

TAUTOLOGY
INNOVATION
SCHOOL

τ TAUTOLOGY

CROSS ENTROPY

CROSS ENTROPY

BY TAUTOLOGY

MADE BY TAUTOLOGY THAILAND
DO NOT PUBLISH WITHOUT PERMISSION

facebook/tautologyai
www.tautology.live

Cross Entropy

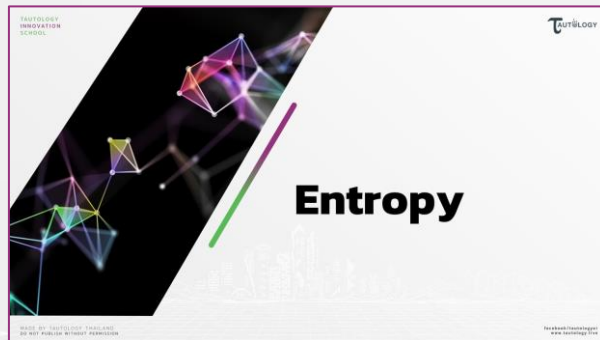
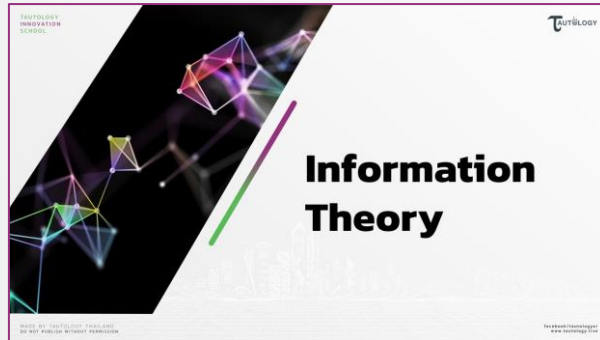
- 2-class

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

- Multi-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^{k-1} [y_{i,c} \log(\hat{y}_{i,c})]$$

Cross Entropy



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

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

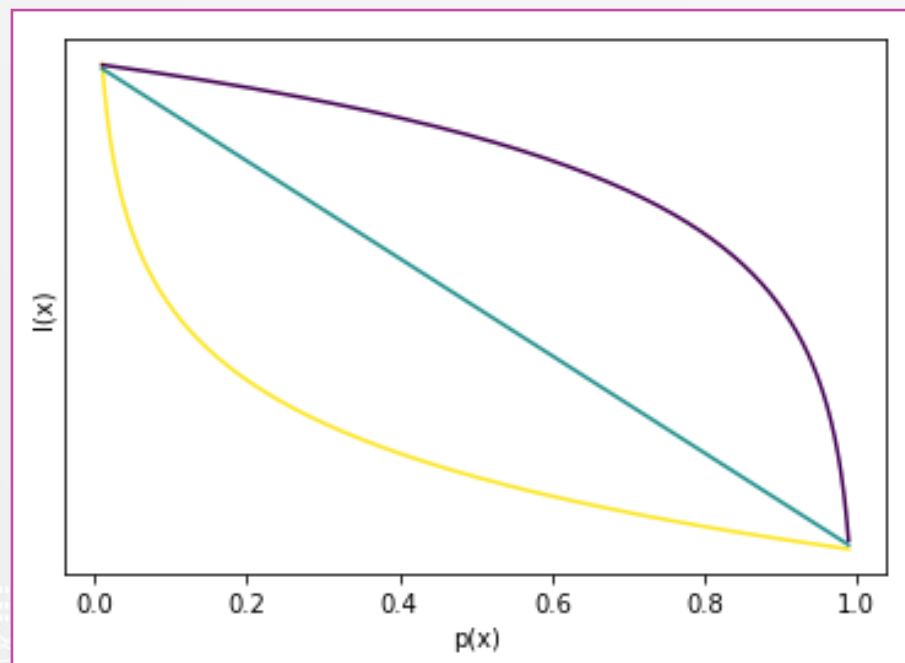
Formation

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

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

Formation

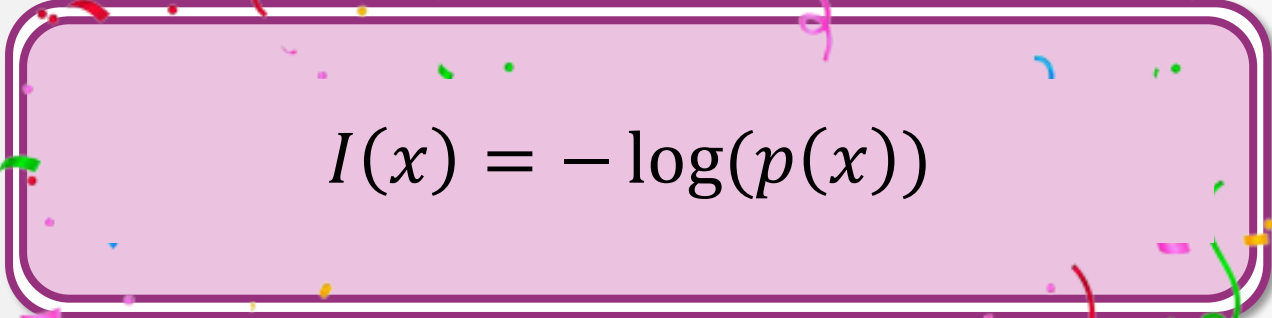
Function เพียงอันเดียวที่มีคุณสมบัติ

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

Formation

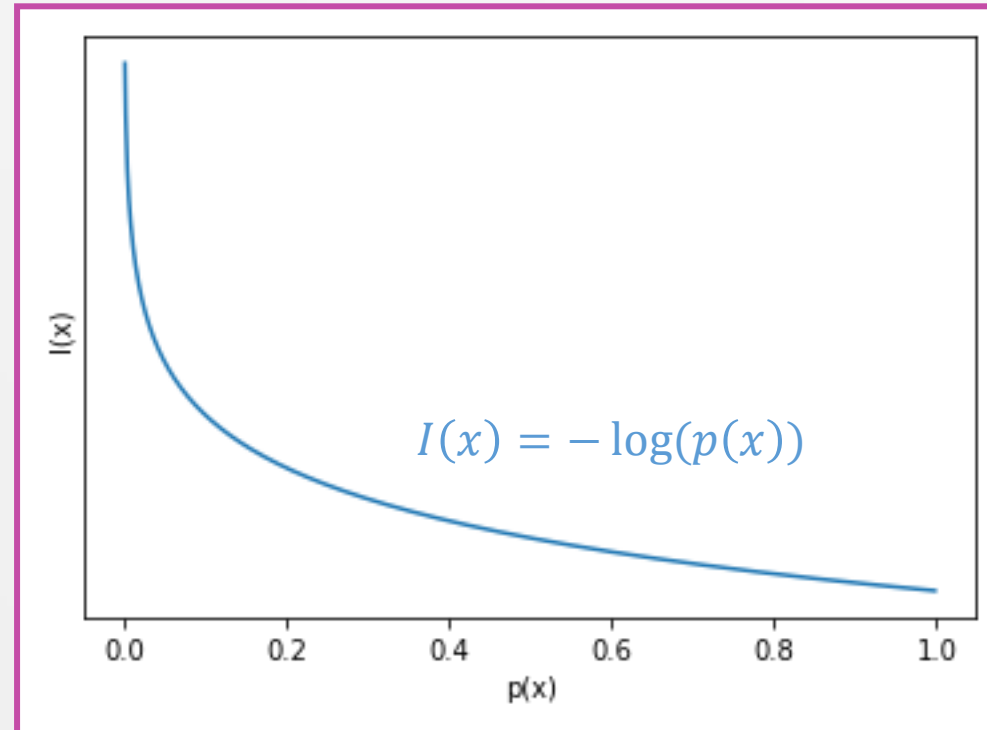
$$f(p(x)) = -\log(p(x))$$

Formation


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

Formation

ดังนั้น



Information Theory

Concept



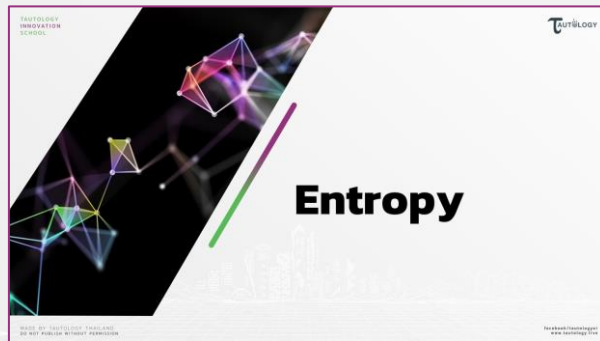
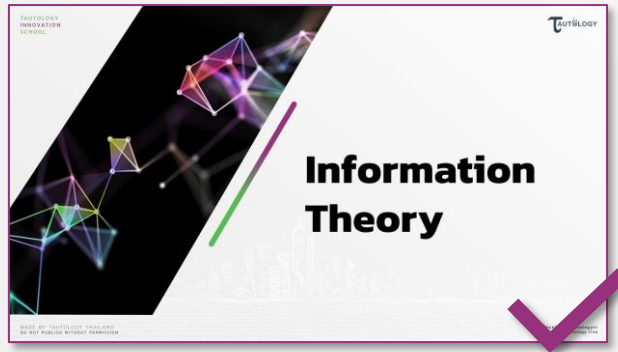
Definition



Formation



Cross Entropy



Uncertainty

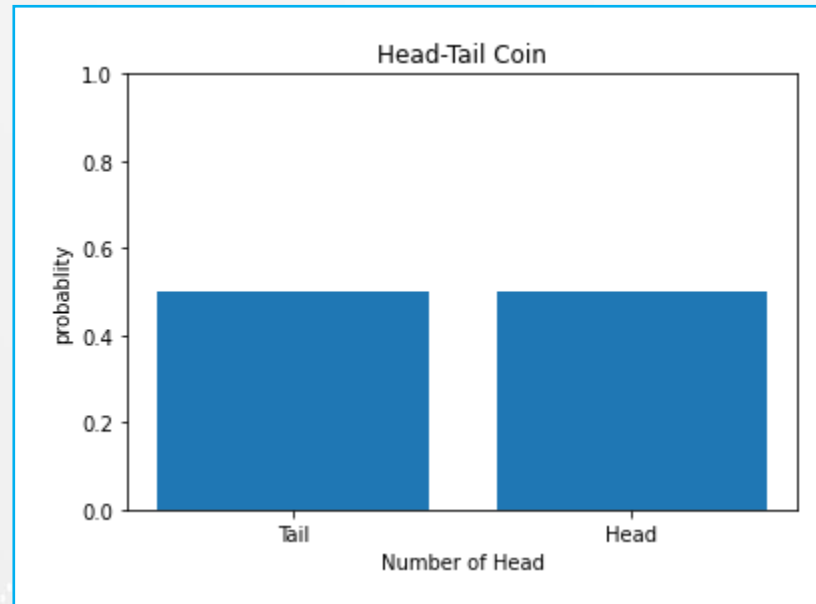
Uncertainty

Uncertainty คือ ค่าที่ใช้บอกความไม่เป็นระเบียบ/ความยุ่งเหยิงของระบบ ซึ่งเป็นอีก
หนึ่งชื่อเรียกของ information

uncertainty = information

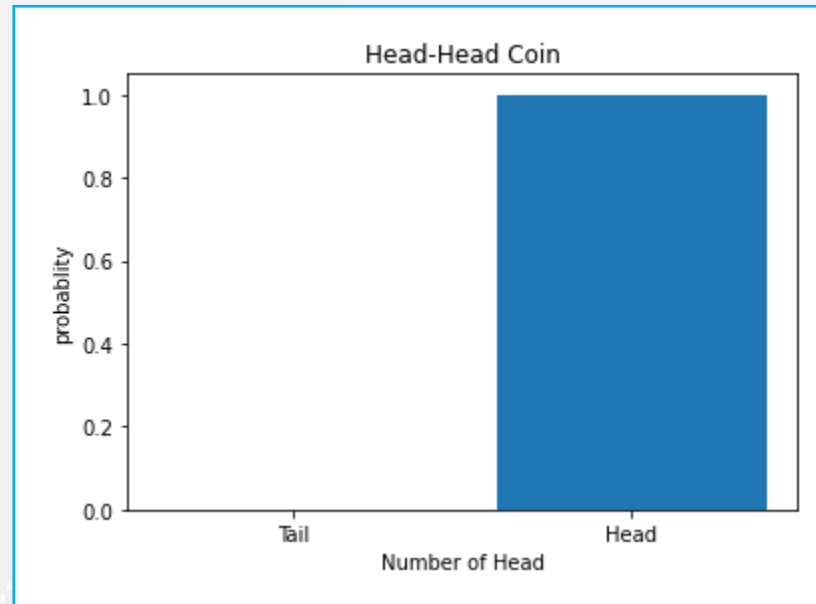
Uncertainty

- ระบบที่มีความยุ่งเหยิง

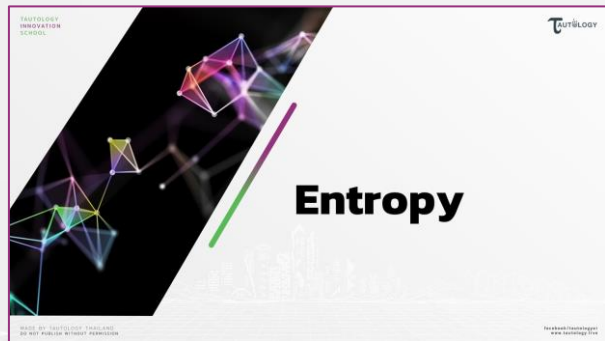
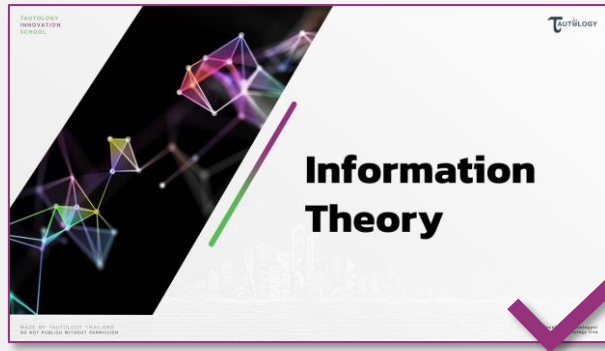


Uncertainty

- ระบบที่ไม่มีคามยุ่งเหยิง



Cross Entropy



Entropy

Entropy

Entropy คือ ค่าที่บอกถึงค่าเฉลี่ยของ information หรือ uncertainty ในระบบ

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

Entropy

$$\begin{aligned} H(P) &= E[I(x)] \\ &= E[-\log(p(x))] \\ &= -E[\log(p(x))] \\ &= -\sum_{c=0}^{k-1} p(x_c) \log(p(x_c)) \end{aligned}$$

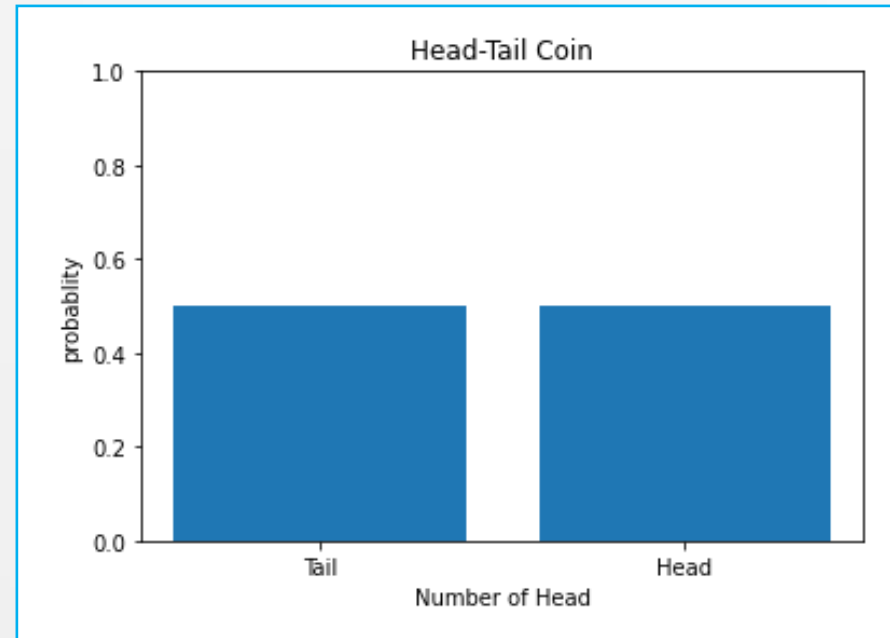
Entropy

Entropy คือ ค่าที่บอกถึงค่าเฉลี่ยของ information หรือ uncertainty ของระบบ

$$H(P) = - \sum_{c=0}^{k-1} p(x_c) \log(p(x_c))$$

Entropy

ตัวอย่าง (1)



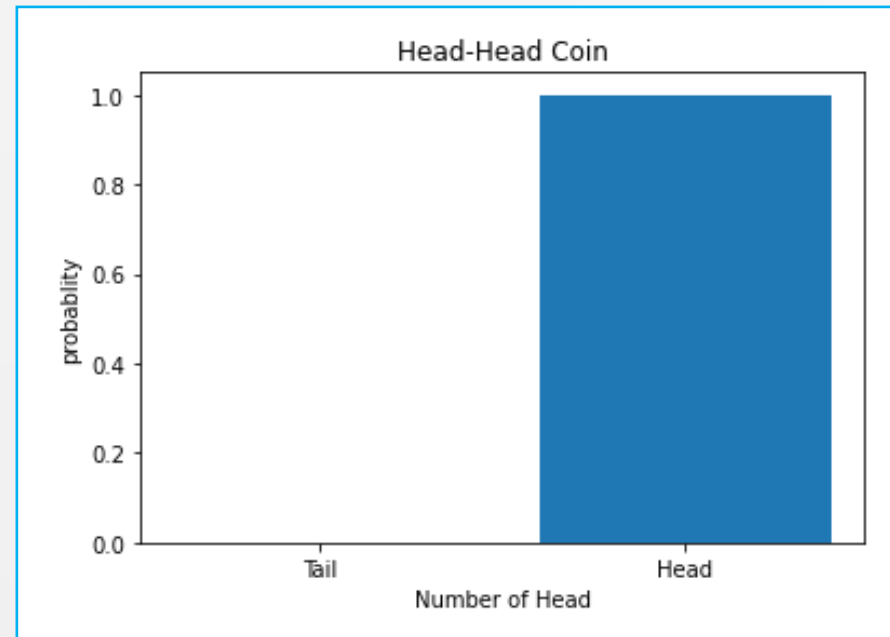
Entropy

ตัวอย่าง (1)

$$\begin{aligned} H(P) &= -\sum_{c=0}^1 p(x_c) \log(p(x_c)) \\ &= -p(x_0) \log(p(x_0)) - p(x_1) \log(p(x_1)) \\ &= -p(\text{Tail}) \log(p(\text{Tail})) - p(\text{Head}) \log(p(\text{Head})) \\ &= -\frac{1}{2} \log\left(\frac{1}{2}\right) - \frac{1}{2} \log\left(\frac{1}{2}\right) \\ &= 0.6931 \end{aligned}$$

Entropy

ตัวอย่าง (2)

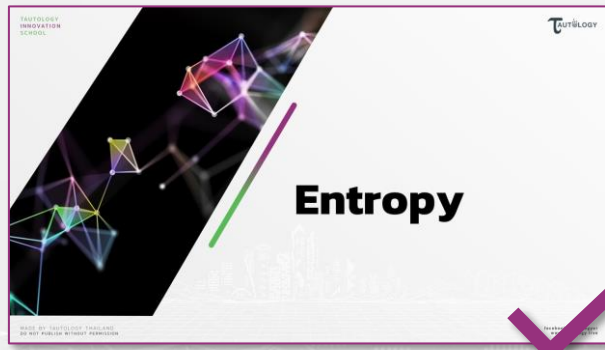
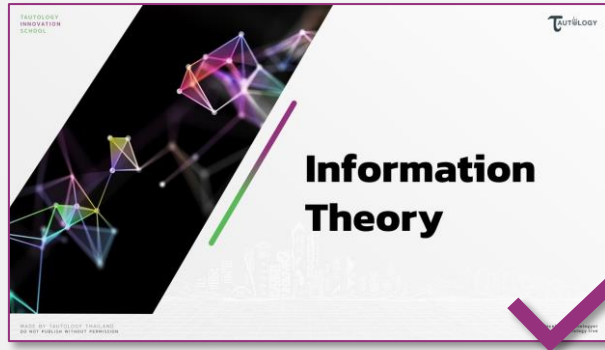


Entropy

ตัวอย่าง (2)

$$\begin{aligned} H(P) &= -\sum_{c=0}^1 p(x_c) \log(p(x_c)) \\ &= -p(x_0) \log(p(x_0)) - p(x_1) \log(p(x_1)) \\ &= -p(\text{Tail}) \log(p(\text{Tail})) - p(\text{Head}) \log(p(\text{Head})) \\ &= -0 \log(0) - 1 \log(1) \\ &= 0 \end{aligned}$$

Cross Entropy



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

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c))$$

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)

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c))$$

KL Divergence

**What is KL
Divergence?**



**Origin of the
Equation**



**KL as Cost
Function**

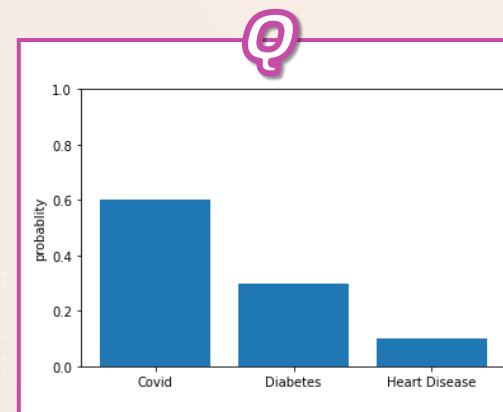
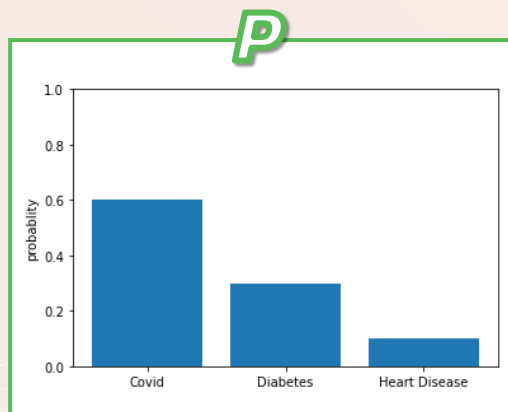


Origin of the Equation

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c))$$

Origin of the Equation

Q : Distribution Q แตกต่างจาก P
เท่าไร? จะต้องดูผ่านอะไรดี?



Origin of the Equation

ดูผ่าน **ratio** ของ distribution P และ Q

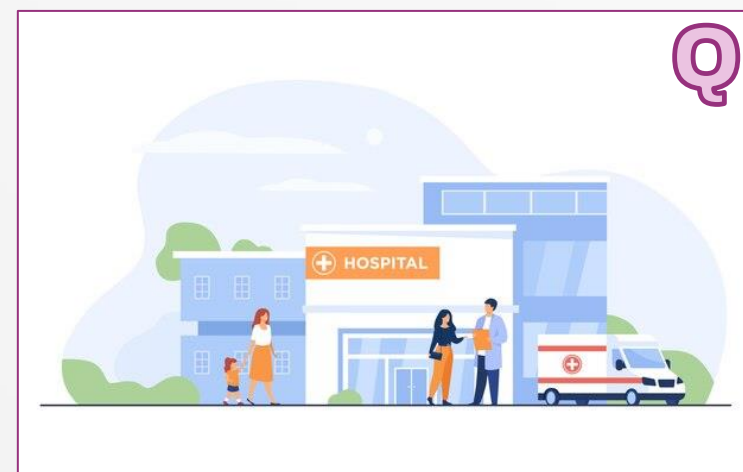
$$\frac{p(x_c)}{q(x_c)}$$

Origin of the Equation

Example 1



- $p(\text{โรคโควิด}) = 0.6$



- $q(\text{โรคโควิด}) = 0.6$

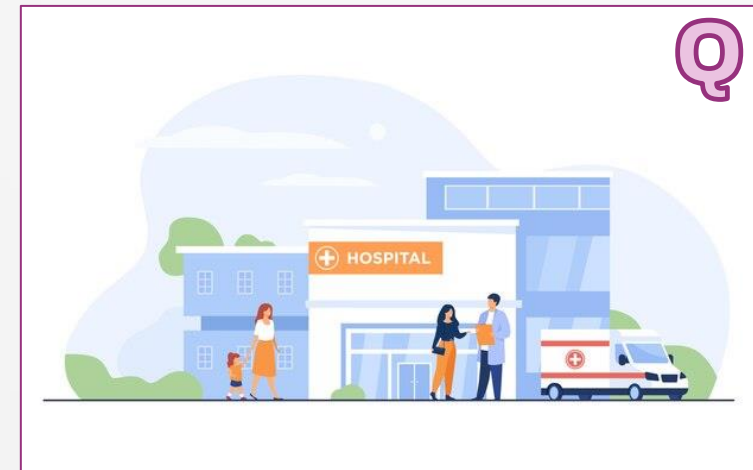
Distribution Q ต่างจาก $P : \frac{0.6}{0.6} = 1$

Origin of the Equation

Example 2



- $p(\text{โรคโควิด}) = 0.6$



- $q(\text{โรคโควิด}) = 0.3$

Distribution Q ต่างจาก $P : \frac{0.6}{0.3} = 2$

Origin of the Equation

NOTE

- ถ้าโอกาสการเกิดโควิด ใน distribution P และ Q **มีค่าเท่ากัน** ค่าของ ratio $\frac{p}{q} = 1$
- ถ้าโอกาสการเกิดโควิด ใน distribution P และ Q **มีค่าไม่เท่ากัน** ค่าของ ratio $\frac{p}{q} \neq 1$

Origin of the Equation

แล้วถ้าข้อมูลของเรามีหลาย class ละ?

Origin of the Equation

สำหรับข้อมูลที่มีหลาย class เราจะนำ ratio ของแต่ละ class มาคูณกัน

$$\frac{p(x_c)}{q(x_c)}$$



$$\prod_{c=0}^{k-1} \frac{p(x_c)}{q(x_c)}$$

Origin of the Equation

Example 3 (P และ Q เป็น distribution เดียวกัน)



- $p(\text{โควิด}) = 0.6$
- $p(\text{เบาหวาน}) = 0.3$
- $p(\text{หัวใจ}) = 0.1$



- $q(\text{โควิด}) = 0.6$
- $q(\text{เบาหวาน}) = 0.3$
- $q(\text{หัวใจ}) = 0.1$

Origin of the Equation

Example 3 (P และ Q เป็น distribution เดียวกัน)

$$\frac{p(\text{โศก})}{q(\text{โศก})}$$



$$\frac{0.6}{0.6}$$



$$1$$

$$\frac{p(\text{มหาหวาน})}{q(\text{มหาหวาน})}$$



$$\frac{0.3}{0.3}$$



$$1$$

$$\frac{p(\text{หิวใจ})}{q(\text{หิวใจ})}$$



$$\frac{0.1}{0.1}$$



$$1$$

Origin of the Equation

Example 3

Distribution Q ต่างจาก P :

$$\begin{aligned}\prod_{c=0}^2 \frac{p(x_c)}{q(x_c)} &= \frac{p(\text{โศภิต})}{q(\text{โศภิต})} \times \frac{p(\text{เมหาหวน})}{q(\text{เมหาหวน})} \times \frac{p(\text{ห้วงใจ})}{q(\text{ห้วงใจ})} \\ &= 1 \times 1 \times 1 \\ &= \mathbf{1}\end{aligned}$$

Origin of the Equation

Example 4 (P และ Q ไม่เป็น distribution เดียวกัน)



- $p(\text{โควิด}) = 0.5$
- $p(\text{เบาหวาน}) = 0.4$
- $p(\text{หัวใจ}) = 0.1$



- $q(\text{โควิด}) = 0.0001$
- $q(\text{เบาหวาน}) = 0.3999$
- $q(\text{หัวใจ}) = 0.6$

Origin of the Equation

Example 4 (P และ Q ไม่เป็น distribution เดียวกัน)

$$\frac{p(\text{โศกวิธ})}{q(\text{โศกวิธ})}$$



$$\frac{0.5}{0.0001}$$



5000

$$\frac{p(\text{เบาสหวาน})}{q(\text{เบาสหวาน})}$$



$$\frac{0.4}{0.3999}$$



1.00025

$$\frac{p(\text{หั่วใจ})}{q(\text{หั่วใจ})}$$



$$\frac{0.1}{0.6}$$



1.66667

Origin of the Equation

Example 4

Distribution Q ต่างจาก P :

$$\begin{aligned}\prod_{c=0}^2 \frac{p(x_c)}{q(x_c)} &= \frac{p(\text{โศภิต})}{q(\text{โศภิต})} \times \frac{p(\text{เมหาชวน})}{q(\text{เมหาชวน})} \times \frac{p(\text{ห้วงใจ})}{q(\text{ห้วงใจ})} \\ &= 5000 \times 1.0003 \times 1.6667 \\ &= \mathbf{8336.00005}\end{aligned}$$

Origin of the Equation

NOTE

- ถ้า P และ Q **เป็น** distribution เดียวกัน $\prod_{c=0}^{k-1} \frac{p(x_c)}{q(x_c)} = 1$
- ถ้า P และ Q **ไม่เป็น** distribution เดียวกัน $\prod_{c=0}^{k-1} \frac{p(x_c)}{q(x_c)} \neq 1$

Origin of the Equation

จาก **Example 4** จะเห็นได้ว่า ในกรณีที่ P และ Q **ไม่เป็น** distribution เดียวกัน ถ้าคำนวณด้วย $\prod_{c=0}^{k-1} \frac{p(x_c)}{q(x_c)}$ จะพบปัญหาว่า

- ① คำนวณค่อนข้างยาก
- ② ค่าที่คำนวณออกมาได้ค่อนข้างเยอะ และเป็นคนละ scale กับ probability function ทำให้ยากต่อการวิเคราะห์



Origin of the Equation

เพื่อแก้ปัญหาลำต้น จึงมีแนวคิดที่จะพัฒนาต่อมาคือ
“Take Log”

Origin of the Equation

$$\prod_{c=0}^{k-1} \frac{p(x_c)}{q(x_c)}$$



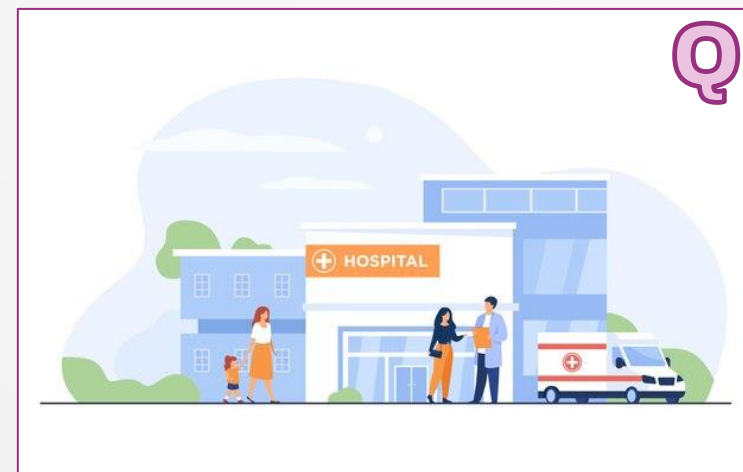
$$\sum_{c=0}^{k-1} \log \frac{p(x_c)}{q(x_c)}$$

Origin of the Equation

Example 1 (P และ Q เป็น distribution เดียวกัน)



- $p(\text{โควิด}) = 0.6$
- $p(\text{เบาหวาน}) = 0.3$
- $p(\text{หัวใจ}) = 0.1$



- $q(\text{โควิด}) = 0.6$
- $q(\text{เบาหวาน}) = 0.3$
- $q(\text{หัวใจ}) = 0.1$

Origin of the Equation

Example 1 (P และ Q เป็น distribution เดียวกัน)

$$\log \left(\frac{p(\text{โศกวิธ})}{q(\text{โศกวิธ})} \right)$$



$$\log \left(\frac{0.6}{0.6} \right) = \log 1$$



$$0$$

$$\log \left(\frac{p(\text{เขาสหวน})}{q(\text{เขาสหวน})} \right)$$



$$\log \left(\frac{0.3}{0.3} \right) = \log 1$$



$$0$$

$$\log \left(\frac{p(\text{ห้วงใจ})}{q(\text{ห้วงใจ})} \right)$$



$$\log \left(\frac{0.1}{0.1} \right) = \log 1$$



$$0$$

Origin of the Equation

Example 1

Distribution Q ต่างจาก P :

$$\begin{aligned}\sum_{c=0}^{k-1} \log \frac{p(x_c)}{q(x_c)} &= \log \frac{p(\text{โศภิต})}{q(\text{โศภิต})} + \log \frac{p(\text{โศภิต})}{q(\text{โศภิต})} + \log \frac{p(\text{โศภิต})}{q(\text{โศภิต})} \\ &= 0 + 0 + 0 \\ &= \mathbf{0}\end{aligned}$$

Origin of the Equation

Example 2 (P และ Q ไม่เป็น distribution เดียวกัน)



- $p(\text{โควิด}) = 0.5$
- $p(\text{เบาหวาน}) = 0.4$
- $p(\text{หัวใจ}) = 0.1$



- $q(\text{โควิด}) = 0.0001$
- $q(\text{เบาหวาน}) = 0.3999$
- $q(\text{หัวใจ}) = 0.6$

Origin of the Equation

Example 2 (P และ Q ไม่เป็น distribution เดียวกัน)

$$\log\left(\frac{p(\text{โศกวิถ})}{q(\text{โศกวิถ})}\right)$$



$$\log\left(\frac{0.5}{0.0001}\right) = \log 5000$$



$$8.517$$

$$\log\left(\frac{p(\text{มหาหวาน})}{q(\text{มหาหวาน})}\right)$$



$$\log\left(\frac{0.4}{0.3999}\right) = \log 1.00025$$



$$0.00025$$

$$\log\left(\frac{p(\text{ห้วงใจ})}{q(\text{ห้วงใจ})}\right)$$



$$\log\left(\frac{0.1}{0.6}\right) = \log 0.16667$$



$$-1.79174$$

Origin of the Equation

Example 2

Distribution Q ต่างจาก P :


$$\begin{aligned}\sum_{c=0}^{k-1} \log \frac{p(x_c)}{q(x_c)} &= \log \frac{p(\text{โศภิต})}{q(\text{โศภิต})} + \log \frac{p(\text{โศภิต})}{q(\text{โศภิต})} + \log \frac{p(\text{โศภิต})}{q(\text{โศภิต})} \\ &= 8.517 + 0.00025 - 1.79174 \\ &= \mathbf{6.72551}\end{aligned}$$

Origin of the Equation

NOTE


- ถ้า P และ Q **เป็น** distribution เดียวกัน $\sum_{c=0}^{k-1} \log \frac{p(x_c)}{q(x_c)} = 0$
- ถ้า P และ Q **ไม่เป็น** distribution เดียวกัน $\sum_{c=0}^{k-1} \log \frac{p(x_c)}{q(x_c)} > 0$

Origin of the Equation



เพื่อที่จะคำนวณว่า distribution Q แตกต่างจาก P เท่าไหร่ ในมุมมองของ distribution P เราหาค่าเฉลี่ยถ่วงน้ำหนักด้วย $p(x_c)$

Origin of the Equation


$$D_{KL}(P \parallel Q) = \sum_{c=0}^{k-1} p(x_c) \log \left(\frac{p(x_c)}{q(x_c)} \right)$$

Origin of the Equation

$$\begin{aligned} D_{KL}(P \parallel Q) &= \sum_{c=0}^{k-1} p(x_c) [\log(p(x_c)) - \log(q(x_c))] \\ &= \sum_{c=0}^{k-1} [p(x_c) \log(p(x_c)) - p(x_c) \log(q(x_c))] \end{aligned}$$

Origin of the Equation

$$\begin{aligned} D_{KL}(P \parallel Q) &= \sum_{c=0}^{k-1} p(x_c) \log(p(x_c)) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c)) \\ &= -H(P) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c)) \\ &(\because H(P) = -\sum_{c=0}^{k-1} p(x_c) \log(p(x_c))) \end{aligned}$$

Origin of the Equation

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c))$$

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

$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^{k-1} [y_{i,c} \log(\hat{y}_{i,c})]$$

KL as Cost Function

Model

| x_1 | x_2 | y_0 | y_1 | y_2 |
|----------|----------|----------|----------|----------|
| 0 | 1 | 1 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| \vdots | \vdots | \vdots | \vdots | \vdots |
| -1 | 0 | 0 | 0 | 1 |

ตารางแสดง dataset

| \hat{y}_0 | \hat{y}_1 | \hat{y}_2 |
|-------------|-------------|-------------|
| 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

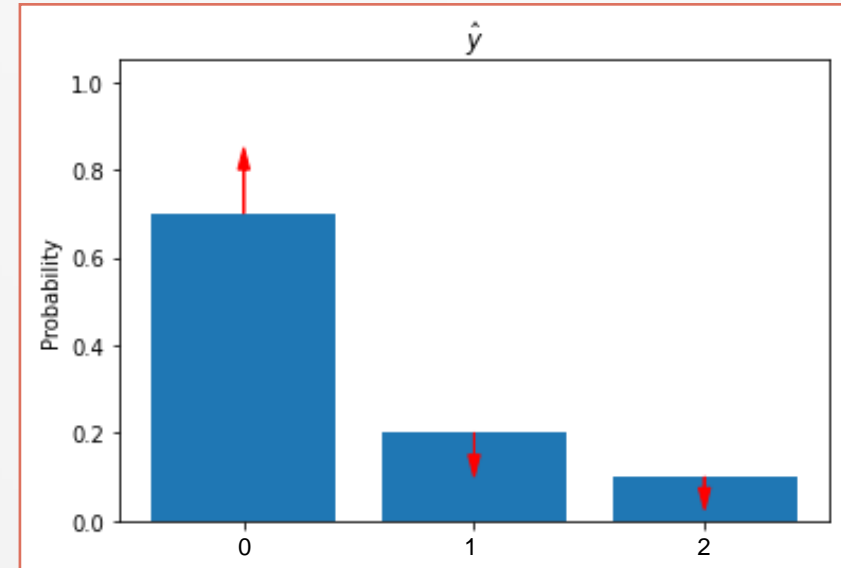
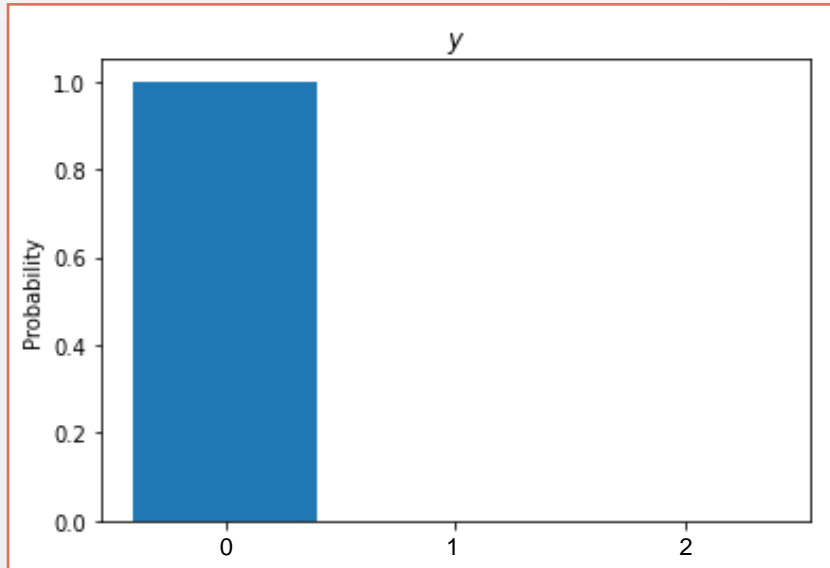
Model

| x_1 | x_2 | y_0 | y_1 | y_2 | | \hat{y}_0 | \hat{y}_1 | \hat{y}_2 |
|----------|----------|----------|----------|----------|--|-------------|-------------|-------------|
| 0 | 1 | 1 | 0 | 0 | | 0.5 | 0.3 | 0.2 |
| 1 | 0 | 0 | 1 | 0 | | 0.2 | 0.7 | 0.1 |
| \vdots | \vdots | \vdots | \vdots | \vdots | | \vdots | \vdots | \vdots |
| -1 | 0 | 0 | 0 | 1 | | 0.1 | 0.3 | 0.6 |

ตารางแสดง dataset

ตารางแสดง \hat{y} ที่ได้จาก model

KL as Cost Function



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

KL as Cost Function

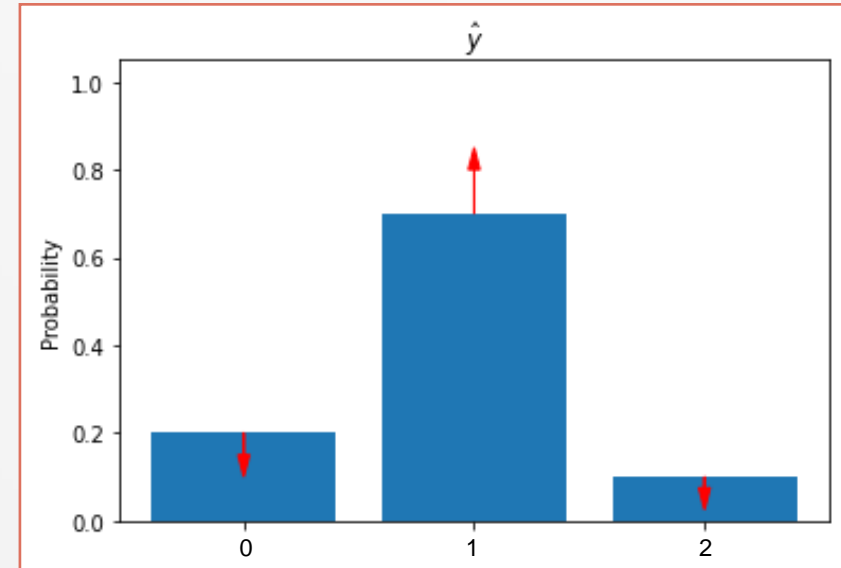
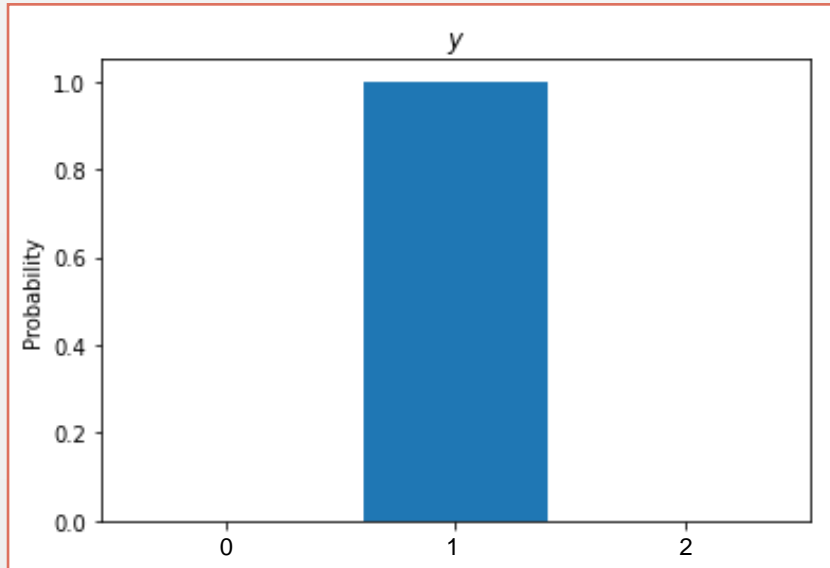
Model

| x_1 | x_2 | y_0 | y_1 | y_2 | | \hat{y}_0 | \hat{y}_1 | \hat{y}_2 |
|----------|----------|----------|----------|----------|--|-------------|-------------|-------------|
| 0 | 1 | 1 | 0 | 0 | | 0.5 | 0.3 | 0.2 |
| 1 | 0 | 0 | 1 | 0 | | 0.2 | 0.7 | 0.1 |
| \vdots | \vdots | \vdots | \vdots | \vdots | | \vdots | \vdots | \vdots |
| -1 | 0 | 0 | 0 | 1 | | 0.1 | 0.3 | 0.6 |

ตารางแสดง dataset

ตารางแสดง \hat{y} ที่ได้จาก model

KL as Cost Function



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

KL as Cost Function

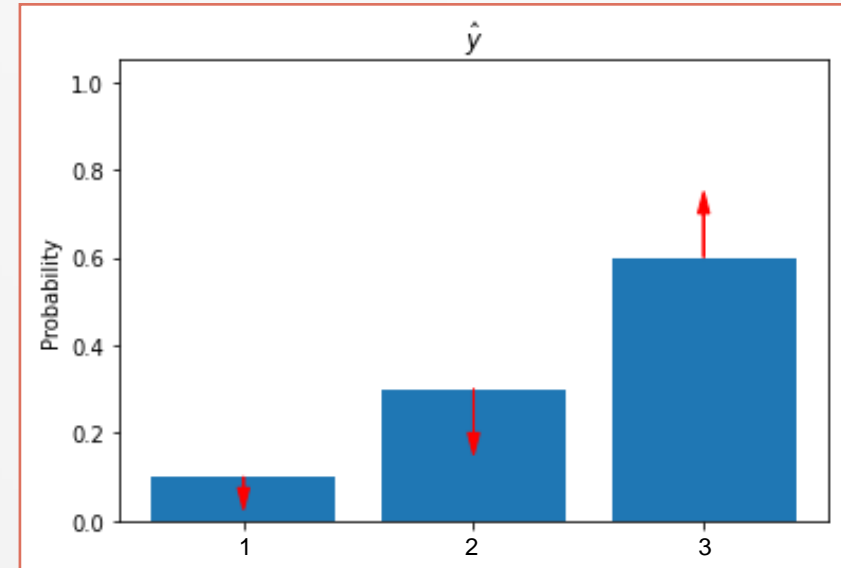
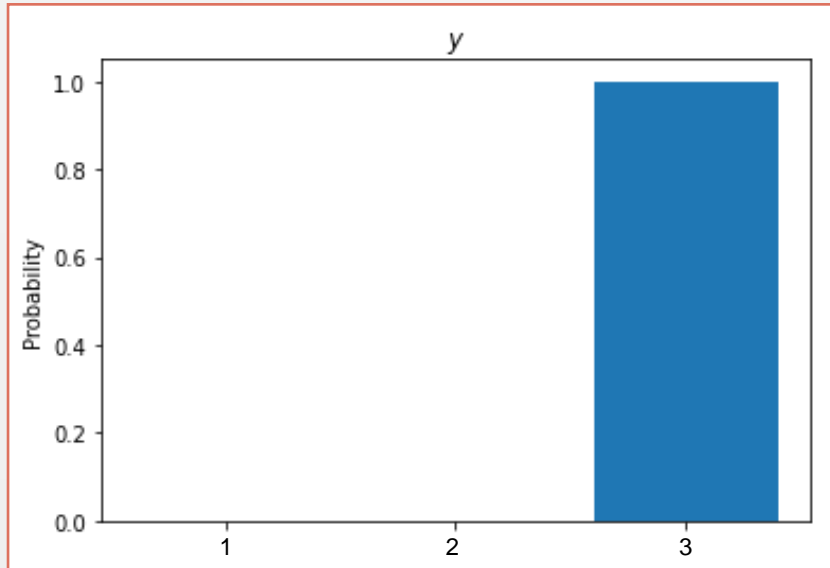
Model

| x_1 | x_2 | y_0 | y_1 | y_2 | \hat{y}_0 | \hat{y}_1 | \hat{y}_2 |
|----------|----------|----------|----------|----------|-------------|-------------|-------------|
| 0 | 1 | 1 | 0 | 0 | 0.5 | 0.3 | 0.2 |
| 1 | 0 | 0 | 1 | 0 | 0.2 | 0.7 | 0.1 |
| \vdots | \vdots | \vdots | \vdots | \vdots | \vdots | \vdots | \vdots |
| -1 | 0 | 0 | 0 | 1 | 0.1 | 0.3 | 0.6 |

ตารางแสดง dataset

ตารางแสดง \hat{y} ที่ได้จาก model

KL as Cost Function



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

KL as Cost Function

$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=0}^{k-1} p(x_c) \log(q(x_c))$$



$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -H(\mathbf{y}_i) - \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

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 |

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -H(\mathbf{y}_i) - \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

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

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

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 |

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto - \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

KL as Cost Function

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto - \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

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 |

$$D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto -\sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

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 |

$$\sum_{i=1}^n D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \propto - \sum_{i=1}^n \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

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 |

$$\min \sum_{i=1}^n D_{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \equiv \min - \sum_{i=1}^n \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

KL as Cost Function

$$\min - \sum_{i=1}^n \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

KL as Cost Function

เพื่อความสะดวกในการใช้ gradient descent เราจึงใช้
ค่าเฉลี่ยของ cross entropy ในการ train model

KL as Cost Function


$$\min -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

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


$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^{k-1} [y_{i,c} \log(\hat{y}_{i,c})]$$

KL as Cost Function

- พิจารณา *Cost* สำหรับ 2-class

$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^{k-1} y_{i,c} \log(\hat{y}_{i,c})$$

$$= -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^1 y_{i,c} \log(\hat{y}_{i,c})$$

$$= -\frac{1}{n} \sum_{i=1}^n [y_{i,0} \log(\hat{y}_{i,0}) + y_{i,1} \log(\hat{y}_{i,1})]$$

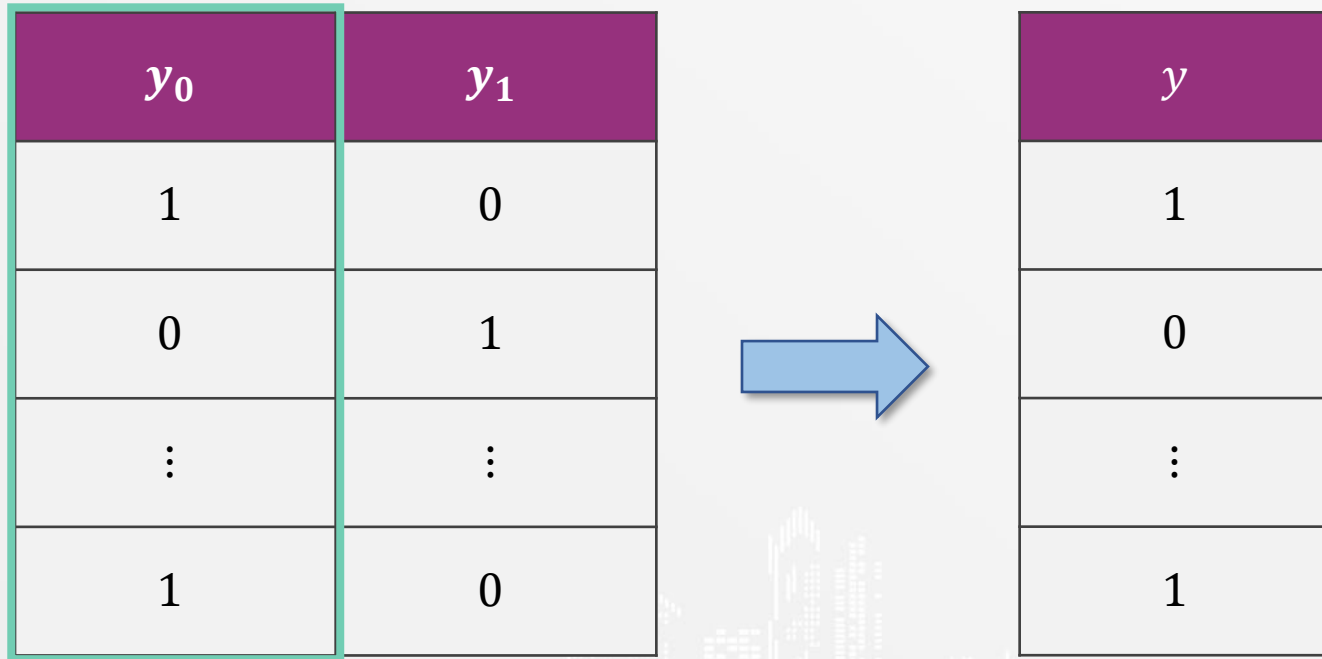
KL as Cost Function

- พิจารณา *Cost* สำหรับ 2-class

$$\begin{aligned} \text{Cost} &= -\frac{1}{n} \sum_{i=1}^n [y_{i,0} \log(\hat{y}_{i,0}) + y_{i,1} \log(\hat{y}_{i,1})] \\ &= -\frac{1}{n} \sum_{i=1}^n [y_{i,0} \log(\hat{y}_{i,0}) + (1 - y_{i,0}) \log(1 - \hat{y}_{i,0})] \end{aligned}$$

$$\begin{aligned} (\because y_{i,0} + y_{i,1} &= 1 \\ \hat{y}_{i,0} + \hat{y}_{i,1} &= 1) \end{aligned}$$

KL as Cost Function



$$\because y_0 + y_1 = 1$$

KL as Cost Function

| \hat{y}_0 | \hat{y}_1 | | \hat{y} |
|-------------|-------------|---|-----------|
| 0.7 | 0.3 |  | 0.7 |
| 0.2 | 0.8 | | 0.2 |
| \vdots | \vdots | | \vdots |
| 0.6 | 0.4 | | 0.6 |

$$\because \hat{y}_0 + \hat{y}_1 = 1$$


KL as Cost Function

- พิจารณา *Cost* สำหรับ 2-class


$$\begin{aligned} \text{Cost} &= -\frac{1}{n} \sum_{i=1}^n [y_{i,0} \log(\hat{y}_{i,0}) + (1 - y_{i,0}) \log(1 - \hat{y}_{i,0})] \\ &= -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \end{aligned}$$

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


$$Cost = -\frac{1}{n} \sum_{i=1}^n \sum_{c=0}^{k-1} [y_{i,c} \log(\hat{y}_{i,c})]$$

KL Divergence

**What is KL
Divergence?**



**Origin of the
Equation**



**KL as Cost
Function**



Cross Entropy

