So Far

1. Linear Regression
2. Polynomial Regression
3. Ridge
4. Lasso
5. Elastic Net

Supervised

Regression ⇔ Conti

Supervised

Regression

Target
↓
continuous

Classification

Target
↓
discrete

Example

Email < spam
not

Transaction < Fraud
not

Gender < male
Female

Sentiment < positive, Negative

Healthcere < Cancer, not

Binary, Multi label

Classification

Logistic Regression

↳ misleading

linear regression

$$\text{Continous outcome} = \beta_0 + \beta_1 x + \cdots + \varepsilon$$

range = $(-\infty, +\infty)$

$0 \le >$

How? probabity $0 - 1$

range

$(0, 1)$

logistic regression → $(0, 0.1, 0.2, \ldots 1)$

$0 - 0.5$, $0.5 - 1$

↓ ↓

$(0 \quad 1)$

$0.2 \quad 0.8$

$$\boxed{y = \beta_0 + \beta_1 x}$$ Linear Regression

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$ ✓ Logistic Regression

1. Input = $\underline{(-\overset{\checkmark}{\infty}, +\overset{\checkmark}{\infty})}$

2. $(0, \infty)$ $e^{\dots} =$ ✓

exponential ⟷ $(0, \infty)$ ⟷

3. $(0, \infty)$

↕

$(0, 1)$

$$\boxed{\frac{p}{p+1}}$$

$p = 561$

$$\boxed{= \frac{561}{561+1}}$$

$(0, 1)$

$$y = \beta_0 + \beta_1 x$$

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{\exp(\beta_0 + \beta_1 x) + 1}$$

$$p = \frac{e^{\wedge} y}{1 + e^{\wedge} y}$$

$p =$ probablity

$q = 1 - P$

$$q = 1 - p = 1 - \frac{e^{\wedge} y}{1 + e^{\wedge} y}$$

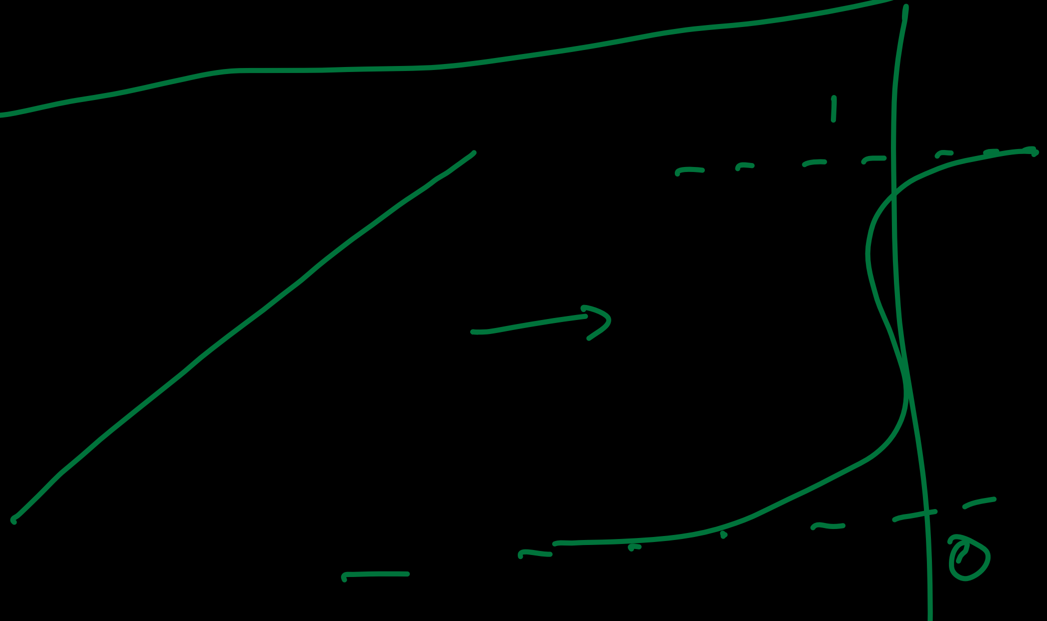
$$p = \frac{e^{y}}{1 + e^{y}}$$

$$p = \frac{\frac{e^{y}}{e^{y}}}{\frac{1 + e^{y}}{e^{y}}} = \frac{1}{\frac{1}{e^{y}} + 1}$$

$$\boxed{p = \frac{1}{1 + e^{-y}}}$$

Sigmoid function

$$\boxed{p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}$$



p $\quad 1 - p = q$ ✓

India vs pakistan

100 → 80, 20

$$\text{Odds} \rightarrow \frac{p}{1-p} = \quad \frac{80}{20} = 4$$

range $(0, \infty)$

$$\frac{p}{1-p}$$

$$p = \frac{\frac{1}{1+e^{-y}}}{1 - \frac{1}{1+e^{-y}}}$$

$$= \frac{\dfrac{1}{1 + e^{-y}}}{\dfrac{\cancel{1} + e^{-y} - \cancel{1}}{1 + e^{-y}}}$$

$$= \frac{1}{\cancel{1 + e^{-y}}} \times \frac{\cancel{1 + e^{-y}}}{e^{-y}}$$

$$= \frac{1}{e^{-y}}$$

$$\boxed{\frac{p}{1-p} = e^{y}} \qquad (0, \infty)$$

Odds

exp => Anti log

apply log on each side

$$\boxed{\log\left(\frac{p}{1-p}\right) = y}$$

$$\boxed{\log\left(\frac{p}{1-p}\right) = \underline{\beta_0 + \beta_1 x}}$$

$$\frac{1}{1+e^{-y}}$$

$\underline{\text{log odds}}$ = linear regression

logit

$(-\infty, +\infty)$

p = probablity $\longrightarrow$ $(1, 0)$

$\frac{p}{1-p}$ = odds $\longrightarrow$ $(0, \infty)$

$\log\left(\frac{p}{1-p}\right)$ = logits or log-odds $\longrightarrow$ $(-\infty, +\infty)$

$$\boxed{\log\left(\frac{p}{1-p}\right) = y}$$

apply exp on both

$$\frac{p}{1-p} = e^{y}$$

$$p = (1-p)\,e^{y}$$

$$p = e^{y} - pe^{y}$$

$$p + pe^{y} = e^{y}$$

$$p(1+e^{y}) = e^{y}$$

$$p = \frac{e^{y}}{1+e^{y}}$$

$\div\, e^{y}$ on num
denon

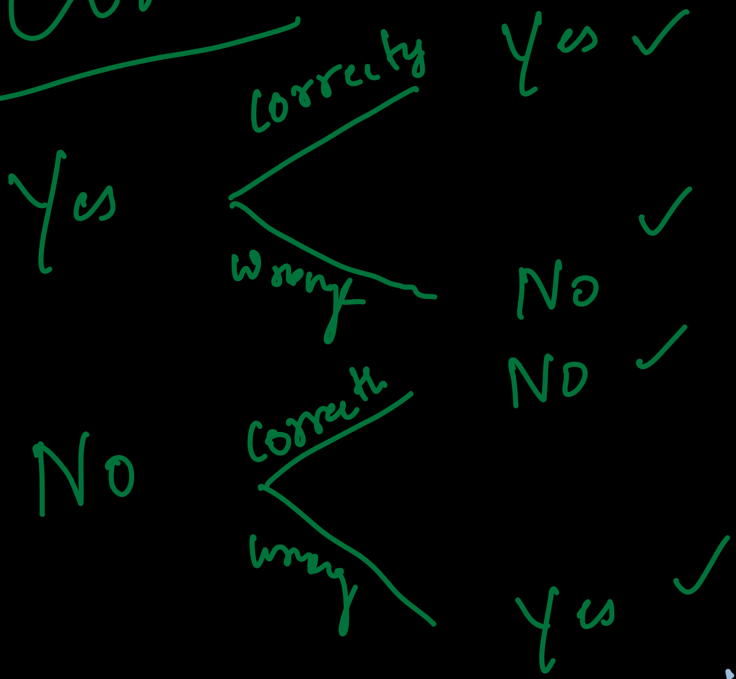
$$p = \frac{\dfrac{e^{y}}{e^{y}}}{\dfrac{1+e^{y}}{e^{y}}} = \boxed{\frac{1}{1+e^{-y}}}$$

Sigmoid function

$$\boxed{\text{Logistic Regression} = \frac{1}{1 + e^{-(\text{Linear Regression})}}}$$

Classification

Metrics



		Prediction class	
		Yes	No
Actual class	Yes	✓	✗
	No	✗	✓

		Prediction class	
		Yes positive	No. ✓ Negative
Actual class	Positive ✓ Yes	✓ Truly Predicting as positive	✗ Falsely Predicting as Negative ✓
	Negative No	✗ Falsely Predicting as positive	Truly Predicting as Negative ✓

wrongly Predicting as NO'

wrongly predicting as 'Yes'

		Predicting class	
		Yes	No
Actual class	Yes	TP	FN
	NO.	FP	TN

Confusion Matrix

		Predicting Class	
		Yes	No
Actual class	Yes	TP	FN
	NO	FP	TN

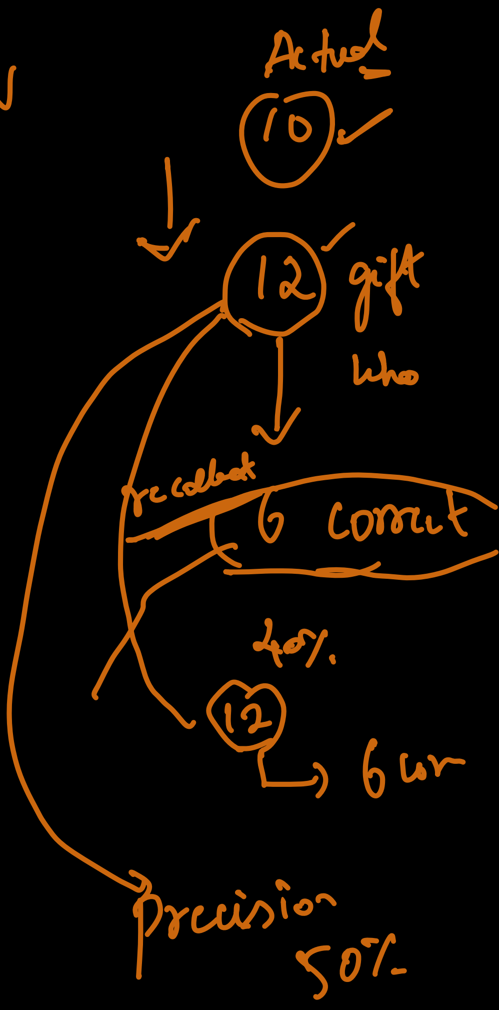
Actual whole Positive — Recall (or) Sensitivity

Actual whole Negative — Specificity

Predicted whole positive Precision

Predicted whole Negative

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



Recall = Based on acutual data

(True positive rate)

(True Negative rate) $= \dfrac{TP}{TP + FN}$

On your Actual data How much you recalled correcty

Precision = Based on Predicted data

$$= \frac{TP}{TP + FP}$$

10,000 — 9990 → Non Cancer ✓ } Training

10 → Cancer ✓

10,000 → Non Cancer → Model Prediction

↳ Accuracy of this model

99.9% ✓

↳ Too good ✓

Worst

Threating → Murders to people

Precision ✓

		Predicting	
	Yes	NO	
Actual Yes	1	9	10
NO	0	9,990 TN	9,990
	1 →	9,999	10,000

$$\text{Acc} = \frac{0 + 9,990}{0 + 9990 + 10 + 0} = \frac{9990}{10000}$$

$$\Rightarrow \underline{99.9\%} ✓$$

$$\text{Precision} = \frac{1}{1+0} = 1.00\%$$

$$\text{Recall} = \frac{1}{1+9} = \frac{1}{10} = 0.1$$

$$= 10\%$$

Security

On a day

10 people → Caught

6 people → prohibited item

How precise

Predicted

$$\frac{6}{10} \Rightarrow \underline{60\%}$$

40% → innocent

---

20 people have prohibited

actual

$$\frac{6}{20} = 30\%$$

recall

Stop everyone

recall ↑ Precison ↓

Trade off

Stop very very sure

recall ↓ precion ↑

Model Screens for a Cancer

High Recall Priority (Don't miss the cancer patient)

Some healty patient will be told to do more test

But you catch almost everyone who actually have cancer

High Precision Priority (Don't scare healty people)

Only flag patients when I have more confident

Fewer healty people get worried unnecessarily

But you might miss - some early stage cancer people

Imagine 1000 emails

700 → legitimate

300 → spam

Scenario A (Catch all spam)

TP = 300 TPR = 100%

FP = 100 ✓

$\frac{600^{TN}}{700}$ = ~~0.86~~ TNR
85%

		Predicted class		
		Spam	Not spam	
Actual class	Spam	300	0	=300
	Not spam	100	600	700
		400	600	1000

$$\frac{600}{100+600} = \frac{600}{700} = 0.85$$

85% <u>TNR</u>

Perfecly catching the Spam but lost 100 important emails

Scenario B: (Protect important mails)

$TN = 690$

$$TNR = \frac{690}{700} = 98\%$$

FP = 10

TP = 210

		Predicter Spam	Not spam	
Actual	Spam	210	90	300
	Not spam	10	690	700
		220	780	1,000

$$TNR = \frac{690}{690+10} = 98\%$$

$$TPR = \frac{210}{300} = 70\%$$

90 mails missed from spam

10 important classified as spam

Metrics of Classification :-

1. Confusion matrix ✓
2. Accuracy ✓
3. Recall or Sensitivity ✓
4. Specificity ✓
5. Precision ✓
6. F1-score ✓
7. ROC-curve
8. AUC-ROC