[SWCON253] Machine Learning – Lec.07b

Information Theory
(Cross-Entropy & KL-Divergence)

김휘용

hykim.v@khu.ac.kr

경희대학교
KYUNG HEE UNIVERSITY

Contents

References

- “*Schaum's Outline of Probability, Random Variables, and Random Processes*,” by Hwei P. Hsu
- "*기계학습*" by 오일석

1. Information

메시지가 지닌 정보를 수량화할 수 있나?

- "고비 사막에 눈이 왔다"와 "대관령에 눈이 왔다"라는 두 메시지 중 어느 것이 더 많은 정보를 가지나?
- 정보이론의 기본 원리 → 확률이 작을수록 (불확실성이 클 수록) 많은 정보

자기 정보self information

- **Definition:** 사건(메시지) x_i의 **정보량** (단위: 비트 또는 나츠) → 불확실성(uncertainty)을 계량화

$$I(x_i) = \log_b \frac{1}{P(x_i)} = -\log_b P(x_i) \qquad \boldsymbol{X} \in \{x_i\}_{i=1..m}$$

- **Properties**

$$I(x_i) = 0 \quad \text{for} \quad P(x_i) = 1$$
$$I(x_i) \geq 0$$
$$I(x_i) > I(x_j) \quad \text{if} \quad P(x_i) < P(x_j)$$
$$I(x_i x_j) = I(x_i) + I(x_j) \quad \text{if } x_i \text{ and } x_j \text{ are independent}$$

2. Entropy

◆ 엔트로피 (평균 정보량)

- **Definition:** ***Average Information*** of a random variable $\boldsymbol{X} \in \{x_i\}_{i=1..m}$

$$H(X) = E[I(x_i)] = \sum_{i=1}^{m} P(x_i) I(x_i)$$

$$= -\sum_{i=1}^{m} P(x_i) \log_2 P(x_i) \quad \text{b/symbol}$$

- **Property:**

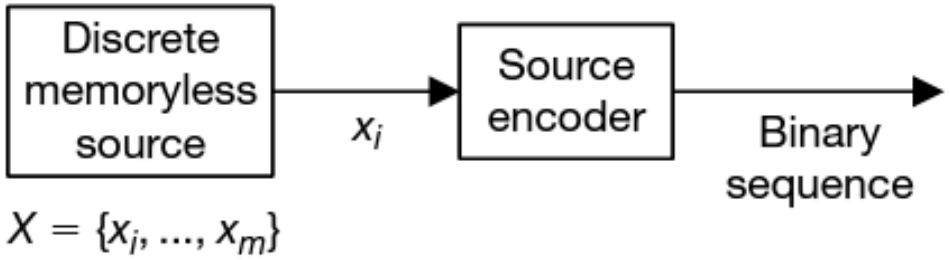
$0 \le H(X) \le \log_2 m$ (*m*: the number of symbols of the source *X*)

★ The maximum occurs when all the symbols are equally likely.

3. Source Coding Theorem

Source Coding

A conversion of the output of a DMS into a sequence of binary symbols (binary code word) is called *source coding*. The device that performs this conversion is called the *source encoder* (Fig. 10-7).



An objective of source coding is to minimize the average bit rate required for representation of the source by reducing the redundancy of the information source.

cf. "memoryless" current output does not depend on previous outputs.

3. Source Coding Theorem (cont'd)

Average Code Length

Let X be a DMS with finite entropy $H(X)$ and an alphabet $\{x_1, \ldots, x_m\}$ with corresponding probabilities of occurrence $P(x_i)(i = 1, \ldots, m)$. Let the binary code word assigned to symbol x_i by the encoder have length n_i, measured in bits. The length of a code word is the number of binary digits in the code word. The average code word length L, per source symbol, is given by

$$L = \sum_{i=1}^{m} P(x_i) n_i$$

Source Coding Theorem

The source coding theorem states that for a DMS X with entropy $H(X)$, the average code word length L per symbol is bounded as (Prob. 10.39)

$$L \geq H(X) = -\sum_{i=1}^{m} P(x_i) \log_2 P(x_i)$$

3. Source Coding Theorem (cont'd)

- **Example:**
  - FLC (Fixed Length Coding) vs. VLC (Variable Length Coding)

X	*a*	*b*	*c*	*d*	*e*	*f*	*g*
P(X)	24/32	2/32	2/32	1/32	1/32	1/32	1/32
I(X)	0.42	4	4	5	5	5	5
FLC (n_X)	000 (3)	001 (3)	010 (3)	011 (3)	100 (3)	101 (3)	110 (3)
VLC (n_X)	0 (1)	10 (2)	110 (3)	1110 (4)	11110 (5)	111110 (6)	1111110 (7)

$\Rightarrow H(X) = E[I(X)] \approx 1.44$

$\Rightarrow L = 3$

$\Rightarrow L \approx 1.75$

4. Cross-Entropy & KL Divergence

Cf. Entropy, Cross-Entropy, KL-Divergence : https://www.youtube.com/watch?v=ErfnhcEV1O8

◆ 교차 엔트로피와 상대 엔트로피

- DMS $X = \{x_1, \dots, x_m\}$에 대한 두 개의 확률분포 $p(X)$와 $q(X)$를 생각하자.
- p: true pdf, q: our guess or approximation

- 이때, 확률분포 p에 대한 확률분포 q의 ***Cross-Entropy*** (교차 엔트로피)를 다음과 같이 정의 한다.

$$H(p,q) = E_p[I_q(X)] = E_p[-\log(q(X))] = -\sum_{i=1}^{m} p(x_i)\log(q(x_i))$$

- 이때, 확률분포 q에서 p로의 ***KL Divergence*** (상대 엔트로피)는 다음과 같이 정의 된다.

$$D_{KL}(p||q) = E_p\left[\log\left(\frac{p(X)}{q(X)}\right)\right] = E_p[-\log(q(X))] - E_p[\log(p(X))] = H(p,q) - H(p,p) \geq 0$$

$$D_{KL}(p||q) = \sum_{i=1}^{m} p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right)$$

4. Cross-Entropy & KL Divergence (cont'd)

Example

X	*a*	*b*	*c*	*d*	*e*	*f*	*g*
p(X)	24/32	2/32	2/32	1/32	1/32	1/32	1/32
$l_p(X)$	0.42	4	4	5	5	5	5
q(X)	16/32	4/32	4/32	4/32	2/32	1/32	1/32
$l_q(X)$	1	3	3	3	4	5	5

$H(p,p) \approx 1.44$

$H(p,q) \approx ?$

$D_{KL}(p\|q) = H(p,q) - H(p,p) \approx ?$

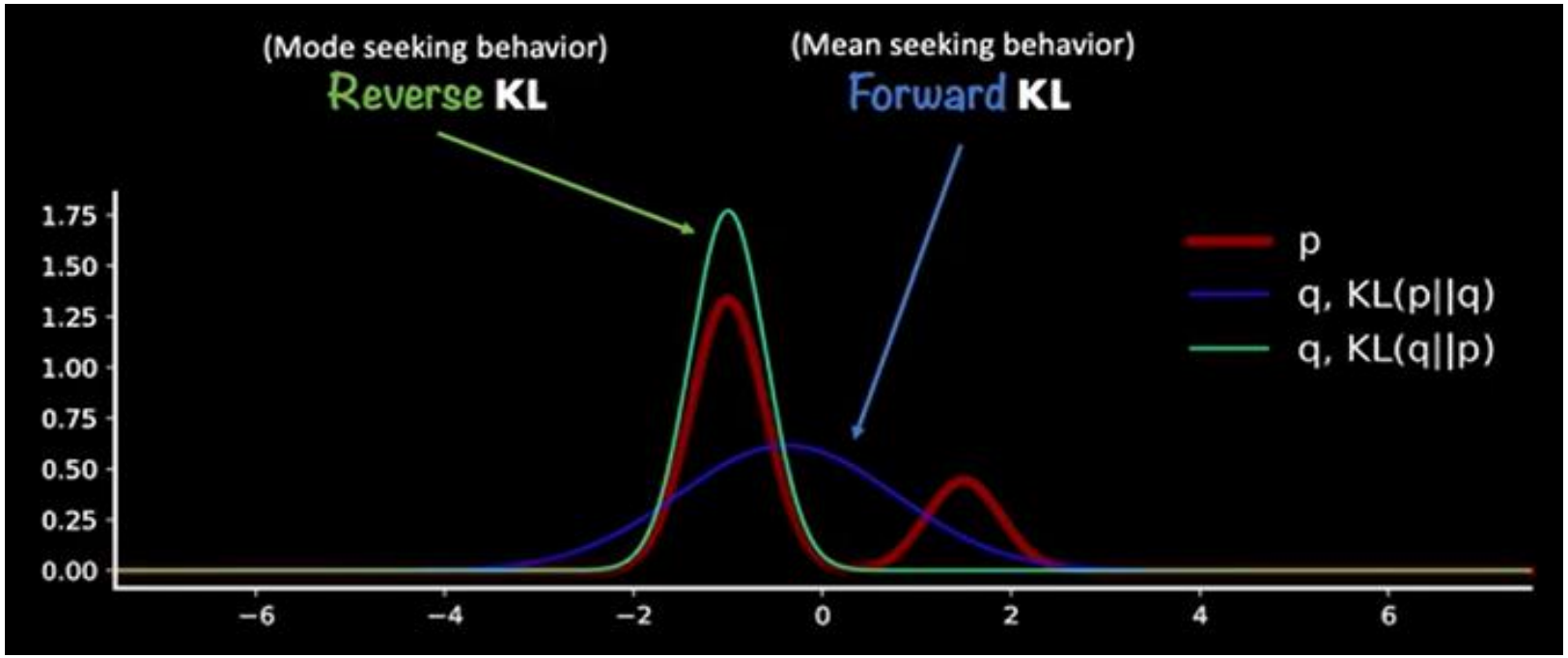
$D_{KL}(p\|q)$: 최소비트량 차이

$H(p,q)$: 늘어난 최소 비트량

$H(p,p)$: 최소 비트량

4. Cross-Entropy & KL Divergence (cont'd)

- **KL-Divergence에서 순서가 중요할까?**
  - Let $p(x)$ and $q(x)$ be *true* distribution and *approximated* distribution of x, respectively.
  - Minimizing the ***Forward KLD*** (i.e., $D_{KL}[\, p \,||\, q \,] = E_p[\, \log(\, p(x) / q(x) \,) \,]$)
    ★ *Mean-seeking* behavior (∵ $p(x) \approx 0$ 일때의 오차를 무시하기 때문)
  - Minimizing the ***Reverse KLD***: (i.e., $D_{KL}[\, q \,||\, p \,] = E_q[\, \log(\, q(x) / p(x) \,) \,]$)
    ★ *Mode-seeking* behavior (∵ $q(x) \approx 0$ 일때의 오차를 무시하기 때문)



Why?
"zero-avoiding" 때문

https://www.youtube.com/watch?v=9_eZHt2qJs4

5. Cross-Entropy Loss in ML

For Binary Classification (# classes=2)

- *Cross-Entropy* for a binary source $\{x_1, x_2\}$
  - ★ p: true pdf
  - ★ q: our guess or approximation of p

$$H(p,q) = -p(x_1)\log\big(q(x_1)\big) - p(x_2)\log\big(q(x_2)\big)$$
$$= -p(x_1)\log\big(q(x_1)\big) - (1-p(x_1))\log\big(1-q(x_1)\big)$$

- *Binary Cross-Entropy Loss*
  - ★ $y \in \{0,1\}$: true label (*true probability*)
  - ★ $h(x)$: our *predicted probability* ($0 \le h(x) \le 1$)

$$H(y,h(x)) = -y\log\big(h(x)\big) - (1-y)\log\big(1-h(x)\big)$$

For Multi-class Classification (# classes=*K*)

- *Multinomial Cross-Entropy*

$$H(p,q) = -\sum_{i=1}^{K} p(x_i)\log(q(x_i))$$

- *Multinomial Cross-Entropy Loss*

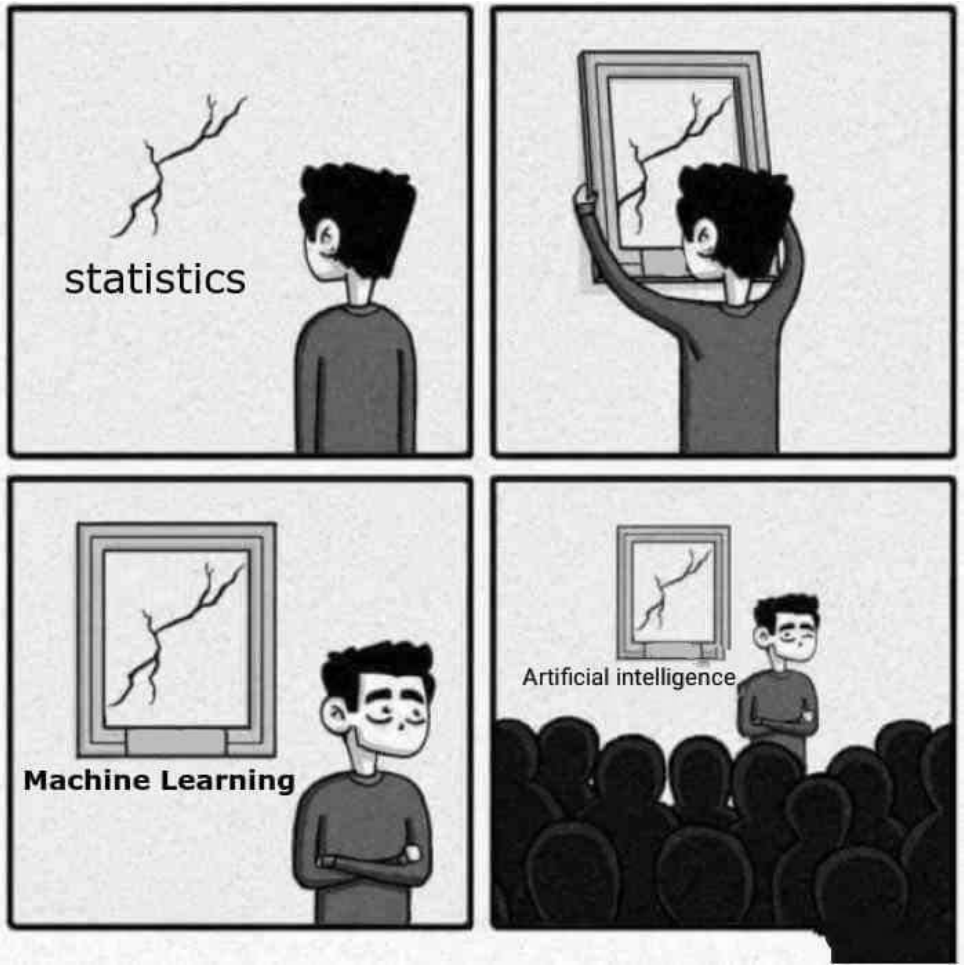
$$H(y_i,h(x_i)) = -\sum_{i=1}^{K} y_i \log(h(x_i))$$

  - ★ $y_i \in \{0,1\}$: true label (*true probability*)
  - ★ $h(x_i)$: our *predicted probability* ($0 \le h(x_i) \le 1$)

Remarks

- In ML context, **'log'** usually denotes ***'natural logarithm'***

Q & A



https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3