

TAUTOLOGY  
INNOVATION  
SCHOOL

$\tau$ TAUTOLOGY

CROSS ENTROPY

# CROSS ENTROPY

BY TAUTOLOGY

MADE BY TAUTOLOGY THAILAND  
DO NOT PUBLISH WITHOUT PERMISSION

facebook/tautologyai  
www.tautology.live

# Cross Entropy

- 2-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- Multi-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^k [y_{i,c} \log(\hat{y}_{i,c})]$$

# Cross Entropy

- Information Theory
- Uncertainty
- Entropy
- KL Divergence

# Information Theory

# Information Theory

- Concept
- Definition
- Formation

# Concept

แนวคิดของ information มี 2 ข้อ ดังต่อไปนี้

1. เหตุการณ์ที่มี**โอกาสเกิดขึ้นต่ำ** (low probability) จะมี **information สูง**
2. เหตุการณ์ที่มี**โอกาสเกิดขึ้นสูง** (high probability) จะมี **information ต่ำ**

# Information Theory

- ✓ • **Concept**
  - Definition
  - Formation



# Definition

1. เหตุการณ์ที่มีความน่าจะเป็น 100% จะไม่มี information ใด ๆ
2. ยิ่งเหตุการณ์มีโอกาสดังขึ้นน้อยเท่าไร information ก็จะมีค่ามากขึ้นเท่านั้น
3. Information รวมของสองเหตุการณ์ที่เป็นอิสระต่อกันจะเท่ากับผลรวมของ information ของสองเหตุการณ์นั้น ๆ



# Information Theory

- ✓ • **Concept**
- ✓ • **Definition**
- **Formation**

# Formation

เราต้องการหา function ที่แสดงความสัมพันธ์ระหว่าง information และ probability

$$\text{information} = f(\text{probability})$$

# Formation

กำหนดให้  $I(x)$  คือ information ของเหตุการณ์  $x$   
และ  $p(x)$  คือ probability ของเหตุการณ์  $x$   
จะได้ว่า

$$I(x) = f(p(x))$$

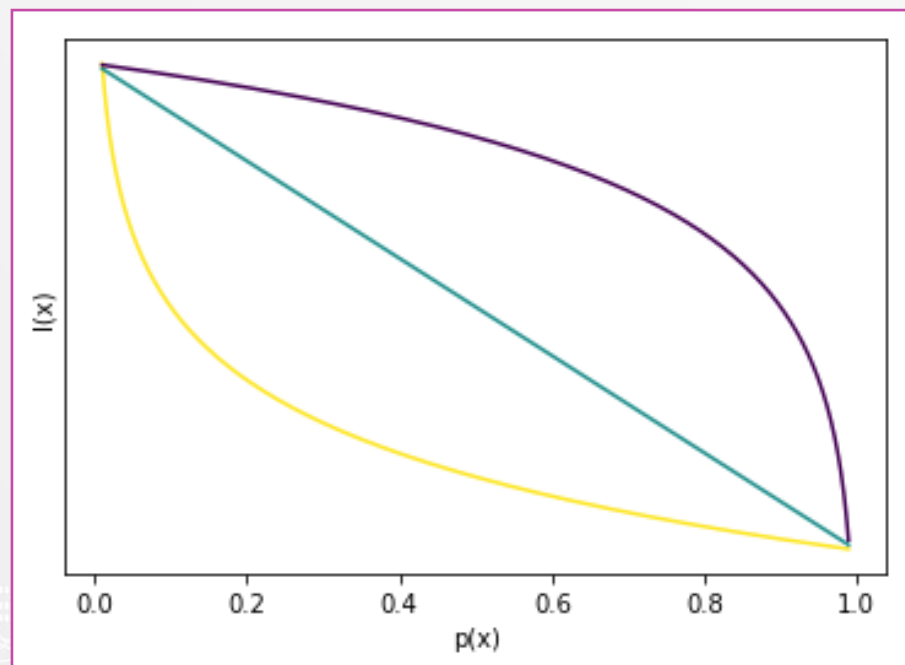
# Formation

จาก definition ข้อที่ 1 “เหตุการณ์ที่มีความน่าจะเป็น 100% จะไม่มี information ใด ๆ” จะได้ว่า

$$\text{ถ้า } p(x) = 1 \text{ แล้ว } I(x) = f(1) = 0$$

# Formation

จาก definition ข้อที่ 2 “ยิ่งเหตุการณ์มีโอกาสเกิดขึ้นน้อยเท่าไร information ก็จะมีค่ามากขึ้นเท่านั้น” จะได้ว่า ความสัมพันธ์ระหว่าง  $I(x)$  และ  $p(x)$  เป็นแบบ monotone function



# Formation

กำหนดให้เหตุการณ์ A และเหตุการณ์ B เป็นอิสระต่อกัน และกำหนดให้เหตุการณ์ C เป็นเหตุการณ์ A และ B เกิดขึ้นพร้อมกัน จะได้ว่า

$$p(c) = p(A \cap B) = p(A) \cdot p(B)$$

# Formation

จาก  $I(x) = f(p(x))$  จะได้ว่า

$$\begin{aligned} I(c) &= f(p(C)) \\ &= f(p(A) \cdot p(B)) \end{aligned}$$



# Formation

จาก definition ข้อที่ 3 “Information รวมของสองเหตุการณ์ที่เป็นอิสระต่อกันจะเท่ากับผลรวมของ information ของสองเหตุการณ์นั้น ๆ” จะได้ว่า

$$\begin{aligned} I(c) &= f(p(C)) \\ &= f(p(A) \cdot p(B)) \\ &= f(p(A)) + f(p(B)) \end{aligned}$$

# Formation

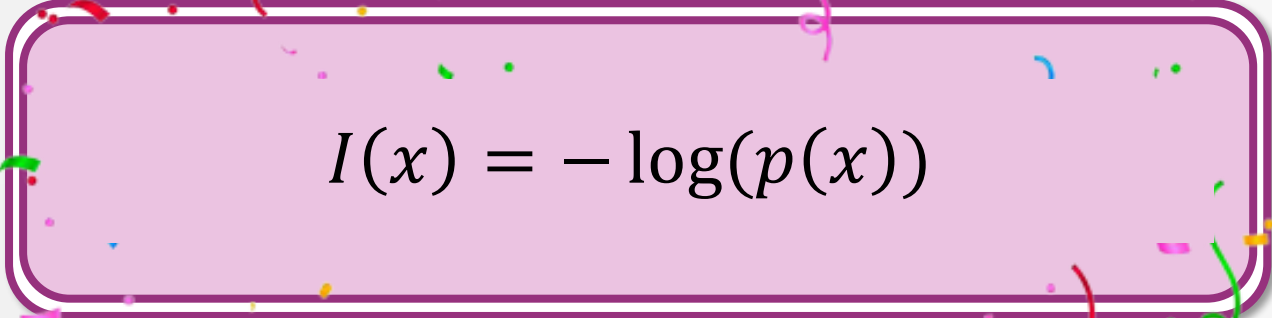
## Function เพียงอันเดียวที่มีคุณสมบัติ

1. เป็น monotone function บนช่วง  $[0,1]$
2.  $f(\square \cdot \triangle) = f(\square) + f(\triangle)$
3.  $f(1) = 0$

# Formation

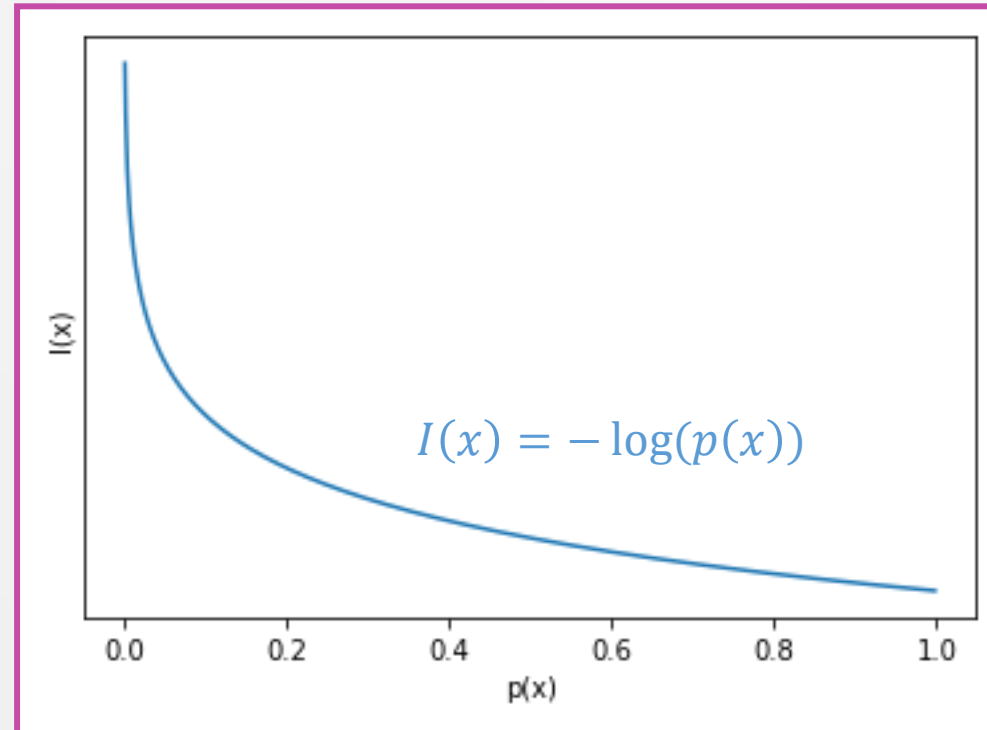
$$f(p(x)) = -\log(p(x))$$

# Formation


$$I(x) = -\log(p(x))$$

# Formation

ดังนั้น



# Information Theory

- ✓ • **Concept**
- ✓ • **Definition**
- ✓ • **Formation**

# Cross Entropy

## ✓ • Information Theory

- Uncertainty
- Entropy
- KL Divergence



# Uncertainty

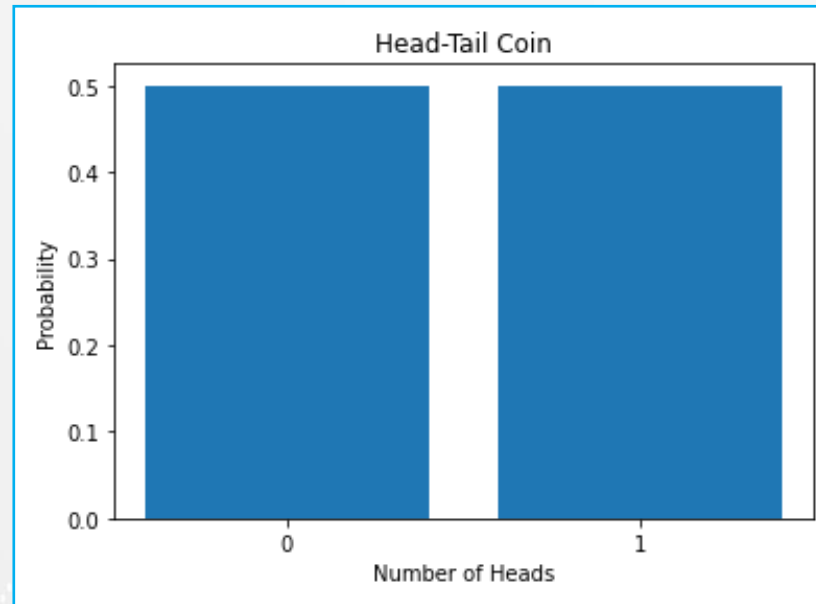
# Uncertainty

Uncertainty คือ คำที่ใช้บอกความไม่เป็นระเบียบ/ความยุ่งเหยิงของระบบ ซึ่งเป็นอีก  
หนึ่งชื่อเรียกของ information

uncertainty = Information

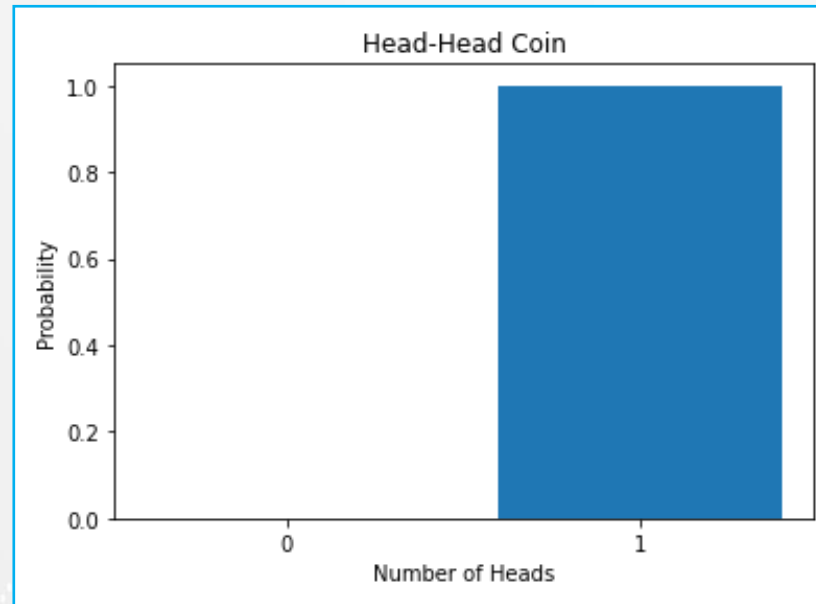
# Uncertainty

- ระบบที่มีความยุ่งเหยิง



# Uncertainty

- ระบบที่ไม่มีคามยุ่งเหยิง



# Cross Entropy

- ✓ • **Information Theory**
- ✓ • **Uncertainty**
  - Entropy
  - KL Divergence

# Entropy

# Entropy

**Entropy** คือ ค่าที่บอกถึงค่าเฉลี่ยของ information หรือ uncertainty ในระบบ

$$H(P) = E[I(x)]$$



# Entropy

$$\begin{aligned} H(P) &= E[I(x)] \\ &= E[-\log(p(x))] \\ &= -E[\log(p(x))] \\ &= -\sum_{c=1}^k p(x_c) \log(p(x_c)) \end{aligned}$$

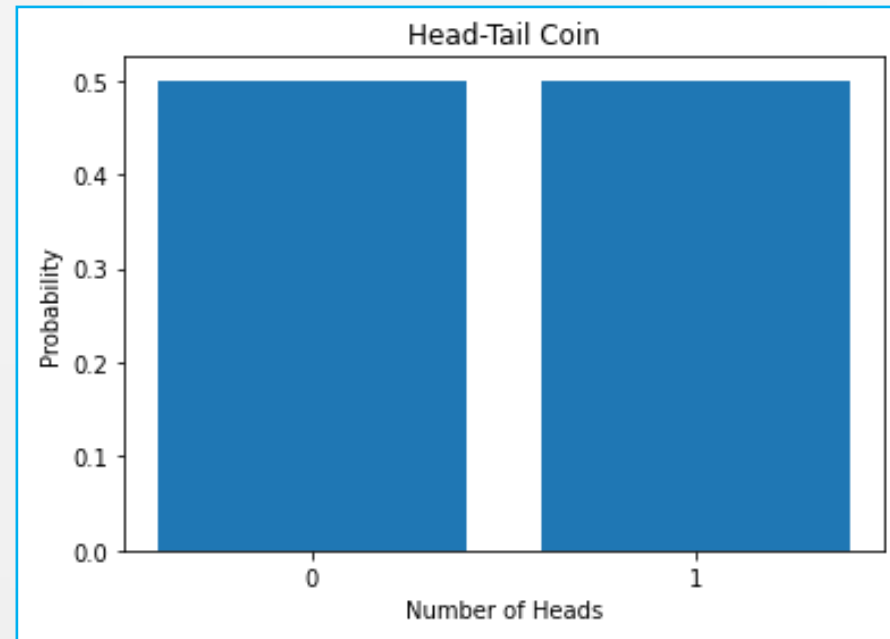
# Entropy

**Entropy** คือ ค่าที่บอกถึงค่าเฉลี่ยของ information หรือ uncertainty ของระบบ

$$H(P) = - \sum_{c=1}^k p(x_c) \log(p(x_c))$$

# Entropy

## ตัวอย่าง (1)



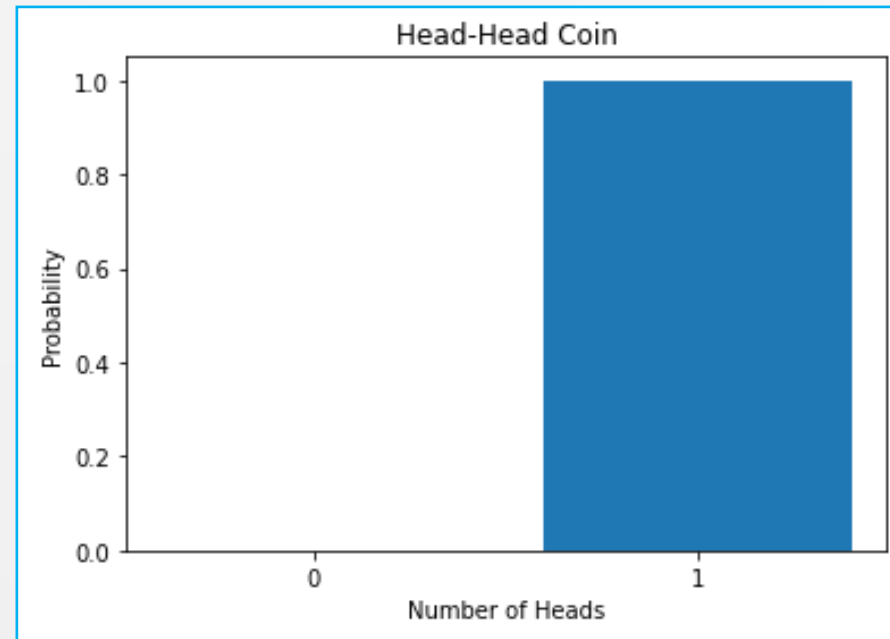
# Entropy

## ตัวอย่าง (1)

$$\begin{aligned} H(P) &= -\sum_{c=1}^2 p(x_c) \log(p(x_c)) \\ &= -p(x_1) \log(p(x_1)) - p(x_2) \log(p(x_2)) \\ &= -p(0) \log(p(0)) - p(1) \log(p(1)) \\ &= -\frac{1}{2} \log\left(\frac{1}{2}\right) - \frac{1}{2} \log\left(\frac{1}{2}\right) \\ &= 0.6931 \end{aligned}$$

# Entropy

## ตัวอย่าง (2)



# Entropy

## ตัวอย่าง (2)

$$\begin{aligned} H(P) &= -\sum_{c=1}^2 p(x_c) \log(p(x_c)) \\ &= -p(x_1) \log(p(x_1)) - p(x_2) \log(p(x_2)) \\ &= -p(0) \log(p(0)) - p(1) \log(p(1)) \\ &= -0 \log(0) - 1 \log(1) \\ &= 0 \end{aligned}$$

# Cross Entropy

- ✓ • **Information Theory**
- ✓ • **Uncertainty**
- ✓ • **Entropy**
  - KL Divergence



# KL Divergence

# KL Divergence

- What is KL Divergence?
- Origin of the Equation
- KL as Cost Function

# What is KL Divergence?

**KL Divergence** คือ เครื่องมือที่ใช้ในการวัดความแตกต่างระหว่าง 2 distribution  $(P, Q)$  ว่า  $Q$  แตกต่างจาก  $P$  เท่าไหร่

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=1}^k p(x_c) \log(q(x_c))$$

# What is KL Divergence?

- ถ้า  $P$  และ  $Q$  เหมือนกันทุกประการ แล้ว  $D_{KL}(P \parallel Q) = 0$
- ถ้า  $P$  และ  $Q$  แตกต่างกัน แล้ว  $D_{KL}(P \parallel Q) > 0$  (ยิ่งแตกต่างมาก  $D_{KL}(P \parallel Q)$  ยิ่งมีค่ามาก)

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=1}^k p(x_c) \log(q(x_c))$$

# KL Divergence

## ✓ • What is KL Divergence?

- Origin of the Equation
- KL as Cost Function

# Origin of the Equation

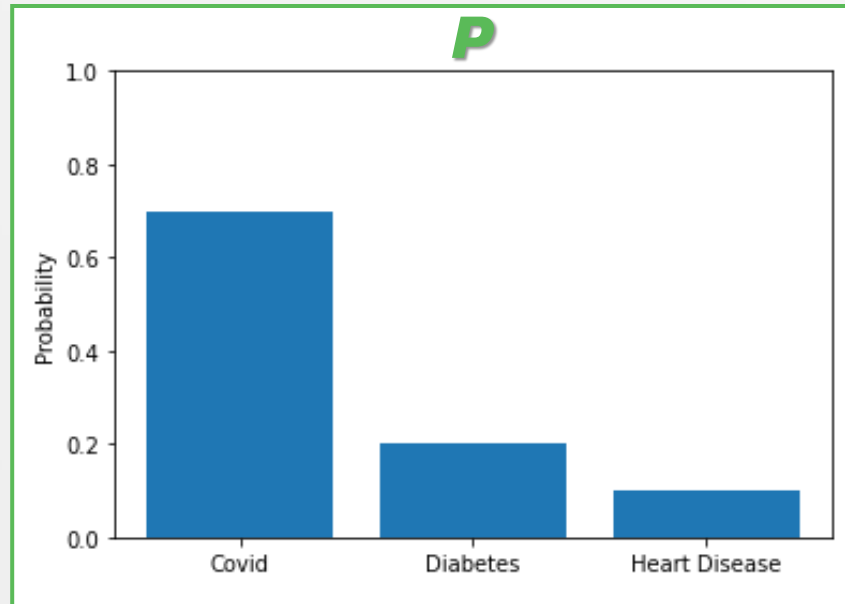
$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=1}^k p(x_c) \log(q(x_c))$$

# Origin of the Equation

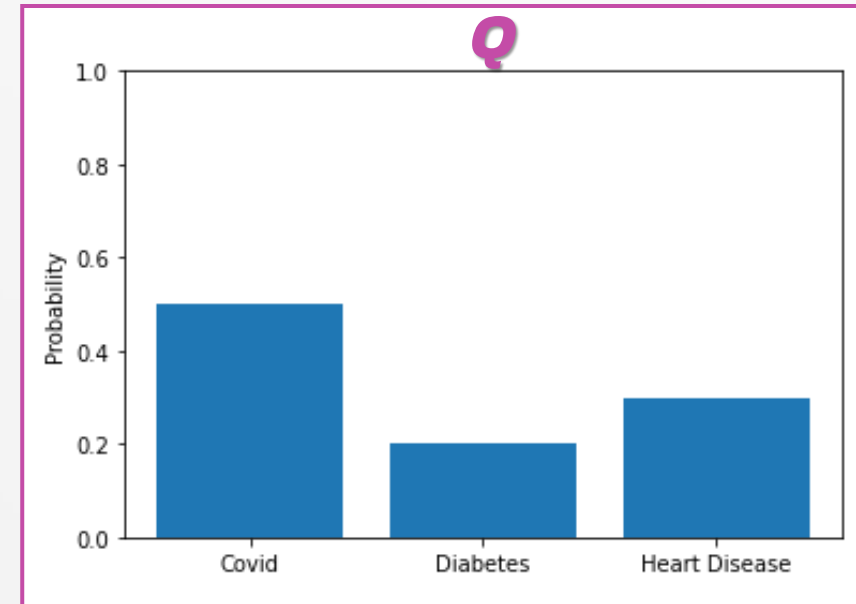




# Origin of the Equation



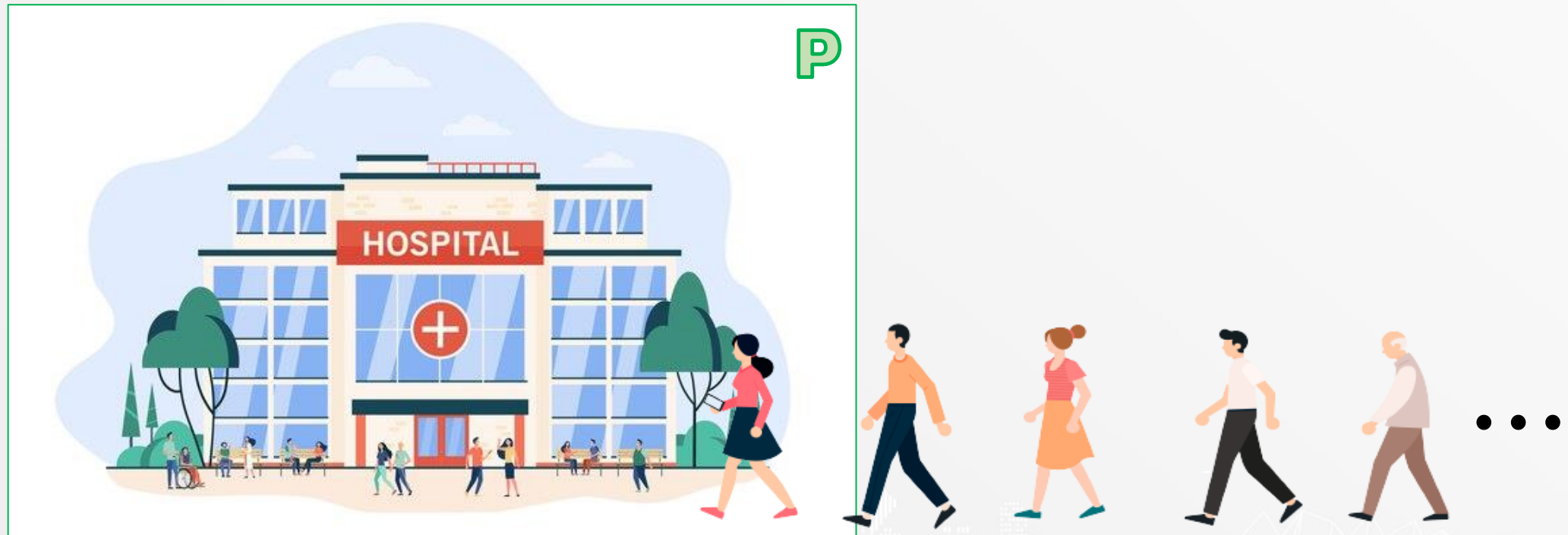
- $p(\text{โควิด}) = 0.7$
- $p(\text{เบาหวาน}) = 0.2$
- $p(\text{หัวใจ}) = 0.1$



- $p(\text{โควิด}) = 0.5$
- $p(\text{เบาหวาน}) = 0.2$
- $p(\text{หัวใจ}) = 0.3$



# Origin of the Equation



# Origin of the Equation



# Origin of the Equation

$$\frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{เบาหวาน})}{q(\text{เบาหวาน})} \quad \frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{หัวใจ})}{q(\text{หัวใจ})} \quad \dots$$

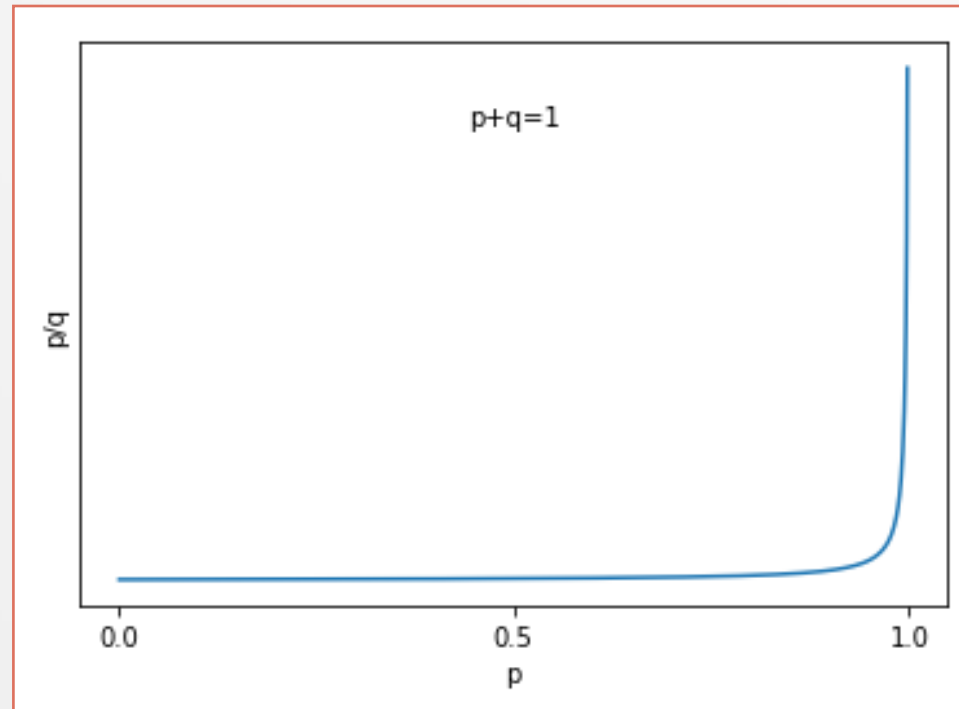


# Origin of the Equation

$$\frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{เบ้าหวาน})}{q(\text{เบ้าหวาน})} \quad \frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{หั่วใจ})}{q(\text{หั่วใจ})} \quad \dots$$

หาค่าเฉลี่ย

# Origin of the Equation



# Origin of the Equation

$$\frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{เบ้าหวาน})}{q(\text{เบ้าหวาน})} \quad \frac{p(\text{โควิด})}{q(\text{โควิด})} \quad \frac{p(\text{หิวใจ})}{q(\text{หิวใจ})} \quad \dots$$

หาค่าเฉลี่ย 

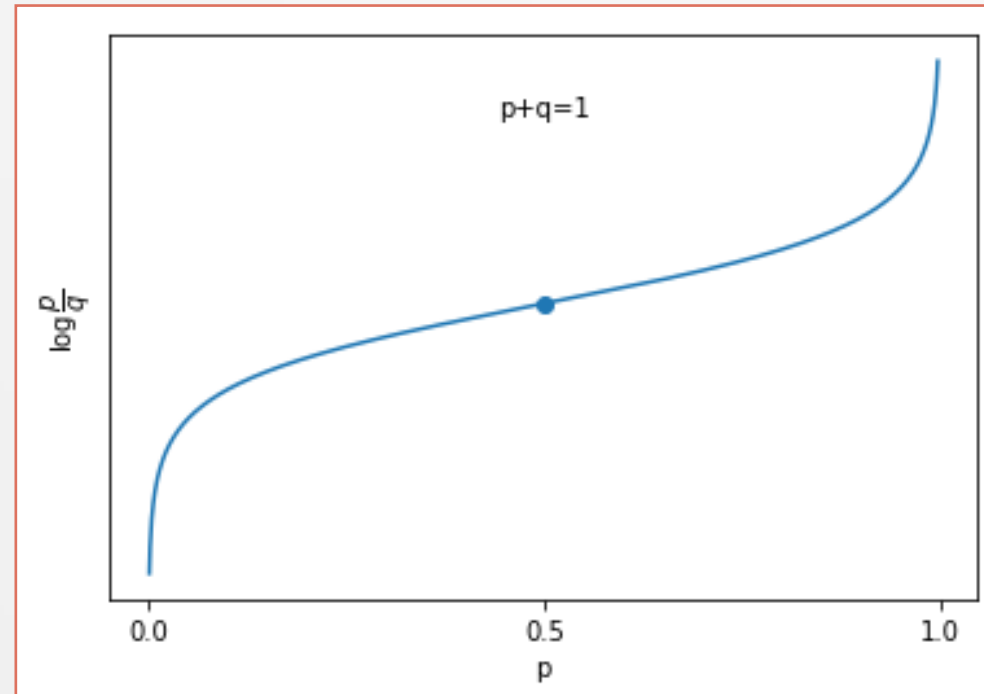
# Origin of the Equation

$$\log\left(\frac{p(\text{โควิด})}{q(\text{โควิด})}\right) \log\left(\frac{p(\text{โควิด})}{q(\text{โควิด})}\right) \log\left(\frac{p(\text{เบาหวาน})}{q(\text{เบาหวาน})}\right) \log\left(\frac{p(\text{โควิด})}{q(\text{โควิด})}\right) \log\left(\frac{p(\text{หัวใจ})}{q(\text{หัวใจ})}\right) \dots$$

หาค่าเฉลี่ย



# Origin of the Equation

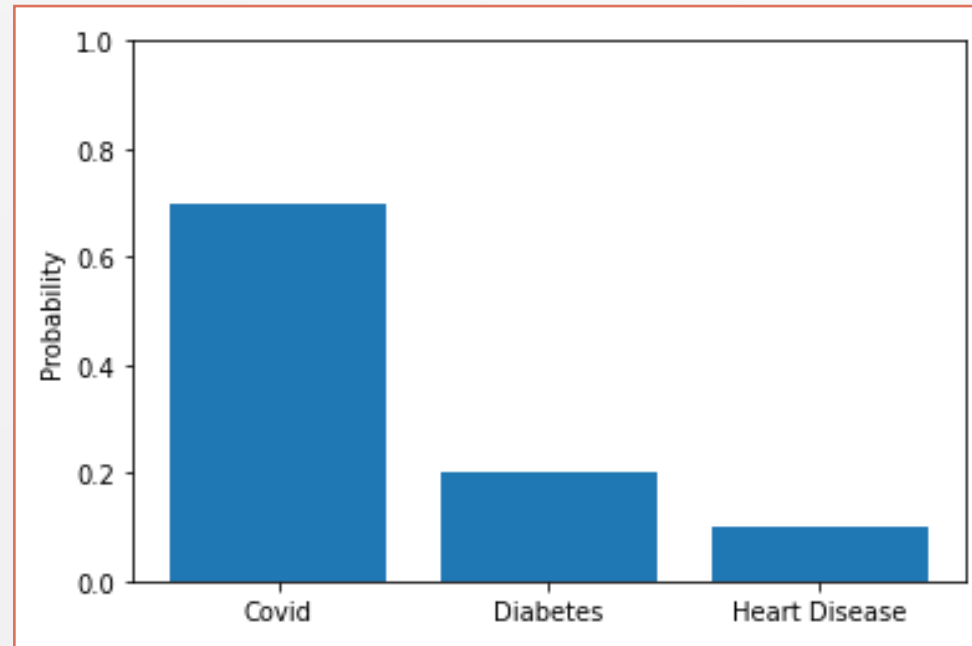




# Origin of the Equation

$$D_{KL}(P \parallel Q) = \frac{1}{n} \sum_{c=1}^n \log \left( \frac{p(x_c)}{q(x_c)} \right)$$

# Origin of the Equation



- $p(\text{โควิด}) = 0.7$
- $p(\text{เบาหวาน}) = 0.2$
- $p(\text{หัวใจ}) = 0.1$

# Origin of the Equation

$$\begin{aligned} D_{KL}(P \parallel Q) = & \frac{1}{n} \left[ p(\text{ໂຄວັດ}) \cdot n \cdot \log \left( \frac{p(\text{ໂຄວັດ})}{q(\text{ໂຄວັດ})} \right) \right. \\ & + p(\text{ເສາຫວານ}) \cdot n \cdot \log \left( \frac{p(\text{ເສາຫວານ})}{q(\text{ເສາຫວານ})} \right) \\ & \left. + p(\text{ຫ້ວໃຈ}) \cdot n \cdot \log \left( \frac{p(\text{ຫ້ວໃຈ})}{q(\text{ຫ້ວໃຈ})} \right) \right] \end{aligned}$$

# Origin of the Equation

$$\begin{aligned} D_{KL}(P \parallel Q) &= p(\text{ໂຄວັດ}) \cdot \log \left( \frac{p(\text{ໂຄວັດ})}{q(\text{ໂຄວັດ})} \right) \\ &+ p(\text{ເສາວານ}) \cdot \log \left( \frac{p(\text{ເສາວານ})}{q(\text{ເສາວານ})} \right) \\ &+ p(\text{ຫ້ວໃຈ}) \cdot \log \left( \frac{p(\text{ຫ້ວໃຈ})}{q(\text{ຫ້ວໃຈ})} \right) \end{aligned}$$

# Origin of the Equation

$$D_{KL}(P \parallel Q) = \sum_{c=1}^k p(x_c) \log \left( \frac{p(x_c)}{q(x_c)} \right)$$

# Origin of the Equation

$$\begin{aligned} D_{KL}(P \parallel Q) &= \sum_{c=1}^k p(x_c) [\log(p(x_c)) - \log(q(x_c))] \\ &= \sum_{c=1}^k [p(x_c) \log(p(x_c)) - p(x_c) \log(q(x_c))] \end{aligned}$$

# Origin of the Equation

$$\begin{aligned} D_{KL}(P \parallel Q) &= \sum_{c=1}^k p(x_c) \log(p(x_c)) - \sum_{c=1}^k p(x_c) \log(q(x_c)) \\ &= -H(P) - \sum_{c=1}^k p(x_c) \log(q(x_c)) \\ &(\because H(P) = -\sum_{c=1}^k p(x_c) \log(p(x_c))) \end{aligned}$$

# Origin of the Equation

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=1}^k p(x_c) \log(q(x_c))$$



# KL Divergence

- ✓ • **What is KL Divergence?**
- ✓ • **Origin of the Equation**
  - KL as Cost Function

# KL as Cost Function

- 2-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- Multi-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^k [y_{i,c} \log(\hat{y}_{i,c})]$$

# KL as Cost Function

Model

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

ตารางแสดง dataset

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

ตารางแสดง  $\hat{y}$  ที่ได้จาก model

# KL as Cost Function

Model

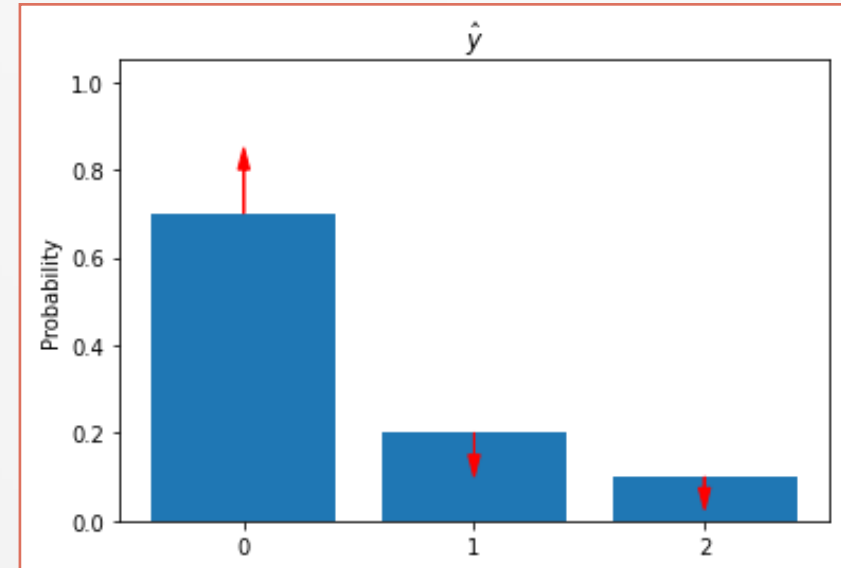
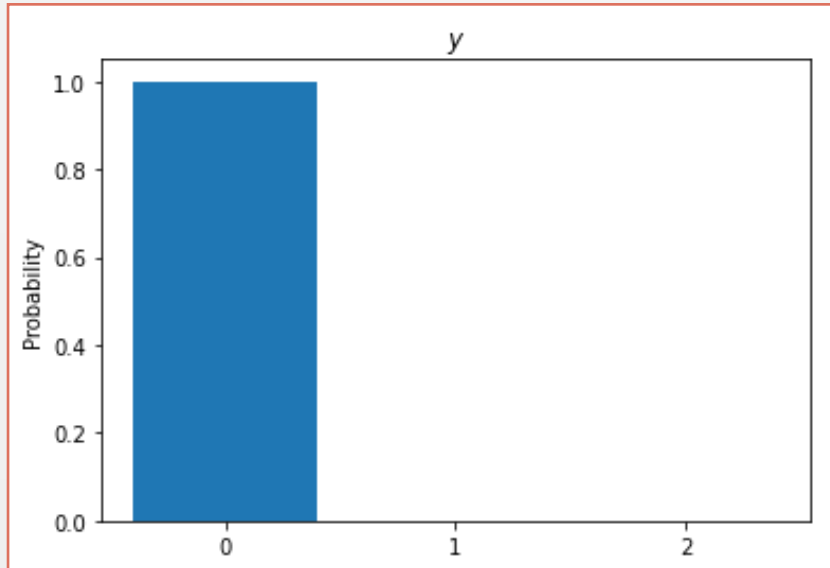
$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

ตารางแสดง dataset

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

ตารางแสดง  $\hat{y}$  ที่ได้จาก model

# KL as Cost Function



$$x_1 = 0, \quad x_2 = 1$$

# KL as Cost Function

Model

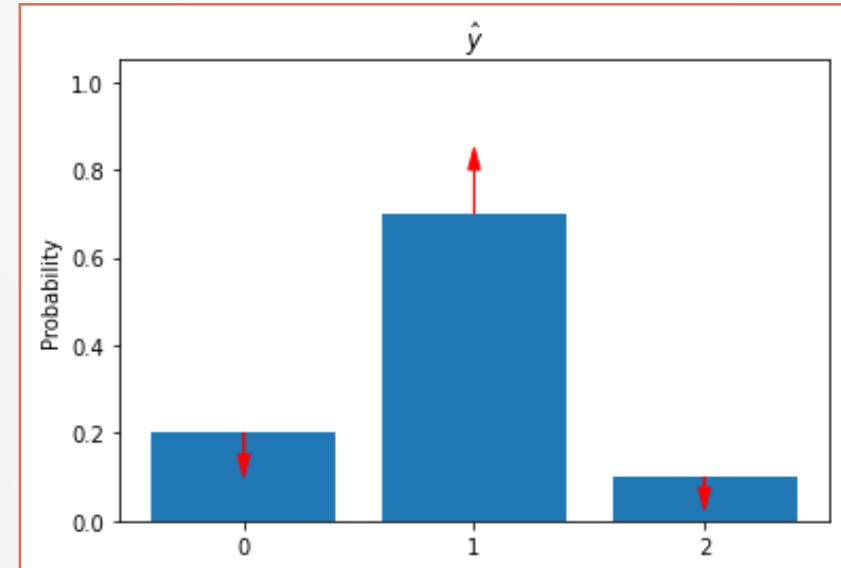
$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

ตารางแสดง dataset

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

ตารางแสดง  $\hat{y}$  ที่ได้จาก model

# KL as Cost Function



$$x_1 = 1, \quad x_2 = 0$$

# KL as Cost Function

Model

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

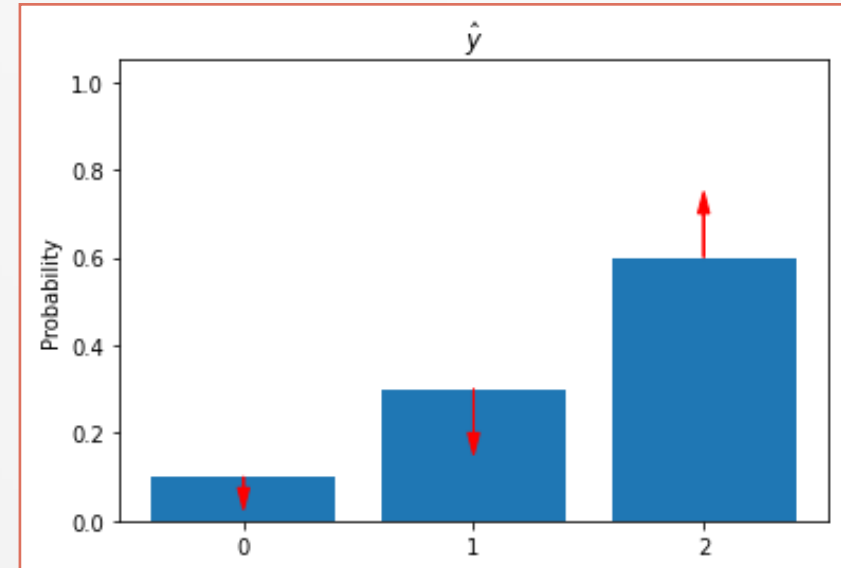
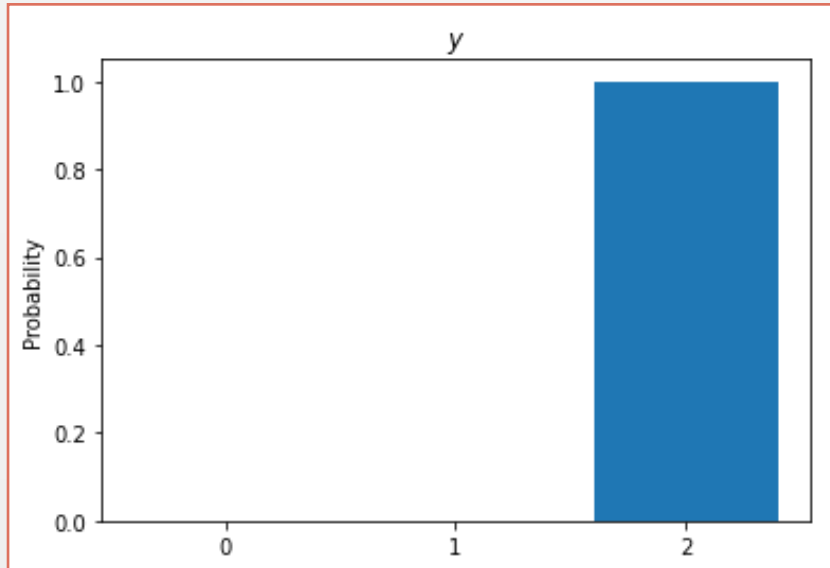
ตารางแสดง dataset

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

ตารางแสดง  $\hat{y}$  ที่ได้จาก model



# KL as Cost Function



$$x_1 = -1, \quad x_2 = 0$$

# KL as Cost Function

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=1}^k p(x_c) \log(q(x_c))$$



$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -H(\mathbf{y}_i) - \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -H(\mathbf{y}_i) - \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

ค่าคงที่

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -H(\mathbf{y}_i) - \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto - \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto - \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto -\sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

$$\sum_{i=1}^n D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto - \sum_{i=1}^n \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$



# KL as Cost Function

เราต้องการ model ที่ทำให้  $\sum_{i=1}^n D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$  มีค่าน้อยที่สุด  
( $\hat{\mathbf{y}}_i$  เหมือนกับ  $\mathbf{y}_i$  บนทุก sample มากที่สุด)

# KL as Cost Function

$x_1$	$x_2$	$y_1$	$y_2$	$y_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
0.5	0.3	0.2
0.2	0.7	0.1
$\vdots$	$\vdots$	$\vdots$
0.1	0.3	0.6

$$\min D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \equiv \min - \sum_{i=1}^n \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

$$\min - \sum_{i=1}^n \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

เพื่อความสะดวกในการใช้ gradient descent เราจึงใช้  
ค่าเฉลี่ยของ cross entropy ในการ train model

# KL as Cost Function


$$\min -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

# KL as Cost Function

- 2-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- Multi-class


$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^k [y_{i,c} \log(\hat{y}_{i,c})]$$

# KL as Cost Function

- พิจารณา *Cost* สำหรับ 2-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^k y_{i,c} \log(\hat{y}_{i,c})$$

$$= -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^2 y_{i,c} \log(\hat{y}_{i,c})$$

$$= -\frac{1}{n} \sum_{i=1}^n [y_{i,1} \log(\hat{y}_{i,1}) + y_{i,2} \log(\hat{y}_{i,2})]$$

# KL as Cost Function

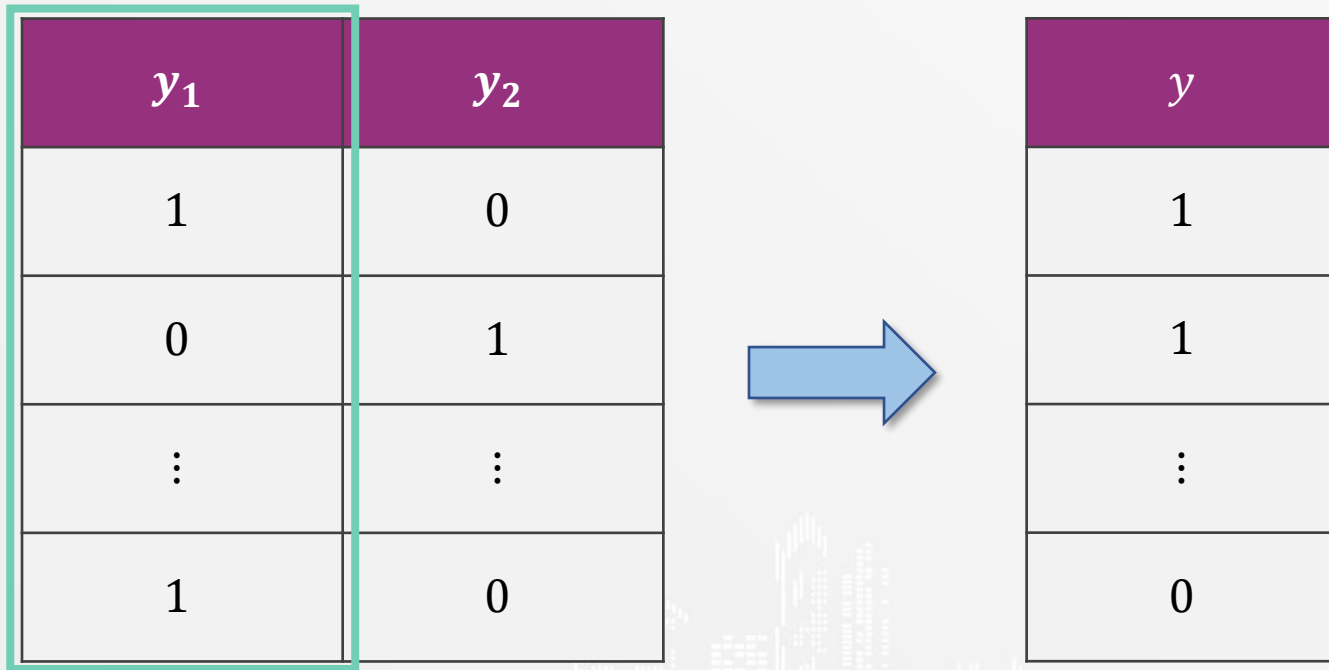
- พิจารณา *Cost* สำหรับ 2-class

$$\begin{aligned} \text{Cost} &= -\frac{1}{n} \sum_{i=1}^n [y_{i,1} \log(\hat{y}_{i,1}) + y_{i,2} \log(\hat{y}_{i,2})] \\ &= -\frac{1}{n} \sum_{i=1}^n [y_{i,1} \log(\hat{y}_{i,1}) + (1 - y_{i,1}) \log(1 - \hat{y}_{i,1})] \end{aligned}$$

$$\begin{aligned} (\because y_{i,1} + y_{i,2} &= 1 \\ \hat{y}_{i,1} + \hat{y}_{i,2} &= 1) \end{aligned}$$



# KL as Cost Function



$$\because y_1 + y_2 = 1$$

# KL as Cost Function



$$\because \hat{y}_1 + \hat{y}_2 = 1$$


# KL as Cost Function

- พิจารณา *Cost* สำหรับ 2-class


$$\begin{aligned} \text{Cost} &= -\frac{1}{n} \sum_{i=1}^n [y_{i,1} \log(\hat{y}_{i,1}) + (1 - y_{i,1}) \log(1 - \hat{y}_{i,1})] \\ &= -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \end{aligned}$$

# KL as Cost Function

- 2-class


$$Cost = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- Multi-class


$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^k [y_{i,c} \log(\hat{y}_{i,c})]$$

# KL Divergence

- ✓ • What is KL Divergence?
- ✓ • Origin of the Equation
- ✓ • KL as Cost Function

# Cross Entropy

- ✓• **Information Theory**
- ✓• **Uncertainty**
- ✓• **Entropy**
- ✓• **KL Divergence**