Statistics Part II

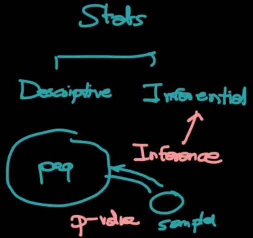
- Hypothesis Test (Inference)
- Binary Classification 0/1
- Clustering → Kmeans ✓ set.seed( )

kmeans( ) ↑

ML

Supervised (Prediction)

Unsupervised (Summarise) ✓

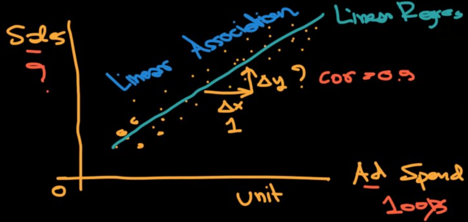


Hypothesis Testing

1. Comparison : AB Test

2. Association (Pearson Correlation)

3. Prediction : Regression

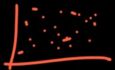


Correlation does not imply causation

[illegible] !

+

Strength



$\text{Sales} = f(\text{Ad})$

$\text{Sales} = 500 + \boxed{50} \cdot \text{Ad}$

$\text{Slope} = \frac{\Delta y}{\Delta x}$

Ronald Fisher (1925) → SPSS Frequentist Approach

Identical Experiment 20 ครั้ง → $1/20$ = 5% → Reject Hyp.

Arbitrary 1% 5% 10%

↙ $P(HHH....95$ ครั้ง$)$

100 ครั้ง

Fair Coin 50% (0.5)

$\frac{1}{2} = 0.5$

Low Chance

↓ $p(H\ 95\% \mid \text{Fair coin})$

$\text{p-value} \leq \underline{5\%}$

$$P(H \geq 95 \mid \text{Fair Coin})$$ ✓

$0.0001 \leq 0.05$

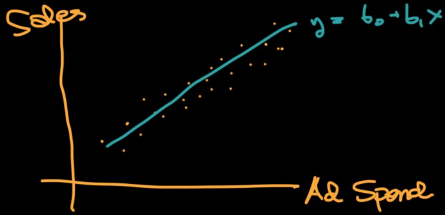
~~H_0: Fair Coin~~

✓ H_a: Not a Fair Coin

Fisher

Reject H_0 : $\text{p-value} \leq 0.05$

Fail to reject H_0 : $\text{p-value} > 0.05$



Significance Test

$$\text{Sales} = f\,(\text{Ad})$$

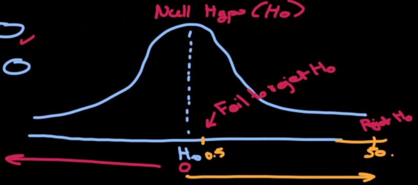
$$\text{Sales} = b_0 + b_1\,\text{Ad}$$

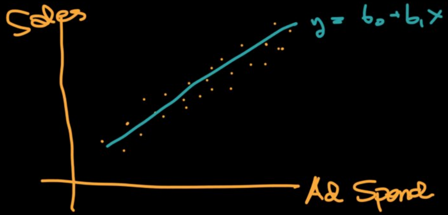
$$\text{Sales} = 500 + 50.\,\text{Ads}$$

Two tailed.

$$H_0:\ b_1^{\text{Ad}} = 0$$

$$H_a:\ b_1 \neq 0$$





Significance Test

$$\underline{\text{Sales}} = f\,(\underline{\text{Ad}})$$

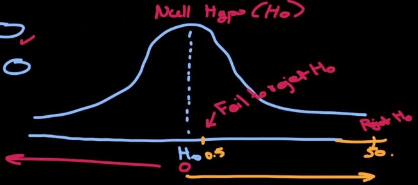
$$\text{Sales} = b_0 + \underline{b_1}\,\text{Ad}$$

$$\text{Sales} = 500 + \boxed{50.}\,\underline{\text{Ads}}$$

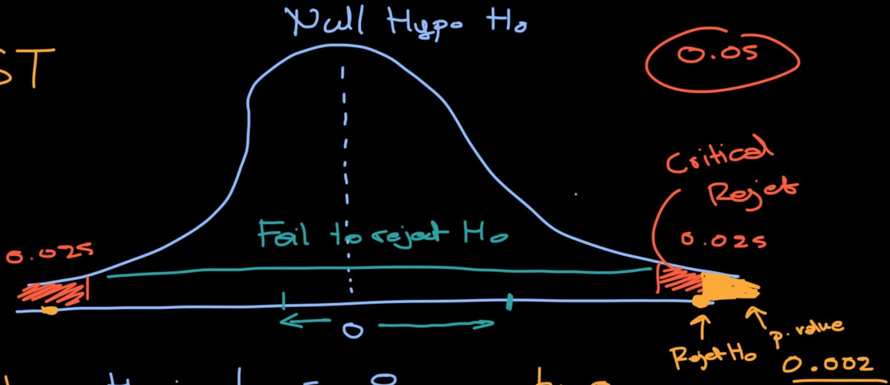
Two tailed.

$$H_0: \overset{Ad}{b_1} = 0$$

$$H_a: b_1 \neq 0$$



NHST

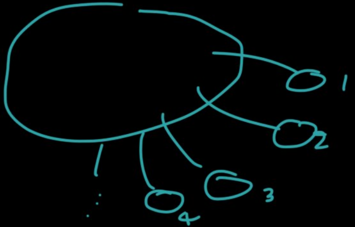


Two tail $H_0 : b_1 = 0$ $b > 0$

$H_a : b_1 \neq 0$ $b < 0$ ≤ 0.05

1. p.value $\leq$ alpha (.05) ✓

2. confidence Interval ✓



$$Sales = b_0 + b_1 \cdot Ad$$

$$= 500 + 50 \cdot Ads$$

estimate

Ad

~~$H_0: b_1 = 0$~~ ✓

$H_a: b_1 \neq 0$

[0]

CI คร่อม H_0 (0)

Fail to reject Ho

CI [ไม่คร่อม 0] Reject Ho

1. H_o, H_a.
2. Collect Data
3. Conclusion

model   Ad $p \leq 5\%$

Sig!

Better Explanation

Ad = 95% [+50, +65]

100↑

AB Test

- Ad_1 ↑
- Ad_2

50–65   95%

H_o: $Ad_1 = Ad_2$

H_a: $Ad_1 \neq Ad_2$ ← T.test / Linear

p-value definition

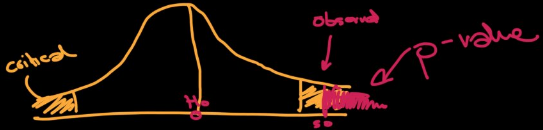
$$P\left(\overset{\geq 55}{\text{observed data}}\ \text{or}\ \underline{\text{more extreme}} \;\middle|\; H_0 \text{ is TRUE}\right)$$

ความน่าจะเป็น
ที่จะเห็น H ≥ 55 ?

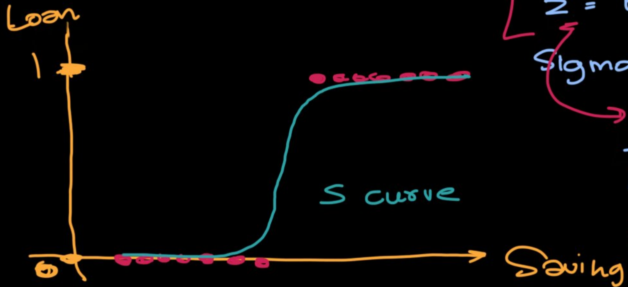
ถ้าเหรียญเป็น Fair coin
จริง.

H_0 : Fair coin

H_a : Not a fair coin



Binary Classification → 0 / 1 yes, no



$[z = b_0 + b_1 x]$

Sigmoid function

$$\frac{\cancel{e}^{z}\,\cancel{\$}100}{1+\cancel{e}^{z}\,\$00} \sim [0, 1]$$

$\frac{5}{6}$ ↓

$\frac{100}{101} \sim .99$

$\frac{0.2}{1.2} \sim 0.26$

Sigmoid Function

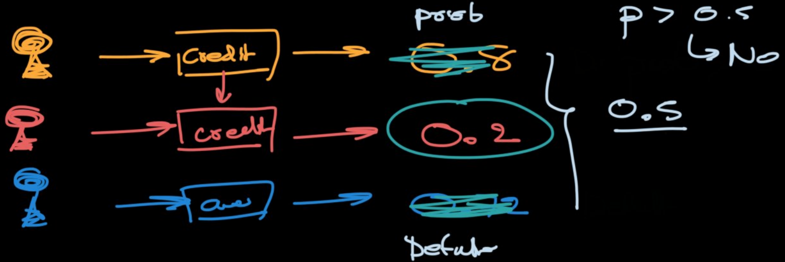
$$\frac{e^{z}}{1+e^{z}} \sim \text{Sigmoid}$$

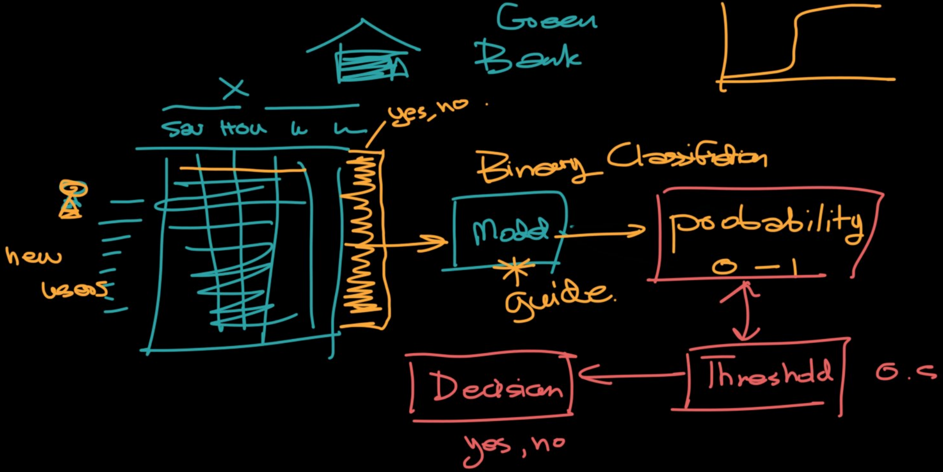
output

$[0, 1] = S(z)$

$$\text{Sigmoid} = \frac{e^{z}}{1+e^{z}} \longleftarrow z = b_0 + b_1 \cdot x$$

$$= [0, 1] \quad \text{probability}$$

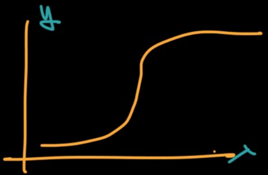



Green
Bank
X
sal
Hou
yes, no
new
users
Binary Classification
Model
guide
probability
0 – 1
Threshold
0.5
Decision
yes, no

glm( )

↑

family = "binomial"



Evaluate Binary Model

Confusion Matrix

		predict	
		no	yes
Actual	no	2	1
	yes	1	1

TRUE Negative

False Positive

False Negative

TRUE POSITIVE

$3/5 = 60\%$

x	y	$\hat{y}$	correct
	0	0	1
	0	0	1
	1	0	0
	0	1	0
	1	1	1

N

P

CONFUSION MAT

$n = 1000$ ✓

		prediction	
		no	yes
Actual	no	TN 300	FP 120
	yes Default	FN 80	TP 500

$$\text{Acc} = \frac{300+500}{1000}$$

$$= 80\%$$

$$\text{Recall} = \frac{TP}{FN+TP}$$

$$= \frac{500}{80+500}$$

$$= \frac{500}{580}$$

$$= 86\%$$

$$\text{precision} = \frac{TP}{Fp + TP} = \frac{500}{120+500} = \frac{500}{620}$$

$$\approx 80.6\%$$

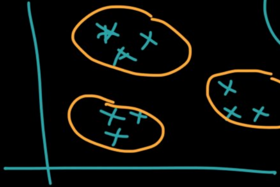
F_1

$$\frac{2 \cdot \text{Prec} \cdot \text{Recall}}{\text{Prec} + \text{Recall}}$$

1. Hypothesis . p-value / CI.

2. Binary Classification

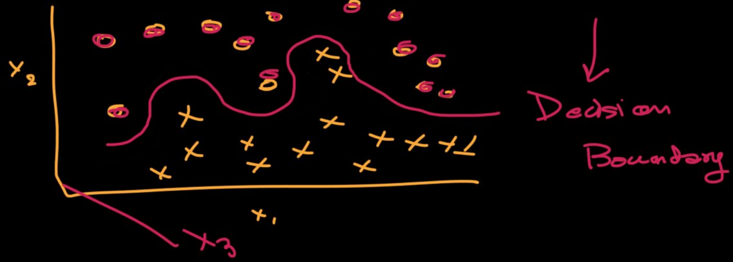
3. Clustering – K. means.



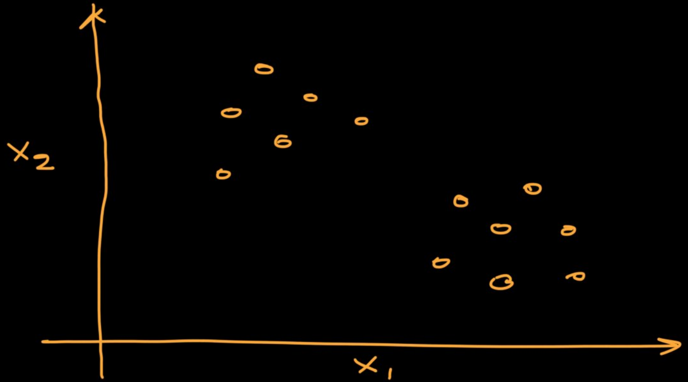
Unsupervised Learning

Regression* ✓ (Parametric ✓)

$$y = b_0 + \underset{\uparrow}{b_1} \cdot \underset{\uparrow}{x} + \underset{\downarrow}{b_2}\underset{\uparrow}{x_2} + \cdots$$



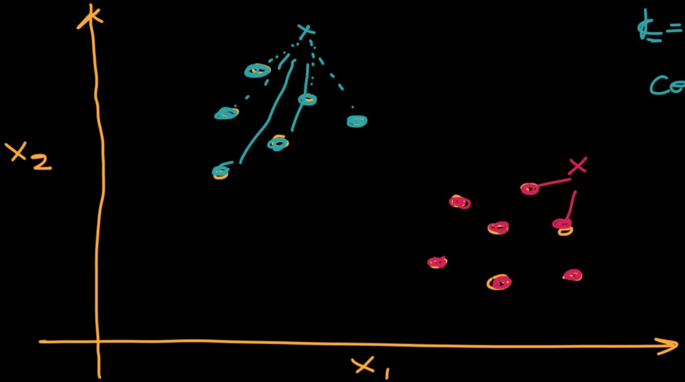
3. Clustering



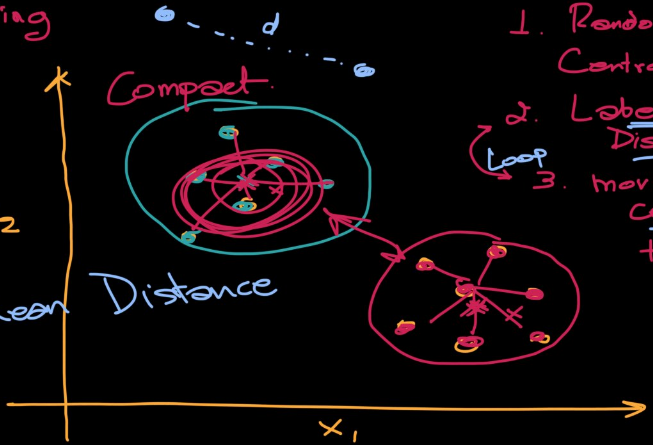
3. Clustering K-Means (Iterative) + Random

$\underline{k} = \underline{2}$

Centroid



3. Clustering



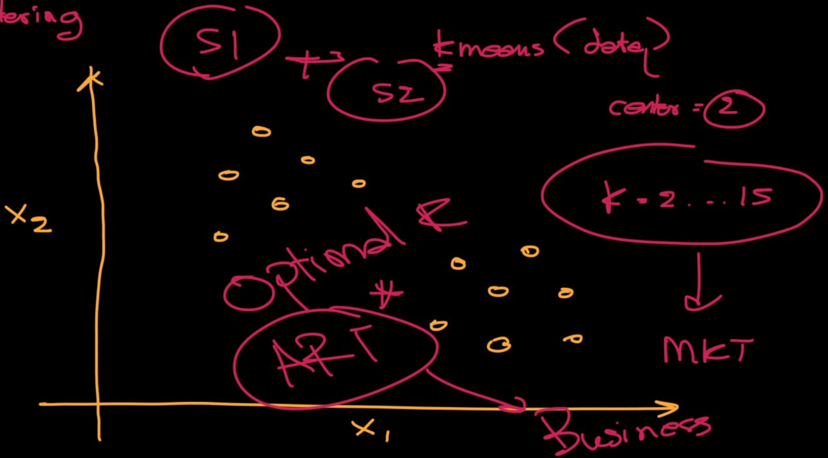
1. Random Centroid
2. Label *.
   Distance

Loop

3. move centroid to centra cluster

Euclidean Distance

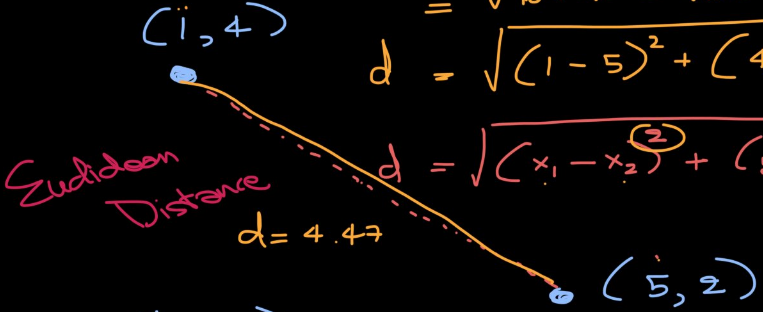
3. Clustering



$$= \sqrt{16+4} = \sqrt{20} = 4.47$$

$$d = \sqrt{(1-5)^2 + (4-2)^2}$$

$$d = \sqrt{(x_1 - x_2)^{\boxed{2}} + (y_1 - y_2)^2} \checkmark$$



$$(1,4,5) \xrightarrow[d]{} (2,10,9)$$

$$d = \sqrt{(1-2)^2 + (4-10)^2 + (5-9)^2} = \dots$$

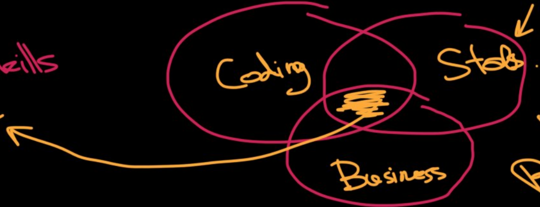
1. Hypothesis A/B Test : t.test( ) =T.TEST( )

2. Binary Classification glm( ) / Confusion Matrix table( )

3. Clustering kmeans

Data Skills

Science



Model

Business.

Project (Open).

#stat.ml

Datalore

1. data.world

→ churn

Explore data.

2. Build Model

└→ Logistic Reg.

└→ glm()

Bonus 3. K-means

2

Acc

Recall

Precision

F1