Probability & Information Theory

Jin Young Choi
Seoul National University

Outline

- Conditional Probability
- Chain Rule
- Independence
- Total Probability
- Bayes Rule
- Information
- Entropy
- Mutual Information
- Kullback Leibler Divergence
- Cross-Entropy

Conditional probability

- Given an event occurred, just as probability changes to conditional probability, the pmf of a random variable changes to the conditional pmf, cpmf.

- example: X is the random variable corresponding to the number on a playing card the opponent is putting face down; B is the event that my hand consists of A(1), 3, 4, 8, and Q(12).

- Does the event B affect the pmf of X?

Conditional probability

- conditional pmf, cpmf:

$$p_{X|Y}(x|y) := \frac{p_{XY}(x,y)}{p_Y(y)}$$

$$X = x \in A, \qquad Y = y \in B$$

- not defined when $p_Y(y) = 0.$

Conditional probability

- example: X is the random variable corresponding to the number on a playing card the opponent is putting face down; B is the event that my hand consists of A(1), 3, 4, 8, and Q(12).

- In the above example, the pmf that is uniformly 1/13(=4/52) changes to the cpmf that is 3/47 for 1, 3, 4, 8, 12 and 4/47 for the rest of the values.

Conditional probability

- conditional pmf, cpmf:

$$p_{X|YZ}(x|y,z) = p_{XYZ}(x,y,z)/p_{YZ}(y,z)$$

$$p_{XY|Z}(x,y|z) = p_{XYZ}(x,y,z)/p_{Z}(z)$$

- independence

$$X,Y \text{ indep } \Rightarrow p_{X|Y}(x|y) = p_X(x)$$

Conditional probability

- chain rule:
  - $p_{XY}(x,y) = p_X(x)\, p_{Y|X}(y|x)$
  - $p_{X_1 \ldots X_k}(x_1, \cdots, x_k)$
    $= p_{X_1}(x_1)\, p_{X_2|X_1}(x_2|x_1) \cdots p_{X_k|X_1 \ldots X_{k-1}}(x_k|x_1, \cdots, x_{k-1})$
  - $p_{XY|W}(x,y|w) = p_{X|W}(x|w)\, p_{Y|WX}(y|w,x)$
  - $p_{XYZ|W}(x,y,z|w) = p_{XY|W}(x,y|w)\, p_{Z|WXY}(z|w,x,y)$

Conditional probability

- total probability law:

$$P(X = x) = p_X(x) = \Sigma_y \, p_{XY}(x, y) = \Sigma_y \, p_{X|Y}(x|y) \, p_Y(y)$$

Conditional probability

- example: X : life expectancy of a 70-year-old.

  - blood condition after 70 years old

    H : having high blood pressure, $P(H) = 2/5$

    R : having normal blood pressure, $P(R) = 3/5$

  - at every year after 70 years old

    survival probability of high blood person: 9/10

    survival probability of normal blood person:19/20

  - What is the probability that a person lives until 90 years old ?

Conditional probability

- example:

$$p_X(x) = p_{X|H}(x)P(H) + p_{X|R}(x)P(R)$$

$$p_{X|H}(x) = \begin{cases} \frac{1}{10}\left(\frac{9}{10}\right)^{x-1}, & x = 1, 2, \cdots \\ 0, & \text{else} \end{cases} \sim \text{geo}(1/10)$$

$$p_{X|R}(x) = \begin{cases} \frac{1}{20}\left(\frac{19}{20}\right)^{x-1}, & x = 1, 2, \cdots \\ 0, & \text{else} \end{cases} \sim \text{geo}(1/20)$$

$$p_X(x) = p_{X|H}(x)P(H) + p_{X|R}(x)P(R)$$

$$= \frac{1}{10}\left(\frac{9}{10}\right)^{x-1} \cdot \frac{2}{5} + \frac{1}{20}\left(\frac{19}{20}\right)^{x-1} \cdot \frac{3}{5}$$

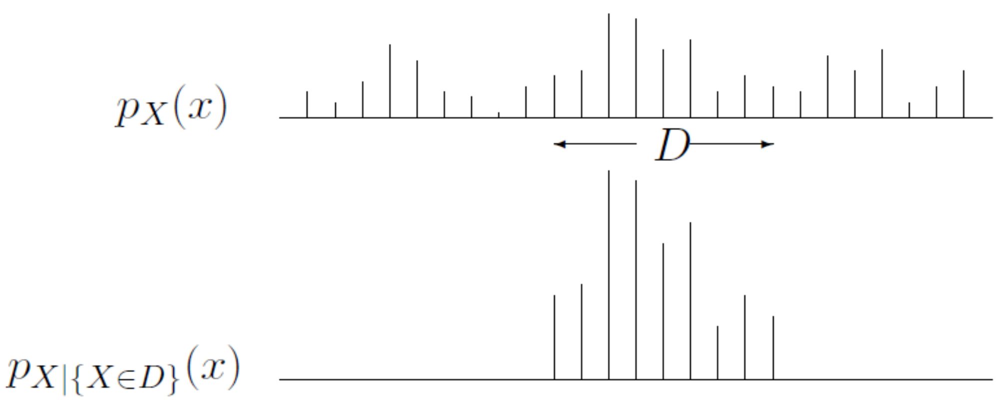
Conditional probability

- Meaning of conditional probability

When the conditioning event is of the form $\{X \in D\}$,

$$p_{X|\{X \in D\}}(x) = P(X = x \mid X \in D) = \frac{P(X = x, X \in D)}{P(X \in D)}$$

$$= \begin{cases} \dfrac{P(X = x)}{P(X \in D)}, & X \in D \\ 0, & \text{else} \end{cases}$$



Conditional probability

- Bayes Rule

$$P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{P_{Y|X}(y|x)P_X(x)}{\sum_X P_{Y|X}(y|X)P_X(X)}$$

$$X = x \in A, \qquad Y = y \in B$$

- Learning and Inference ?

Information

- Discrete random variable X is defined in the sample set Ψ

$$\Psi = \left\{ x_k \middle| k = 0, \pm 1, \ldots, \pm K \right\}$$

- Event $X = x_k$ occurs with probability $p_k = P(X = x_k)$

- **Information** ≡ surprise ≡ uncertainty
  The amount of information of the event is related to the *inverse* of the probability of occurrence. That is, the lower the probability p_k is, the more "surprise" there is, and the more "information". $I(x_k) = \log(\frac{1}{p_k}) = -\log p_k$

$$\begin{cases} \text{내일도 지구가 돈다} & p_k = 1 \text{ : 정보(x), surprise(x)} \\ \text{내일 미국이 북한을 공격한다} & p_k << 1\text{: 정보(o), surprise(o)} \end{cases}$$

Information

base=2 ⇒ 정보단위 bits
base=e ⇒ 정보단위 nats

32 bit : 한 code의 정보는 $I(x_k) = -\log(\frac{1}{2^{32}}) = 32$

① $I(x_k) = 0$ for $p_k = 1$

② $I(x_k) \geq 0$ for $0 \leq p_k \leq 1$

③ $I(x_k) \geq I(x_i)$ for $p_k \leq p_i$ (희귀한 정보)

Entropy : a measure of the *average amount of information conveyed per message*, i.e., expectation of Information

$$H(X) = E[I(X)] = \sum_{k=-K}^{K} p_k I(x_k) = -\sum_{k=-K}^{K} p_k \log p_k$$

Information

- Maximum entropy : when p_k is equiprobable.

$$0 \le H(X) \le -\sum_{k=-K}^{K} \frac{1}{2K+1} \log \frac{1}{(2K+1)} = \log(2K+1)$$

$H(X) = 0$ for an event $p_k = 1$ o/w $p_k = 0$

- Theorem (Gray 1990)

$$\sum_k p_k \log\left(\frac{p_k}{q_k}\right) \ge 0$$

- Relative entropy (or Kullback – Leibler divergence)

$$D_{p\|q} = \sum_{x \in \mathrm{X}} \underbrace{p_X(x)}_{\text{probability mass ftn.}} \log\left(\frac{p_X(x)}{q_X(x)}\right)$$

probability mass ftn. $q_X(x)$: reference pmf

Information

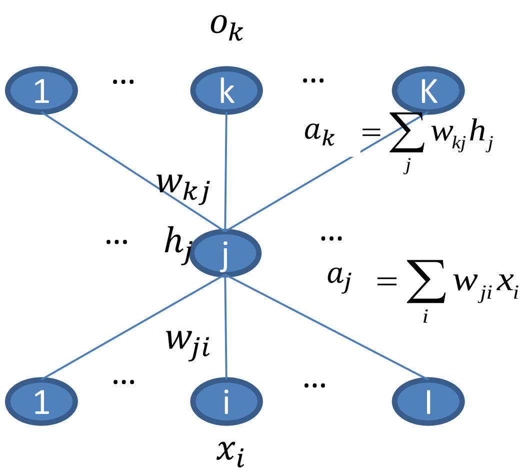
- Relative entropy (or Kullback – Leibler divergence)

$$D_{p\|q} = \sum_{x \in X} p_X(x) \log\left(\frac{p_X(x)}{q_X(x)}\right)$$

- Cross entropy

$$C_{p||q}(x;W) = -\sum_x p(x)\ \log q(x;W)$$

- Cross entropy for classification by deep learning



$$C_{p||q}(X;\ W) = -\sum_i [p(x_i)\ \log q(x_i;W) + (1-p(x_i)) \log(1-q(x_i;W))]$$

Mutual Information

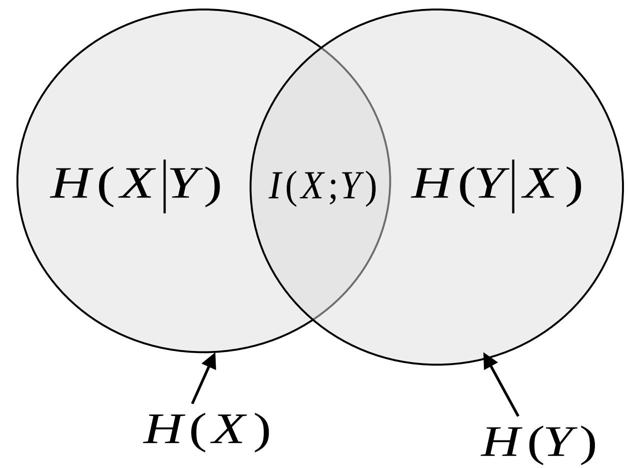
- Conditional Entropy (조건부 불확실성의 량)

  Y 가 관측되고 난 후의 X 의 정보기대치 (Entropy)

  Y 와 연관이 있는 X 의 정보는 제외

- Theorem (Gray 1990)

$$H(X|Y) = H(X,Y) - H(Y)$$
$$0 \le H(X|Y) \le H(X)$$



- Joint Entropy

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$

Joint probability mass function

Mutual Information

- Mutual Information: Output *Y* 의 관측에 의해 알 수 있는 *X* 의 uncertainty (정보)

$$I(X;Y) = H(X) - H(X|Y)$$

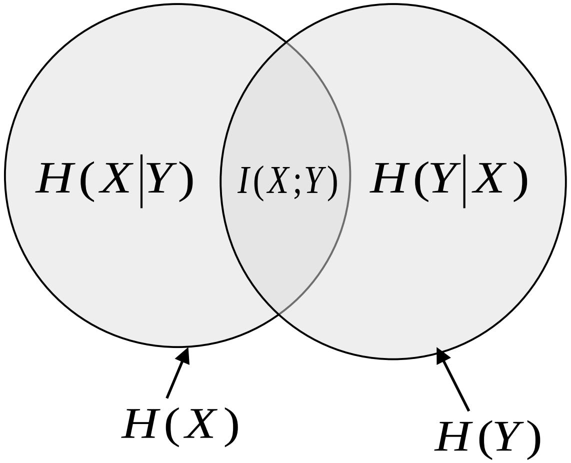
$$= H(X) + H(Y) - H(X,Y)$$

$$= -\sum_{x\in X} p(x)\log\left(p(x)\right) - \sum_{y\in Y} p(y)\log\left(p(y)\right)$$

$$+\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(p(x,y)\right)$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

QL-divergence
Independence ?



$$H(X) = I(X,X)$$

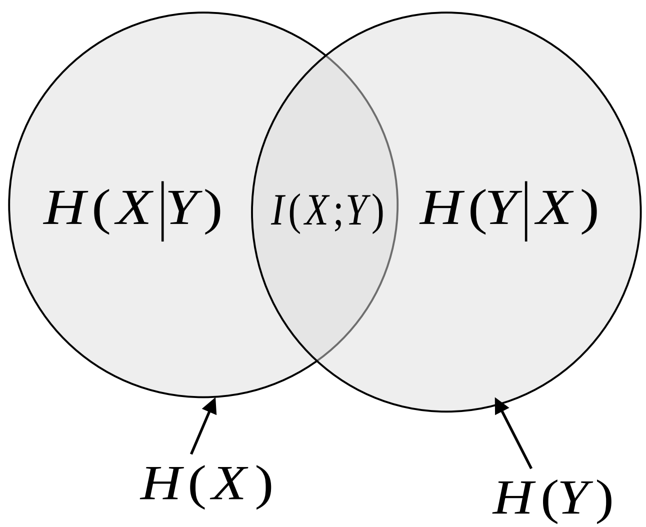
Mutual Information

- Properties of $I(X,Y)$

① $I(Y;X) = I(X;Y)$

② $I(X;Y) \geq 0$

③ $I(X;Y) = H(Y) - H(Y|X)$



Mutual Information

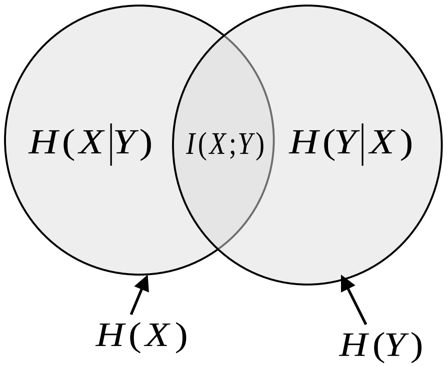
- Mutual Information for Continuous Random Variables

$$I(X;Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}\right)dxdy$$

$$\begin{aligned}I(X;Y)&=h(X)-h(X|Y)=h(Y)-h(Y|X)\\&=h(X)+h(Y)-h(X,Y)\end{aligned}$$

$$I(X;Y)=I(Y;X)$$

$$I(X;Y)\geq 0$$



Mutual Information

Exercise:

- In computer science(CS) department, the probability of dropping the machine learning(ML) course in March is 1/6, that in April is 1/3, and the probability of taking ML course to the end without dropping is 1/2, whereas those in Electrical engineering(EE) department are 1/8, 1/8, and 3/4, respectively. Meanwhile, the portions of CS & EE students in ML course are 1/5 & 4/5, respectively. Letting X be the random variable on dropping or not of a student, and Y be the random variable on the department of a student, find the followings.

1. Conditional entropy $H(X|Y)$.
2. Mutual information $I(X;Y)$.

Mutual Information

1. We get the joint distribution $p(X,Y) = p(X|Y) \times p(Y)$. Then, we compute the conditional entropy $H(X|Y)$ by $H(X|Y) = H(X,Y) - H(Y)$.

$$H(Y) = -\Sigma_{y \in Y} p(y) logp(y) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} = 0.7219$$

$$H(X,Y) = \ -\Sigma_{x \in X}\Sigma_{y \in Y} p(x,y) logp(x,y)$$

$$H(X,Y) = \ -\Sigma_{x \in X}\Sigma_{y \in Y} p(x|y) \times p(y) logp(x|y) \times p(y)$$

$$= -\frac{1}{6} * \frac{1}{5}\log\left(\frac{1}{6} * \frac{1}{5}\right) - \frac{1}{3} * \frac{1}{5}\log\left(\frac{1}{3} * \frac{1}{5}\right) - \frac{1}{2} * \frac{1}{5}\log\left(\frac{1}{2} * \frac{1}{5}\right)$$
$$-\frac{1}{8} * \frac{4}{5}\log\left(\frac{1}{8} * \frac{4}{5}\right) - \frac{1}{8} * \frac{4}{5}\log\left(\frac{1}{8} * \frac{4}{5}\right) - \frac{3}{4} * \frac{4}{5}\log\left(\frac{3}{4} * \frac{4}{5}\right)$$
$$= \ 1.8628$$

Hence, the conditional entropy $H(X|Y)$ is 1.1409

Mutual Information

2. Compute the mutual information $I(X;Y)$ using the equation $I(X;Y) = H(X) + H(Y) - \; H(X,Y)$. Since $p(X = March\ drop) = \frac{1}{6} * \frac{1}{5} + \frac{1}{8} * \frac{4}{5} = \frac{2}{15}$, $p(X = April\ \ drop) = \frac{1}{3} * \frac{1}{5} + \frac{1}{8} * \frac{4}{5} = \frac{1}{6}$, $p(NO\ drop) = \frac{1}{2} * \frac{1}{5} + \frac{3}{4} * \frac{4}{5} = \frac{7}{10}$

$$H(X) = \; -\Sigma_{x \in X} p(x) log p(x)$$

$$= -\left[\frac{2}{15}\log\left(\frac{2}{15}\right) + \frac{1}{6}\log\left(\frac{1}{6}\right) + \frac{7}{10}\log\left(\frac{7}{10}\right) + \right] = 1.1786$$

Using $H(Y)$ and $H(X,Y)$ calculated in (1), we compute $I(X;Y)$ as below:

$I(X;Y) = H(X) - \; H(X|Y) = 1.1786 - 1.1409 = 0.038$

This means X and Y are dependent to each other.

Interim Summary

- Conditional Probability
- Chain Rule
- Independence
- Total Probability
- Bayes Rule
- Information
- Entropy
- Mutual Information
- Kullback Leibler Divergence
- Cross-Entropy