 x_1}{w_2} \boxed{- \frac{w_0}{w_2}}$$

c → 0, intercept term

If line passes through origin ⟩ $$-\frac{w_0}{w_2} = 0$$

$$\boxed{w_0 = 0}$$

$\boxed{w_0 = 0}$ ✓

→ General eqⁿ of a line that passes through origin ↓

(2D) → $w_1 x_1 + w_2 x_2 = 0 \Rightarrow \sum_{i=1}^{2} w_i x_i = 0$

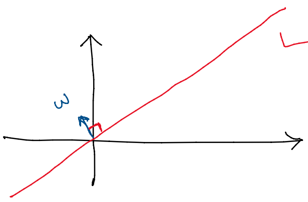
$\pi_2 : w^T x$

→ General eqⁿ of a hyperplane that passes through origin ↓

(4D) → $w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 = 0$

$\Rightarrow \sum_{i=1}^{4} w_i x_i = 0$

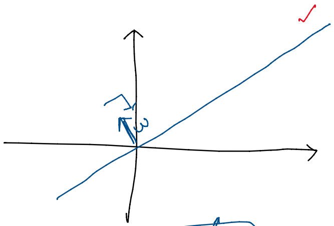
$\Rightarrow \pi_4 : w^T x$ → norm of a plane



→ $\boxed{\pi_d : w^T x}$

OR

$\boxed{\sum_{i=1}^{d} w_i x_i}$



↳ $\boxed{\theta_{w,x} = 90^\circ}$

$\boxed{w^T x = 0}$

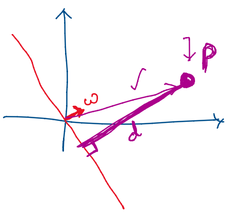
→ $\boxed{w \cdot x = 0}$

$\|w\| \, \|x\| \cos\theta_{w,x} = 0$

+ve, +ve, 0

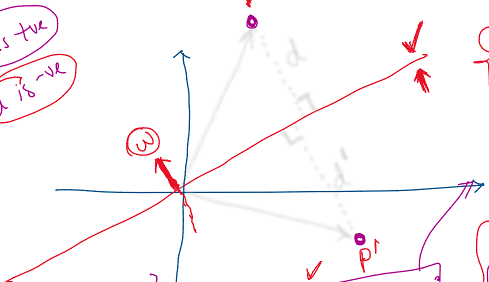
$\boxed{\cos 90 = 0}$

Distance of a point from a Plane ↓



$$\boxed{d = \frac{w \cdot P}{\|w\|}}$$ ✓

↓



d is +ve

d is -ve

$d = \frac{w \cdot P}{\|w\|}$ (+ve)

$= \frac{\|w\|\,\|P\|\cos\theta}{\|w\|}$ (+ve)

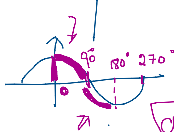
$\boxed{0^\circ \le \theta_{w,P} \le 90^\circ}$

$\boxed{90^\circ \le \theta_{w,P'} \le 180^\circ}$

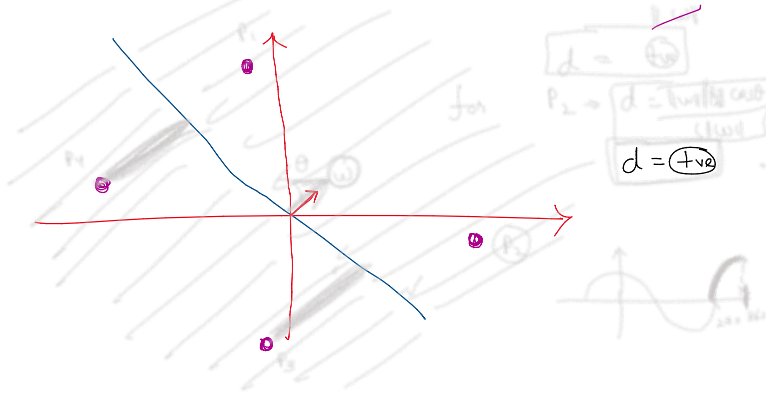
neg

$d = \frac{w \cdot P'}{\|w\|}$ (-ve)

$= \frac{\|w\|\,\|P'\|\cos\theta_{w,P'}}{\|w\|}$ (+ve)



$\cos\theta$ is -ve



$d = (+ve)$

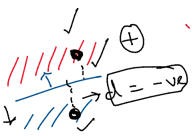
---

1. If a π passes through the origin
$\pi_d : w^T x = 0$, where $w_0 = 0$

2. Distance of a point 'P' from a plane 'π'

$$d = \frac{w \cdot P}{\|w\|}$$

3. If a point 'P' lies on the same size of the plane, then the distance of point from plane will be +ve

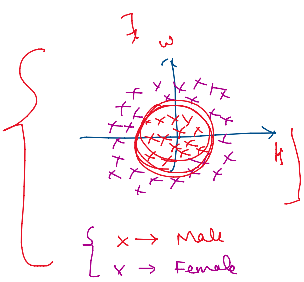


$\boxed{d} = \begin{cases} \boxed{+ve}, & \text{if P is in the same direction of plane} \\ \boxed{-ve}, & \text{if P lies on the opposite direction of plane} \end{cases}$

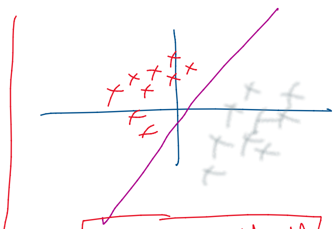
Logistic Regression :

{ Task → Find a line that BEST SEPERATES +ve's from -ve's.

{ Assumption → * Classes should be perfectly / almost linearly seperable.



(Can't use logistic because assumption doesn't hold true)



Linearly seperable

(Can use Logistic Reg)

Linear Reg	Logistic Reg
• Task → Find the $\boxed{\text{BEST FIT}}$ line ✓ $\min \sum e^2$	• Task → Find the $\boxed{\text{BEST}}$ $\boxed{\text{SEPERABLE}}$ Line
• Assumption : 1. Linear Relationship ✓	• Assumption : 1. $\boxed{\text{Classes}}$ → linearly seperable

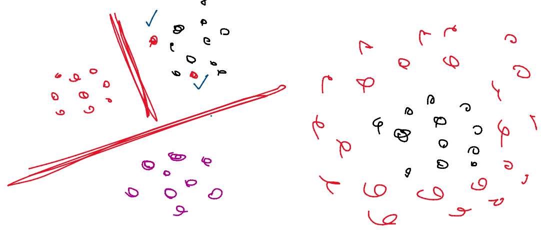
- Assumption ⌉
  1. Linear Relationship ✓
  2. Normal dist of errors ✓
- Regression Task
- Supervised Learning

- Assumption ⌉
  1. Classes → linearly seperable
- Classification Task
- Supervised Learning

Starting @ 4:02 PM

→ Logistic Reg (Cont..)

↳ Case Study (Cont..)



Logistic Regression ⌉

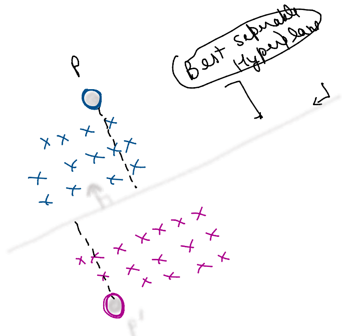
Given → $D_n = \{ (x_i, y_i) \mid x_i \in \mathbb{R}^d, \; y_i \in \{+ve, -ve\} \}$

Task → Find a line that BEST SEPERATES the class from -ve class.

Assumption → Classes should be almost | perfectly linearly seperable.

✓ Assumptions while deriving Logistic Reg ⌉

1. The normal of the plane is always going to be a unit vector. $\|w\| = 1$
2. Classifier / Predictor will pass through the origin



$$d_i = \frac{w \cdot x_i}{\|w\|}$$

Because of Assumption - 1

$$\boxed{d_i = w \cdot x_i} = \|w\| \|x_i\| \cos\theta$$

- $\boxed{+ve}$ → pt lies in the same dir of the plane
- $\boxed{-ve}$ → pt. lies on the opposite side

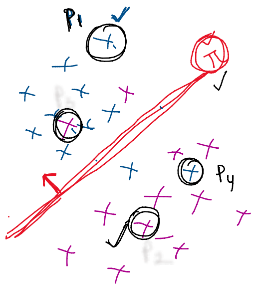
Line

actual ↑ $\boxed{y_i}$ * $\boxed{\hat{y}_i}$ ↑ pred actual

$\boxed{y_i} * \boxed{\hat{y}_i}$

$\boxed{w^T x_i} \to d_i$

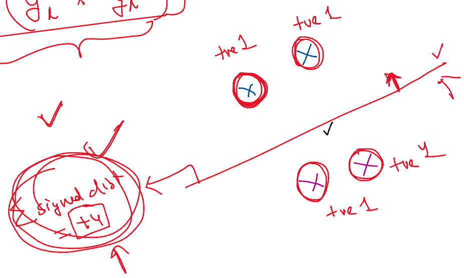
Line

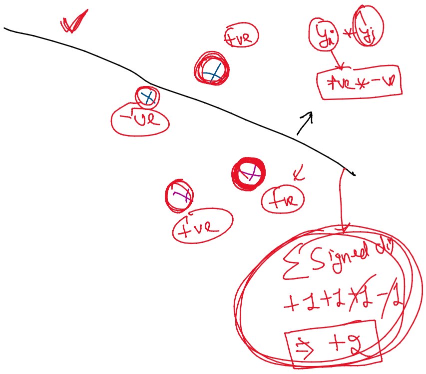


$\boxed{y_1 * \hat{y}_1} + \boxed{y_2 * \hat{y}_2} +$

$\boxed{y_3 * \hat{y}_3} + \boxed{y_4 * \hat{y}_4}$

$\boxed{\max \sum (y_i * \hat{y}_i)}$





Linear Reg ⟧

Task → Find $\boxed{\text{BEST FIT}}$ Line

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

$\boxed{\min \sum (y_i - w^T x_i)^2}$

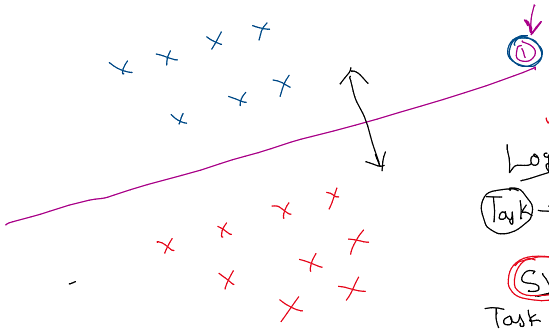
Logistic Reg

Task → Find $\boxed{\text{BEST SEPERABLE}}$ Line

→ maximize correct classified pt.

$\boxed{\max \sum (y_i * \hat{y}_i)}$

$\boxed{\max \sum (y_i * (w^T x_i))}$



Logistic Reg ⟧

Task → Find best seperable line

SVM ⟧

Task → Find best seperable line that $\boxed{\text{maximizing the margin}}$

$\boxed{x_i} \boxed{y_i}$

Logistic Reg ⟧ (Classification)

sum of signed distance

$$\left\{ w^* = \underset{w}{\arg\max} \left\{ \sum (y_i * w^T x_i) \right\} \right.$$

Linear Reg ⟧ (Regression) sum sq error

$$(\qquad (\sum (\qquad - \qquad)^2)$$

Linear Reg → (Regression) sum sq error

$$w^* = \arg\min_w \left\{ \sum \left(y_i - w^T x_i\right)^2 \right\}$$

$-1 * -1$

sum of signed distance $= +1 +2 +3 +4 +5 -1 -2 -3 -4 -5 +1 = +1$

← outlier

$y_i * \hat{y}_i$ → distance

$w^T x_i$

$\hat{y}_i$

100

1

w

$\{ = -90 \}$

$-1 * -1$

$x \to +1$
$x \to -1$
→ y_i

1

2

3

4

5

$\sum$ Signed dist $= +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 -100 = -90$

$-1 * +1$

$$w^* = \arg\max_w \left\{ \sum \left(y_i * \hat{y}_i\right) \right\}$$

$$\boxed{w^* = \arg\max_w \left\{ \sum \left(y_i * w^T x_i\right) \right\}}$$

Problem with above formulation is because of OUTLIER

identify the outlier

central tendency → mean, median

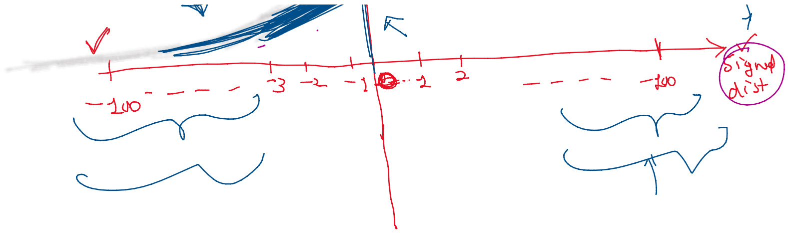
Signed distance → $y_i * \hat{y}_i$ OR $y_i * w^T x_i$

$f(y_i * \hat{y}_i) =$

Sigmoid function

-3 -2 -1 0 1 2 ---- +100

Signed dist



Logreg ⇁

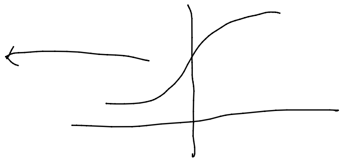
$$w^* = \arg\max_{w} \left\{ \sum (y_i * w^T x_i) \right\}$$

Problem → Impacted because of outliers

$$w^* = \arg\max_{w} \left\{ \sum \sigma (y_i * w^T x_i) \right\}$$

Sigmoid ⇁

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



$$0 \leq \sigma(x) \leq 1$$

$$\boxed{w^* = \arg\max_{w} \left\{ \sum \left(\frac{1}{1+e^{-(y_i * w^T x_i)}} \right) \right\}}$$

OR

$$w^* = \arg\max_{w} \left\{ \sum \left(\frac{1}{1+\exp(-y_i w^T x_i)} \right) \right\}$$

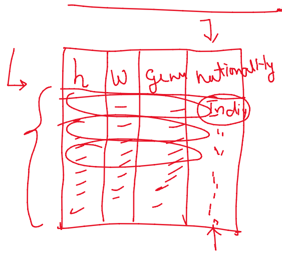
Less impacted by outliers

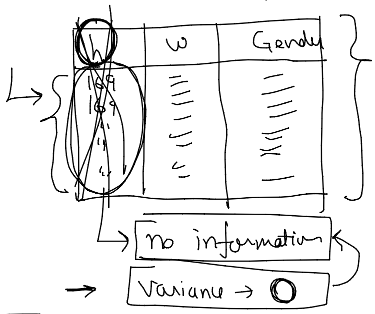
Starting @ 4:02 PM

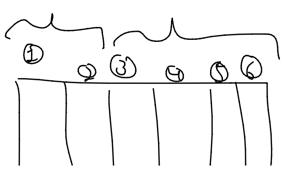
Topics ⇁

- → Logistic Regression (Code)
- → Dimensionality Reduction Techniques
- → Car Price Prediction (Case study Cont.)

→ Dimensionality Reduction ⇁ ⇁







Variance

$\sigma^2 = 1000$ $\sigma^2 = 100$ 0.5 1 0.1

Variance

0.5

6

Dimensionality Reduction

→ Variance Threshold

a b c d e f g — columns — 7 column

target — 1 column

RFE — Automatic feature elimination — top n cols

→ Forward Feature Elimination
→ Backward " "
} Manual

7 column → target → Accur
→ 6 colum → target → Acur_1
→ 6 colum → target → Acur_2
→ 6 Colum → target → Acur_3

max

a → target → Acc
b → target → Acc
c → targ → Acc
d → targ → Acc
e → targ → Acc
f → targ → Acc

max

2 Best feat/column

ca b/d/e/f

3

c, a, d

→ c, a → targ → Acc
c, b → targ → Acc
⋮

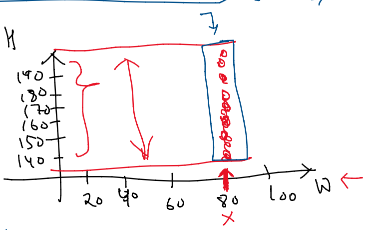
5. Principal Component Analysis

scatterplot

2D

H	W
170	80
160	80
150	80
140	80
190	80
180	80

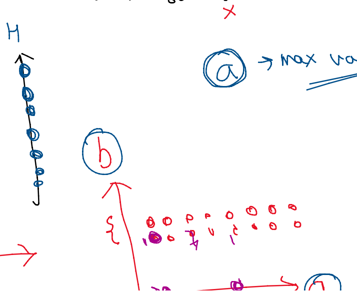
No variance → drop

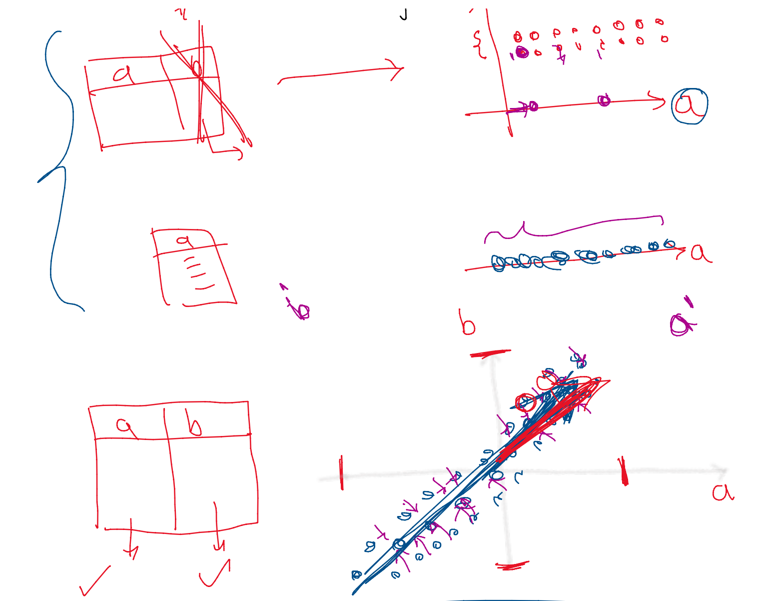


1D

H
170
160
150
140
190
180

a → max variance



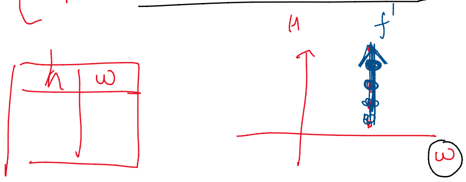


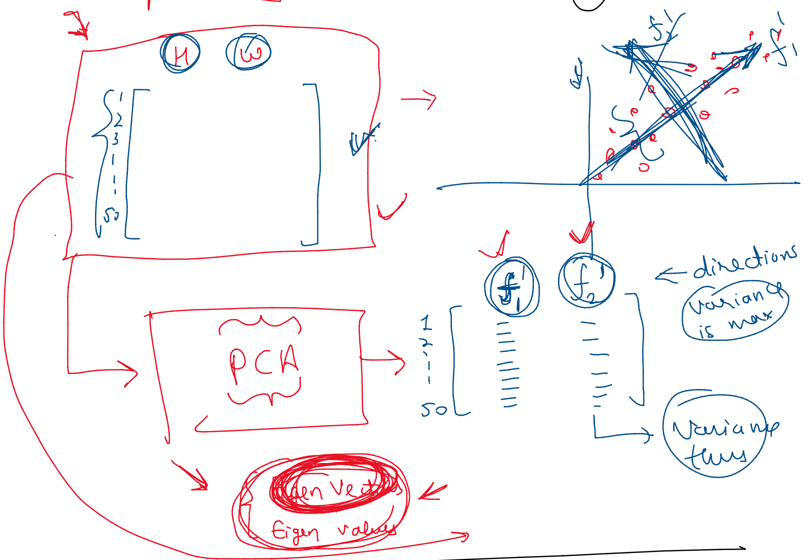
Find the direction where Variance is maximum

Eigen vectors

PCA →

Task → We want to find the direction f' such that the variance of x_i's projected onto f' is maximum





t-sne → t-distributed Stochastic neighbourhood embedding

(Code → PCA

Code → PCA

Corr Pred → ~~RFE~~ → PCA