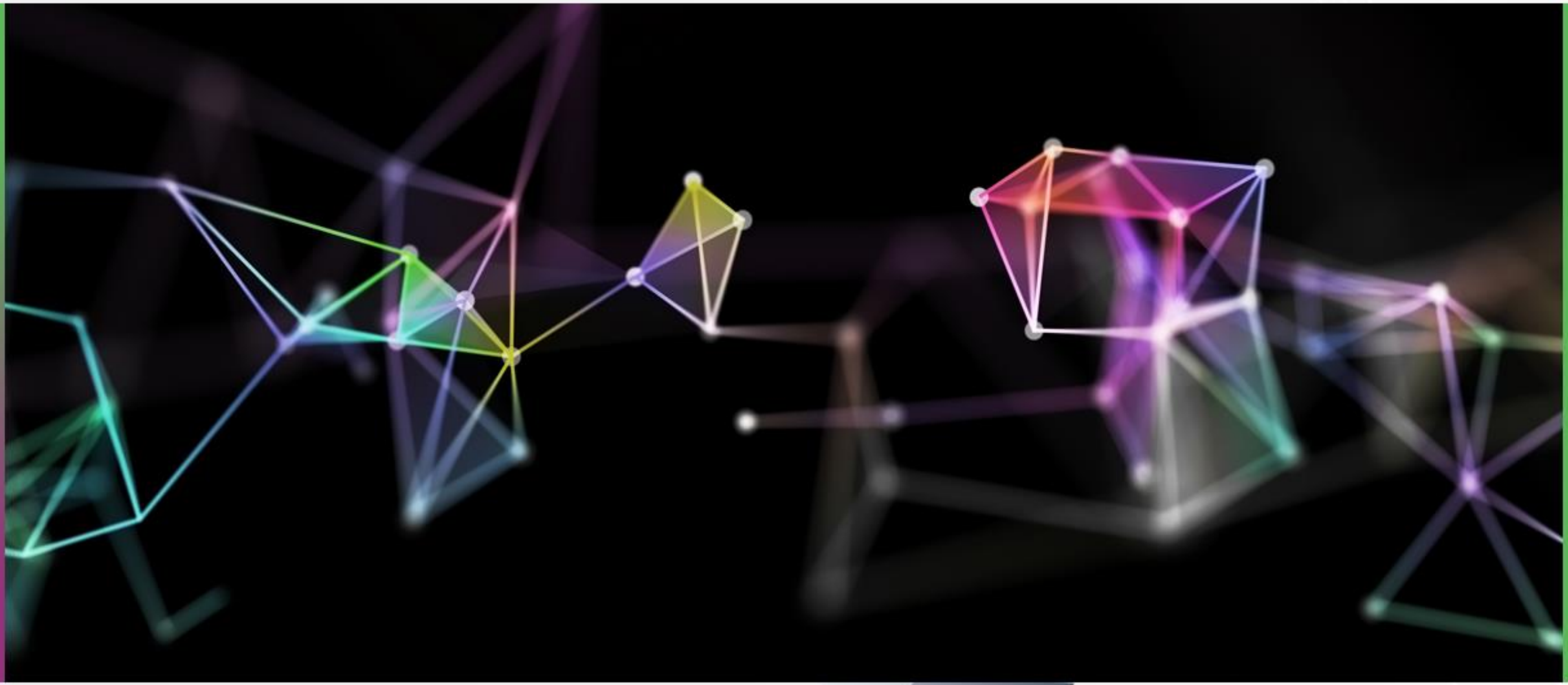
TAUTOLOGY
INNOVATION
SCHOOL

TAUTOLOGY

CROSS ENTROPY



CROSS ENTROPY

BY TAUTOLOGY

facebook/tautologyai
www.tautology.live

Cross Entropy

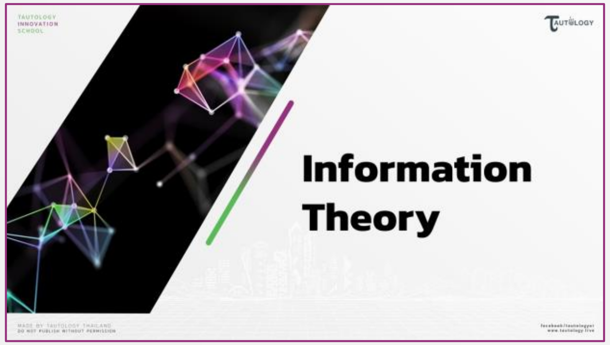
- 2-class

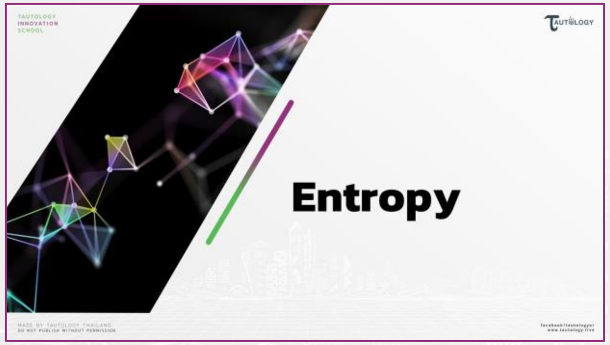
$$Cost = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

- Multi-class

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

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. ยิ่งเหตุการณ์มีโอกาสเกิดขึ้นน้อยเท่าไหร่ 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

$$\boxed{I(x) = f(p(x))}$$

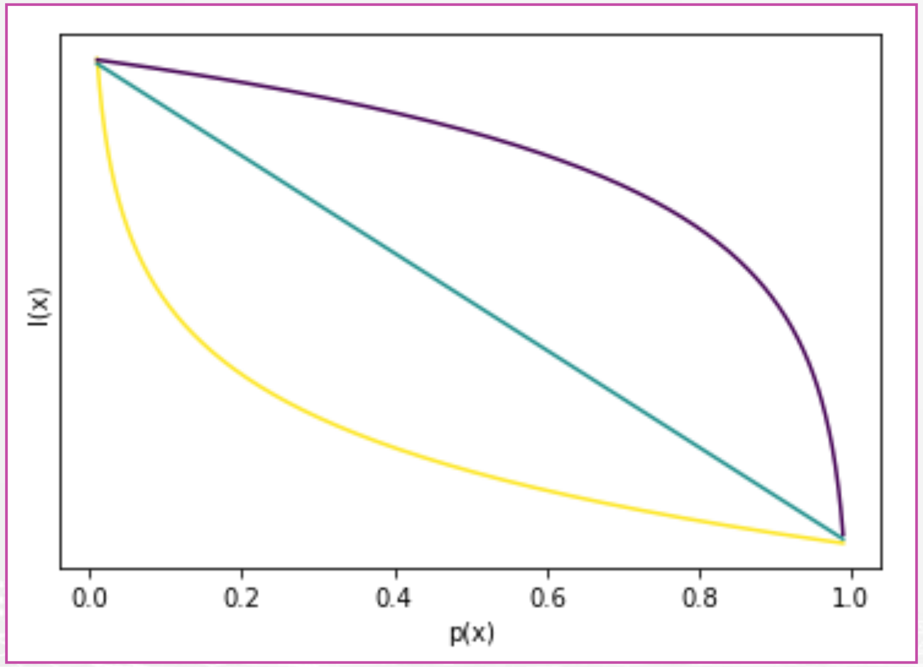
Formation

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

ถ้า $p(x) = 1$ แล้ว $I(x) = f(1) = 0$

Formation

จาก definition ข้อที่ 2 “ยิ่งเหตุการณ์มีโอกาสเกิดขึ้นน้อยเท่าไหร่ 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

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

Formation

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

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

Formation

Function เพียงอันเดียวที่มีคุณสมบัติ

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

Formation

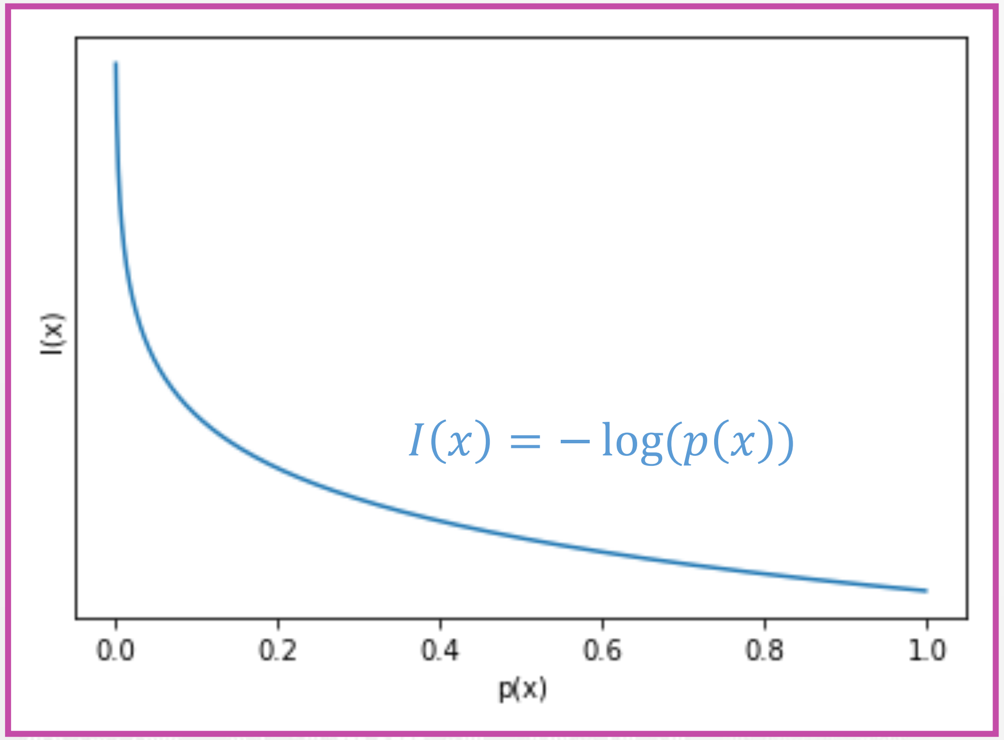
$$f\big(p(x)\big) = -\log(p(x))$$

Formation

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

Formation

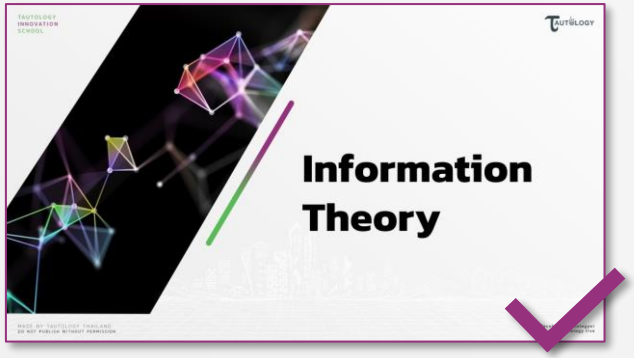
ดังนั้น

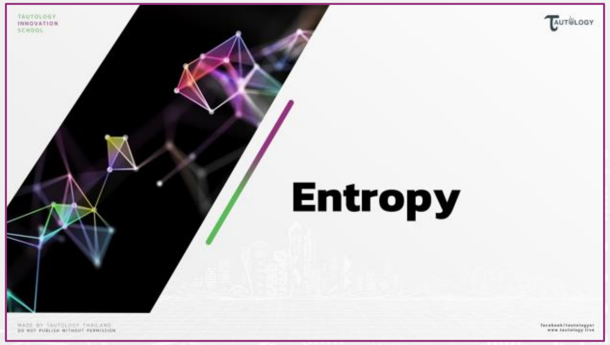


Information Theory

- Concept
- Definition
- Formation

Cross Entropy





Uncertainty

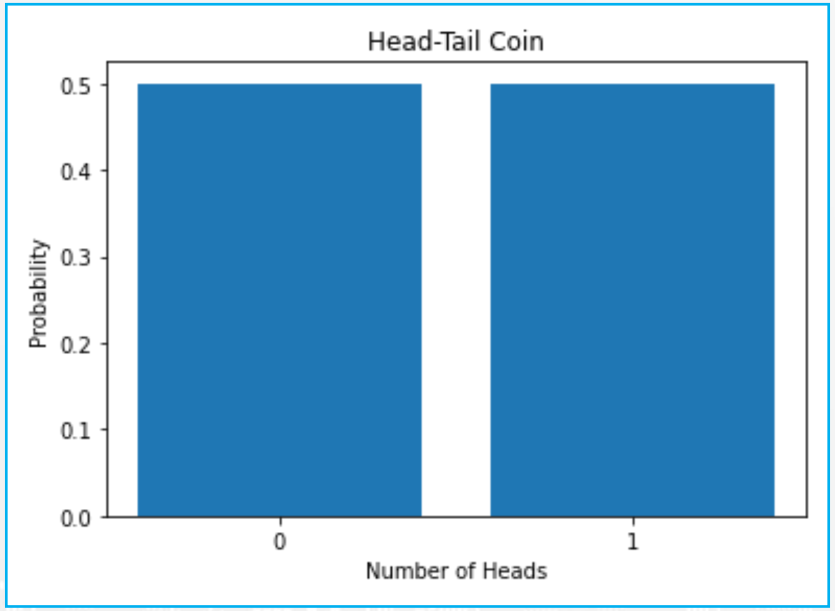
Uncertainty

Uncertainty คือ ค่าที่ใช้บอกความไม่เป็นระเบียบ/ความยุ่งเหยิงของระบบ ซึ่งเป็นอีกหนึ่งชื่อเรียกของ information

uncertainty = information

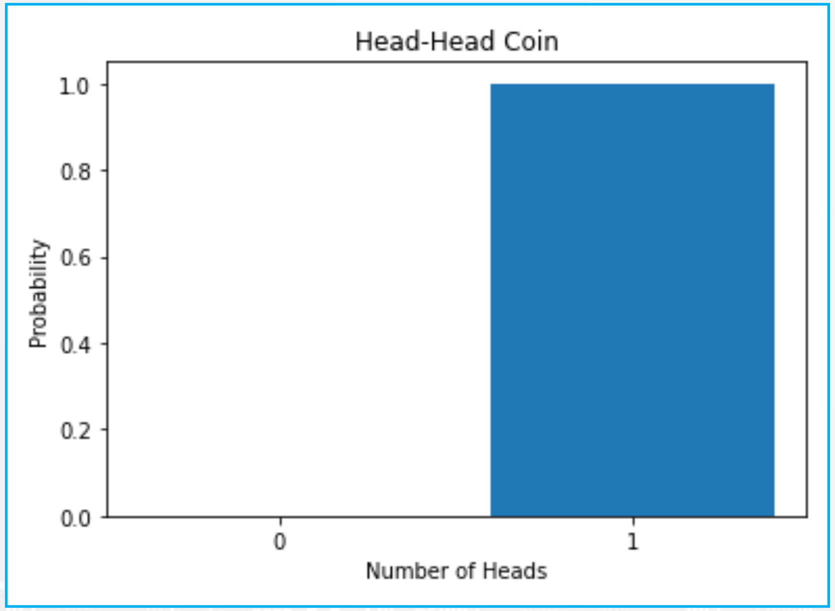
Uncertainty

- <u>ระบบที่มีความยุ่งเหยิง</u>

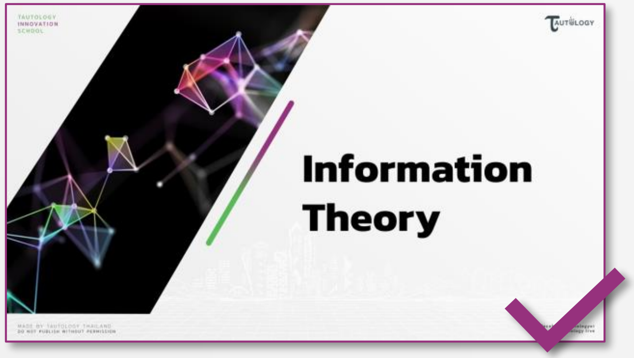


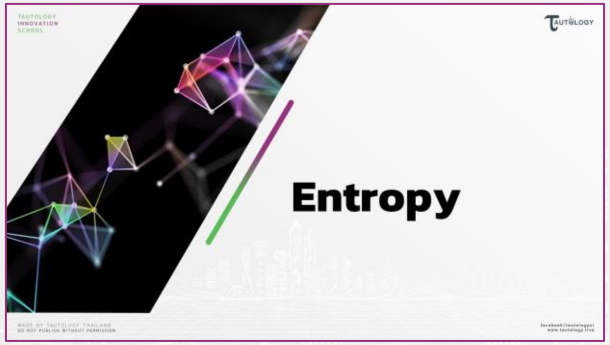
Uncertainty

- ระบบที่ไม่มีความยุ่งเหยิง



Cross Entropy





Entropy

Entropy

Entropy คือ ค่าที่บอกถึงค่าเฉลี่ยของ information หรือ uncertainty ในระบบ

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

Entropy

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

$$= E[-\log(p(x))]$$

$$= -E[\log(p(x))]$$

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

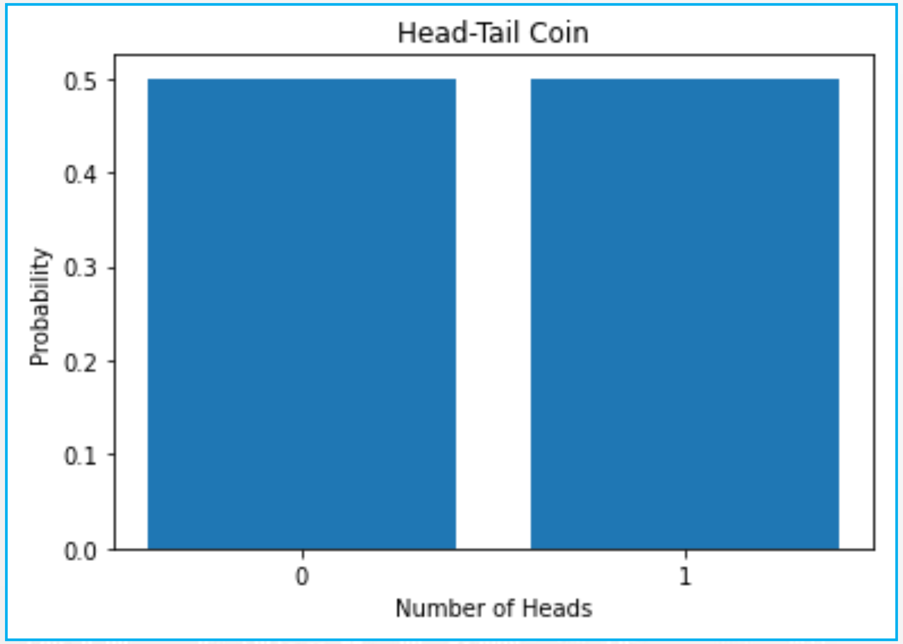
Entropy

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)

$$H(P) = -\sum_{c=1}^{2} p(x_c)\log(p(x_c))$$

$$= -p(x_1)\log(p(x_1)) - p(x_2)\log(p(x_2))$$

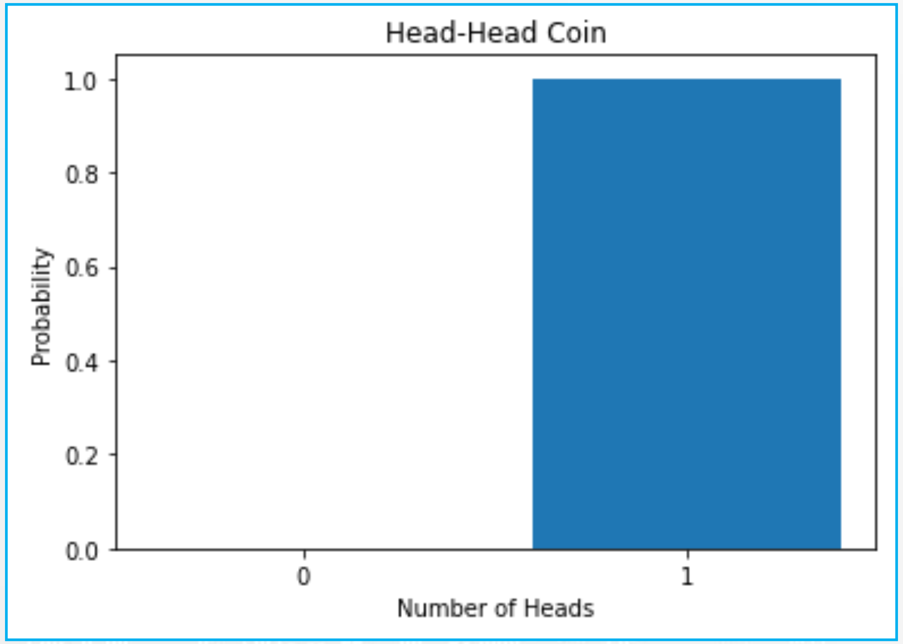
$$= -p(0)\log(p(0)) - p(1)\log(p(1))$$

$$= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$$

$$= 0.6931$$

Entropy

ตัวอย่าง (2)



Entropy

ตัวอย่าง (2)

$$H(P) = -\sum_{c=1}^{2} p(x_c)\log(p(x_c))$$

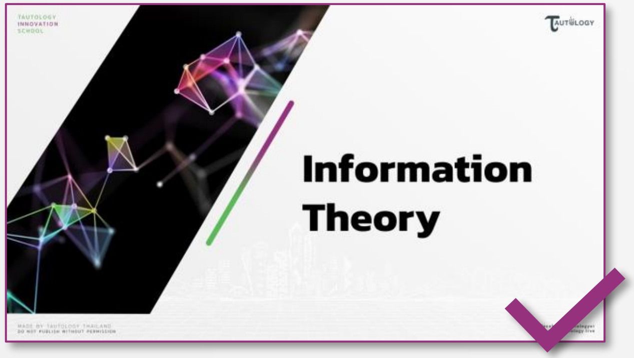
$$= -p(x_1)\log(p(x_1)) - p(x_2)\log(p(x_2))$$

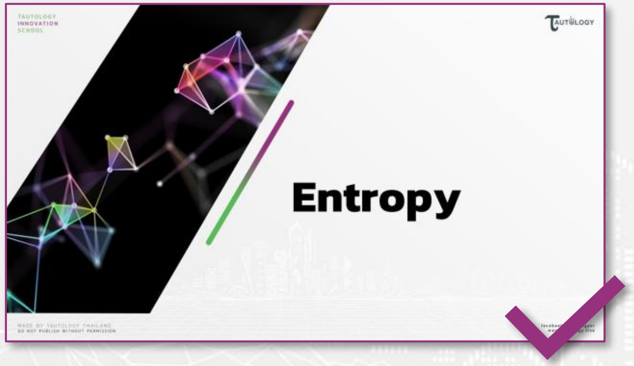
$$= -p(0)\log(p(0)) - p(1)\log(p(1))$$

$$= -0\log(0) - 1\log(1)$$

$$= 0$$

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

entropy = ค่าที่ใช้วัดความไม่แน่นอนในระบบ

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

D_{KL} //

What is KL Divergence?

ถ้า distribution เหมือนกัน
→ ค่าน้อย
ถ้า dis___ ต่างกัน → ค่ามาก

- 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)

ใน sense ของการสร้าง model เราต้องการให้ distribution (y, $\hat{y}$ เหมือนกัน)

☆ KL มีค่าน้อยๆ ☆

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

KL Divergence

What is KL Divergence?

Origin of the Equation

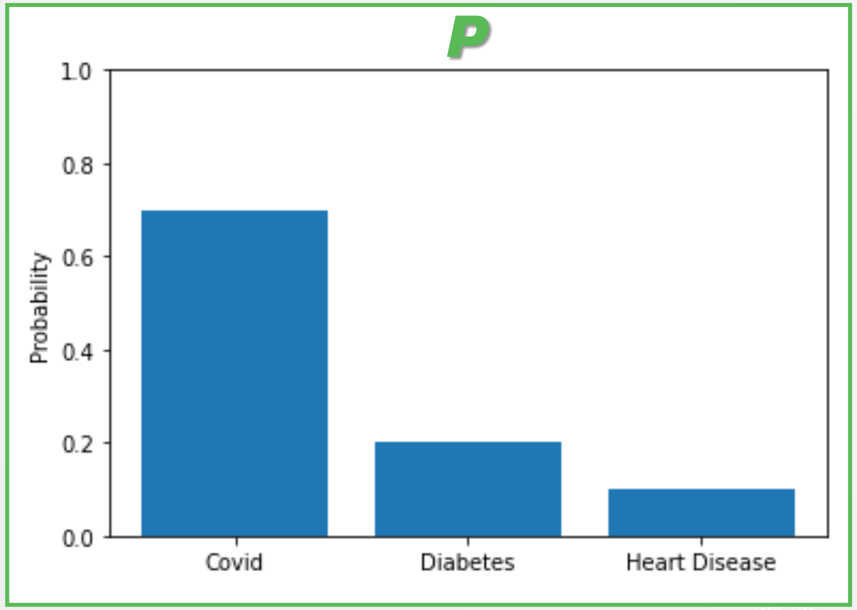
KL as Cost Function

Origin of the Equation

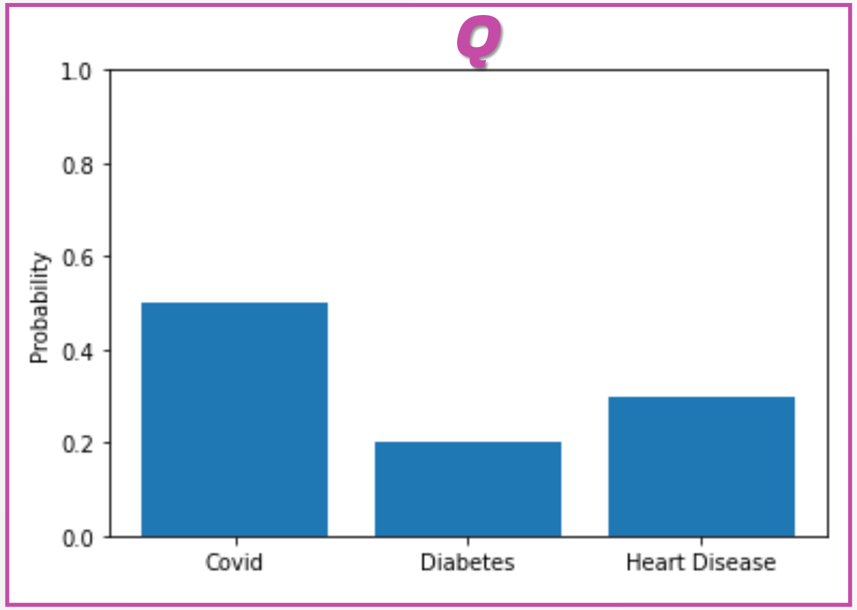
$$D_{KL}(P \parallel Q) = -H(P) - \sum_{c=1}^{k} p(x_c) \log(q(x_c))$$

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

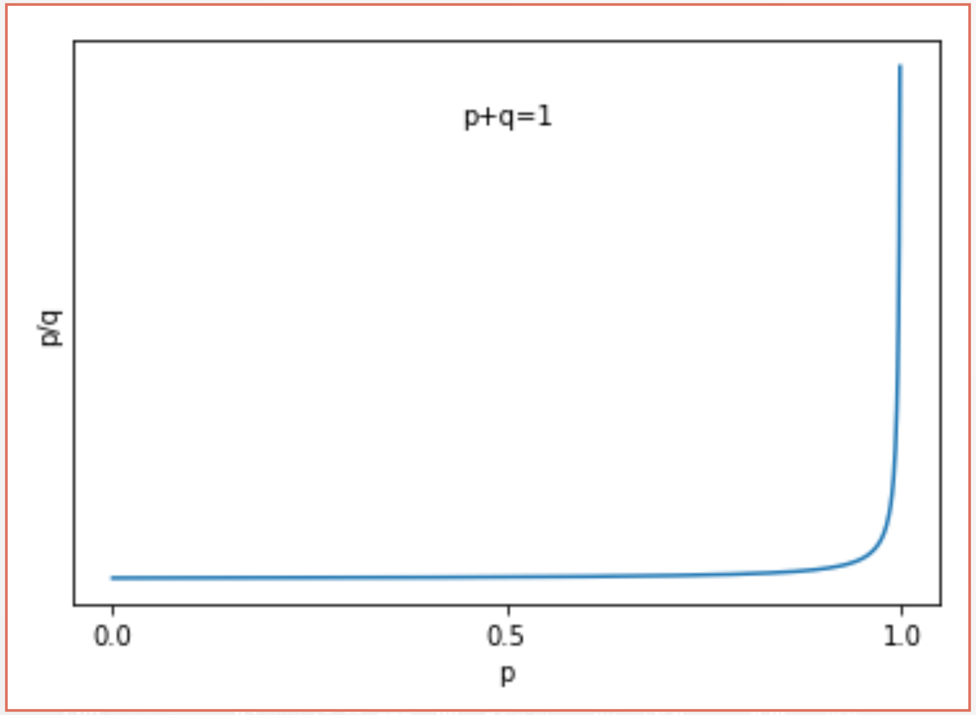
$$\frac{p(\text{โควิด})}{q(\text{โควิด})} \; \frac{p(\text{โควิด})}{q(\text{โควิด})} \; \frac{p(\text{เบาหวาน})}{q(\text{เบาหวาน})} \; \frac{p(\text{โควิด})}{q(\text{โควิด})} \; \frac{p(\text{หัวใจ})}{q(\text{หัวใจ})} \; \cdots$$

Origin of the Equation

$$\frac{p(\text{โควิด})}{q(\text{โควิด})} \; \frac{p(\text{โควิด})}{q(\text{โควิด})} \; \frac{p(\text{เบาหวาน})}{q(\text{เบาหวาน})} \; \frac{p(\text{โควิด})}{q(\text{โควิด})} \; \frac{p(\text{หัวใจ})}{q(\text{หัวใจ})} \; \cdots$$

หาค่าเฉลี่ย

Origin of the Equation



Origin of the Equation

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

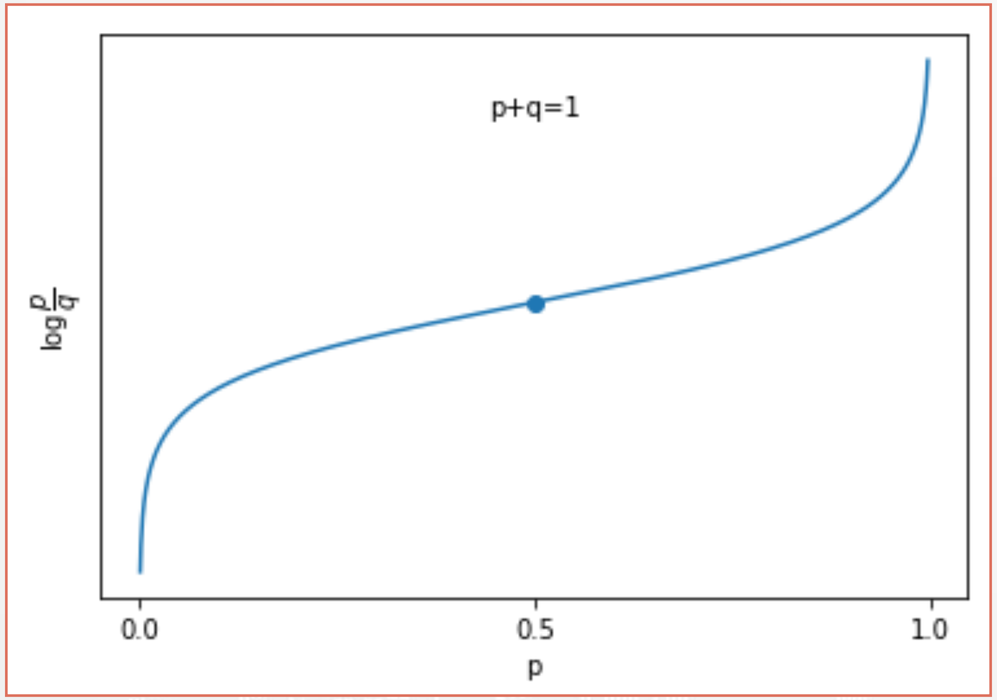
หาค่าเฉลี่ย ❌

Origin of the Equation

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

หาค่าเฉลี่ย

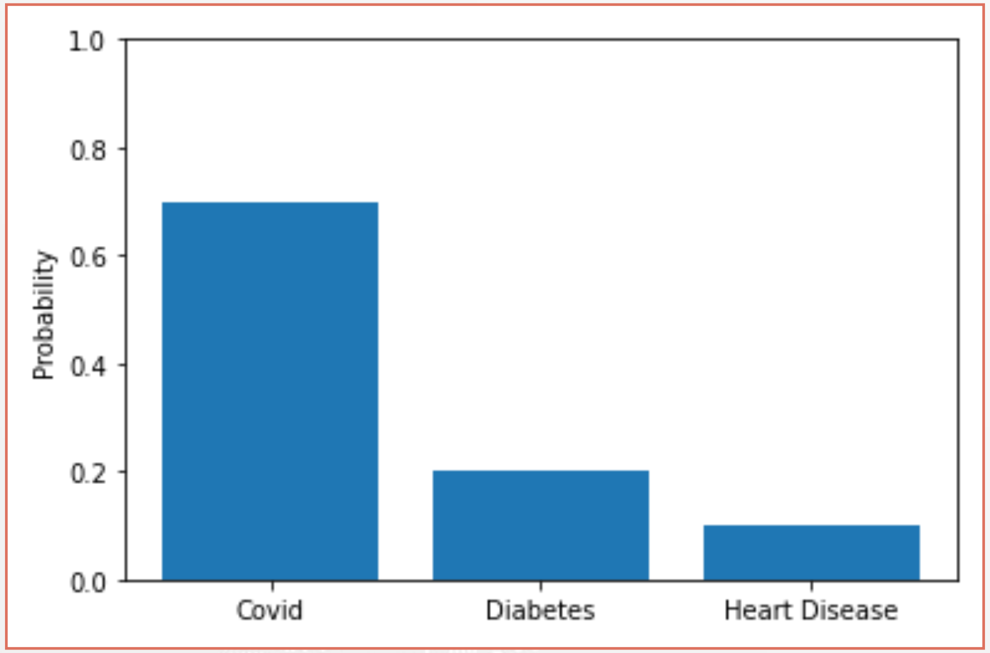
Origin of the Equation



Origin of the Equation

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

Origin of the Equation



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

Origin of the Equation

$$
\begin{aligned}
D_{KL}(P \parallel Q) = \frac{1}{n}\Bigg[& p(\text{โควิด}) \cdot n \cdot \log\left(\frac{p(\text{โควิด})}{q(\text{โควิด})}\right) \\
& + p(\text{เบาหวาน}) \cdot n \cdot \log\left(\frac{p(\text{เบาหวาน})}{q(\text{เบาหวาน})}\right) \\
& + p(\text{หัวใจ}) \cdot n \cdot \log\left(\frac{p(\text{หัวใจ})}{q(\text{หัวใจ})}\right)\Bigg]
\end{aligned}
$$

Origin of the Equation

$$D_{KL}(P \parallel Q) = p(\text{โควิด}) \cdot \log\left(\frac{p(\text{โควิด})}{q(\text{โควิด})}\right)$$
$$+p(\text{เบาหวาน}) \cdot \log\left(\frac{p(\text{เบาหวาน})}{q(\text{เบาหวาน})}\right)$$
$$+p(\text{หัวใจ}) \cdot \log\left(\frac{p(\text{หัวใจ})}{q(\text{หัวใจ})}\right)$$

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

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

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

Origin of the Equation

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

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

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

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

for LoR, NN, DL

- 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

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

KL as Cost Function

x_1	x_2	y_1	y_2	y_3
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
−1	0	0	0	1

ตารางแสดง dataset

Model

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

ตารางแสดง $\hat{y}$ ที่ได้จาก model

TAUTOLOGY

KL as Cost Function

พิจารณาทีละ 1 sample

Model

x_1	x_2	y_1	y_2	y_3
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
−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
⋮	⋮	⋮
0.1	0.3	0.6

ตารางแสดง $\hat{y}$ ที่ได้จาก model

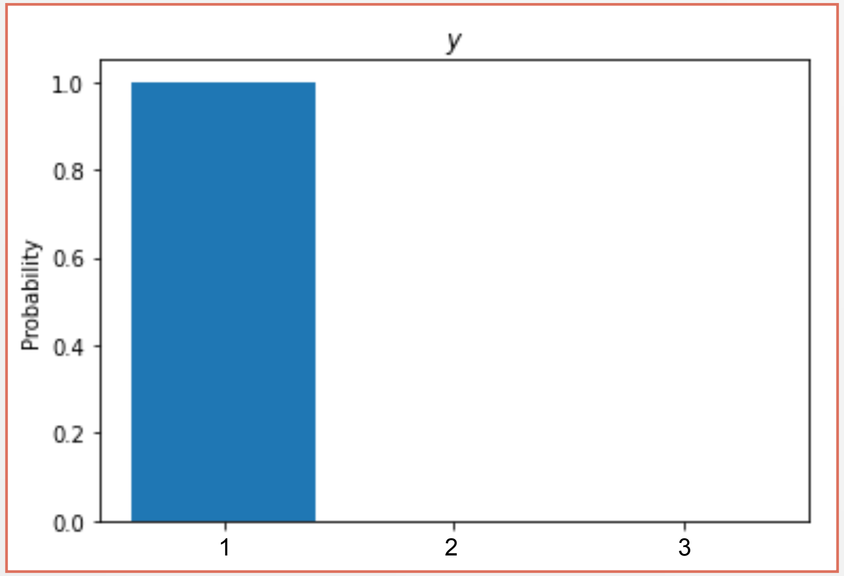
minimize cost $\rightarrow W \rightarrow Z_1 \rightarrow \hat{y}_1$

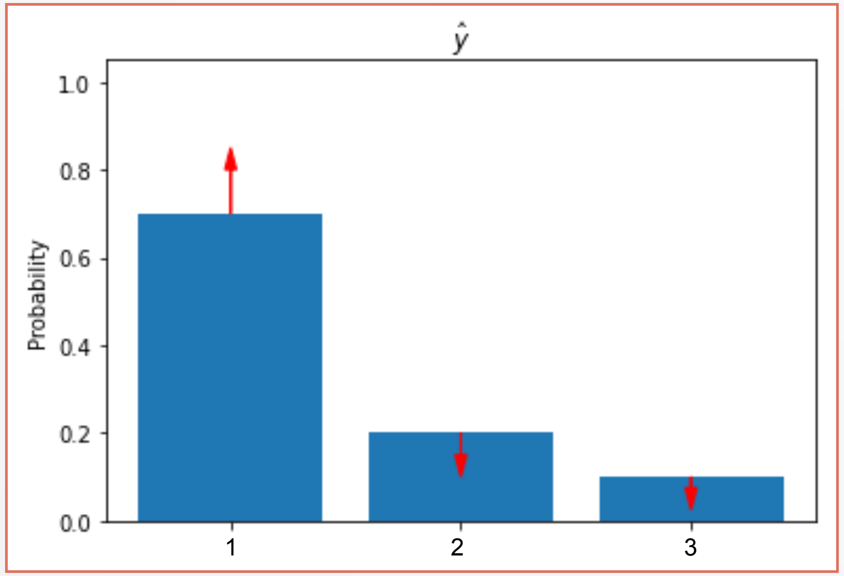
$Z_2 \rightarrow \hat{y}_2$

$Z_3 \rightarrow \hat{y}_3$

KL as Cost Function

sample 1





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

TAUTOLOGY

KL as Cost Function

Model

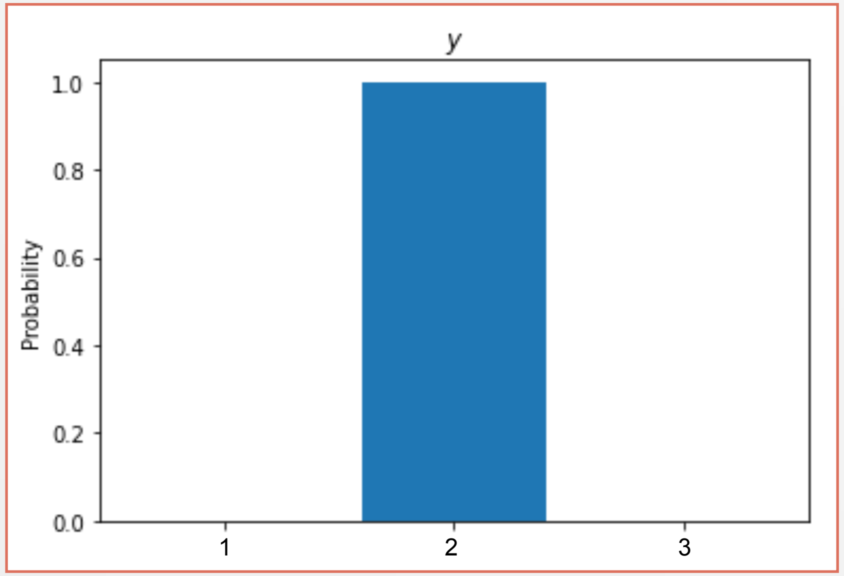
x_1	x_2	y_1	y_2	y_3
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
-1	0	0	0	1

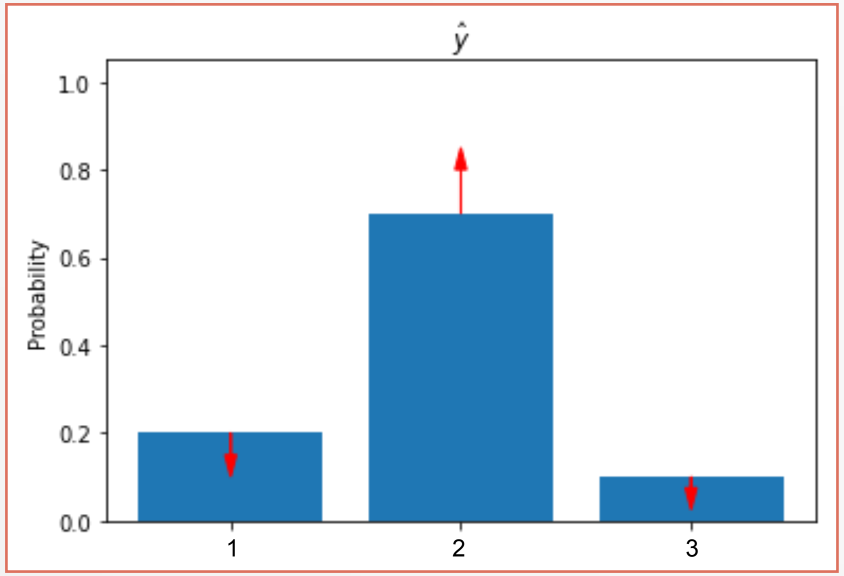
ตารางแสดง dataset

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

ตารางแสดง $\hat{y}$ ที่ได้จาก model

KL as Cost Function





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

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
⋮	⋮	⋮	⋮	⋮
−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
⋮	⋮	⋮
0.1	0.3	0.6

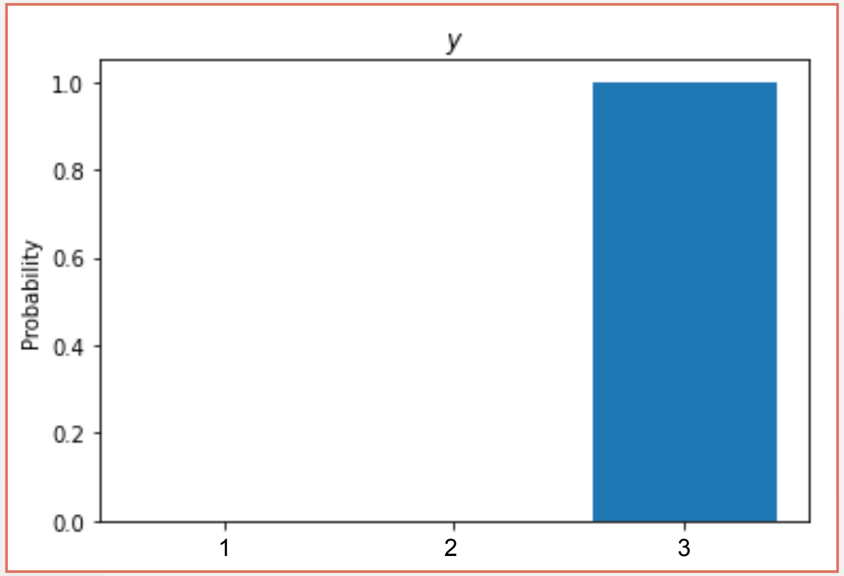
ตารางแสดง $\hat{y}$ ที่ได้จาก model

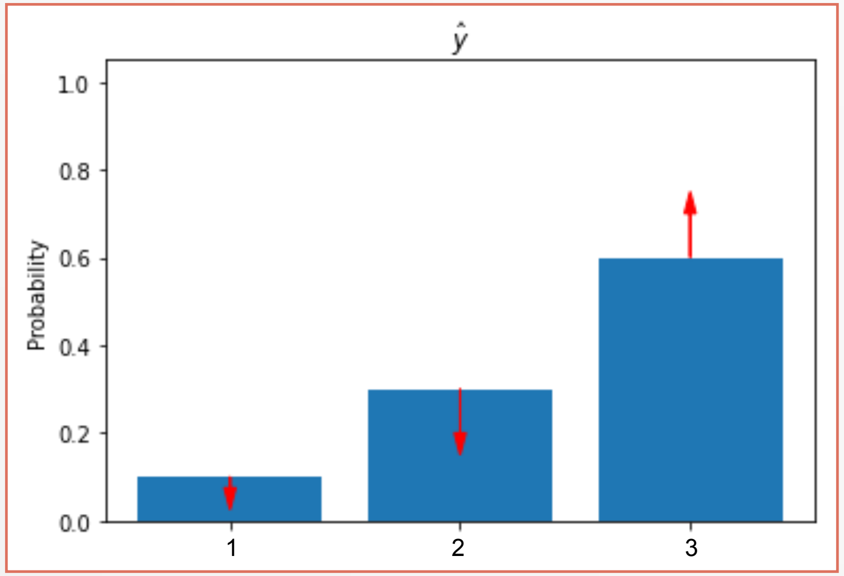
KL as Cost Function

P

sample ที่ 3

Q





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

KL as Cost Function

$P \rightarrow \vec{y}_i$

$Q \rightarrow \hat{\vec{y}}_i$

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

ค.น่าจะเป็น ที่จะเป็น class c ของ sample ที่ i (data)

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

ค.น่าจะเป็น ที่จะเป็น class c ของ sample ที่ i (model)

y ของ sample ที่ i

เพราะ y แต่ละตัวถูกใช้เป็นตัวแทน distribution

KL as Cost Function

KL = Entropy + Cross Entropy

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
−1	0	0	0	1

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

KL as Cost Function

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
0	1	1	0	0
1	0	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-1	0	0	0	1

ค่าคงที่

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

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

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

KL as Cost Function

สำคัญ

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

เป็นสูตรใช้วัด ค. ต่าง dis
สำหรับ 1 sample

KL as Cost Function

x_1	x_2	y_1	y_2	y_3
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
-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
⋮	⋮	⋮
0.1	0.3	0.6

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

KL as Cost Function

x_1	x_2	y_1	y_2	y_3
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
−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
⋮	⋮	⋮
0.1	0.3	0.6

ผลรวมความแตกต่างของ distr— ระหว่าง y_i กับ $\hat{y}_i$
ของทุก sample

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

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)

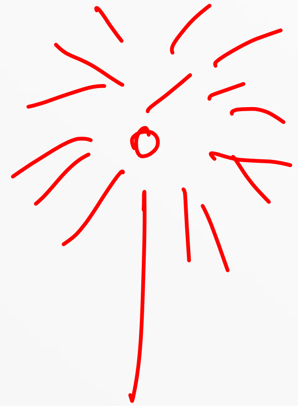
KL as Cost Function

x_1	x_2	y_1	y_2	y_3
0	1	1	0	0
1	0	0	1	0
⋮	⋮	⋮	⋮	⋮
−1	0	0	0	1

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

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

KL as Cost Function



Cross Entropy

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

$(\vec{y}_i, \hat{\vec{y}}_i)$

☆ พยายามทำให้ distribution ของแต่ละ sample เหมือนกันมากที่สุด

KL as Cost Function

เพื่อความสะดวกในการใช้ gradient descent เราจึงใช้

ค่าเฉลี่ยของ cross entropy ในการ train model

KL as Cost Function

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

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

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

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

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

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

$$(\because y_{i,1} + y_{i,2} = 1$$

$$\hat{y}_{i,1} + \hat{y}_{i,2} = 1)$$

KL as Cost Function

เป็น ไม่เป็น

y_1	y_2
1	0
0	1
⋮	⋮
1	0

เป็น

y
1
0
⋮
1

$\because y_1 + y_2 = 1$

KL as Cost Function

เป็น $\hat{y}_1$	ไม่เป็น $\hat{y}_2$
0.7	0.3
0.2	0.8
⋮	⋮
0.6	0.4

เป็น $\hat{y}$
0.7
0.2
⋮
0.6

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

KL as Cost Function

- พิจารณา $Cost$ สำหรับ 2-class

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

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

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

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

KL Divergence

What is KL Divergence?

Origin of the Equation

KL as Cost Function

สรุป

1. KL Divergence ใช้วัดความแตกต่างกัน distribution
2. KL $= -$Entropy $+$ Cross Entropy
3. Entropy เป็นค่าคงที่ เพราะ ถูกคำนวณมาจาก $(\vec{y}_i)$
4. KL $\propto$ Cross Entropy
5. min KL $\equiv$ min Cross Entropy
6. CE (2-class) $\leftarrow$ derive CE (general) ให้อยู่ในรูปแบบที่ specific

☆ เราทำการ minimize ค.แตกต่างระหว่าง 2 distri___ Y กับ $\hat{Y}$ ให้น้อยที่สุด (distri___ เหมือนกันมากที่สุด) !

Cross Entropy