Artificial Intelligence 2: Information Theory II – Measures for more variables

Shan He

School for Computer Science
University of Birmingham

Outline of Topics

1. Joint entropy and conditional entropy
    - Chain rule for conditional entropy
2. Relative entropy (aka. Kullback-Leibler divergence)
3. Mutual Information

Joint/Conditional entropy

Note 1: The Joint entropy and conditional entropy you are going to learn is based on joint PMF and conditional PMF, or most basically joint probability and conditional probability. If you have forgot what they are, please pause this video and read this text book Introduction to Probability, Statistics, and Random Processes .

Note 2: To simplify the notations, we are using the following shorthand mathematical notations for joint probability distributions of two random variables X and Y using notations of probability:

- The marginal PMF of X: $P(X)$
- The marginal probability of X takes the value x: $p(x)$.
- Joint PMF: $P(X, Y)$
- The value of joint PMF $p(X, Y)$ at (x, y), i.e., $P_{XY}(X = x, Y = y)$: $p(x, y)$
- Conditional PMF: $P(X|Y)$
- The value of conditional PMF $p(X|Y)$ at (x, y), i.e., $P_{X|Y}(X = x, Y = y)$: $p(x|y)$

Joint entropy

Joint entropy

"*A measure of the uncertainty associated with a set of variables.*". For two discrete random variables X and Y, the joint entropy is defined as:

$$H(X,Y) = -E[\log p(X,Y)] = -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log p(x_i, y_j), \qquad (1)$$

Conditional entropy

Conditional entropy

"*Quantifies uncertainty of the outcome of a random variable Y given the outcome of another random variable X.*", which is defined as

$$H(Y|X) \equiv -E[\log p(Y|X)] = -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log p(y_j|x_i) \qquad (2)$$

Conditional entropy

Denote $H(Y|X=x)$ as the entropy of the discrete random variable Y conditioned on the discrete random variable X taking a certain value x, then $H(Y|X)$ is the result of averaging $H(Y|X=x)$ over all possible values x that X may take:

$$\begin{aligned} H(Y|X) &\equiv \sum_{x_i \in R_X} p(x_i)\, H(Y|X=x_i) \\ &= -\sum_{x_i \in R_X} p(x_i) \sum_{y_j \in R_Y} p(y_j|x_i) \log\, p(y_j|x_i) \\ &= -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log\, p(y_i|x_j) \\ &= -E\left[\log p(Y|X)\right]. \end{aligned} \tag{3}$$

Chain rule for conditional entropy

From equation 2 and the conditional probability: $p(y|x) = \frac{p(x,y)}{p(x)}$, we have

$$
\begin{aligned}
H(Y|X) &= -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log p(y_j|x_i) \\
&= -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i)} \right) \\
&= -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \left[\log(p(x_i, y_j)) - \log(p(x_i)) \right] \\
&= -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log(p(x_i, y_j)) + \sum_{x_i \in R_X} \underbrace{\sum_{y_j \in R_Y} p(x_i, y_j)}_{p(x_i)} \log(p(x_i) \\
&= H(X, Y) + \sum_{x_i \in R_X} p(x_i) \log(p(x_i)) \\
&= H(X, Y) - H(X)
\end{aligned}
$$

Chain rule for conditional entropy

Chain rule for conditional entropy:

$$H(Y|X) = H(X,Y) - H(X) \tag{4}$$

$$H(X|Y) = H(X,Y) - H(Y) \tag{5}$$

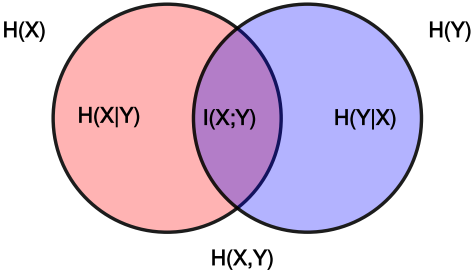

Figure: By KonradVoelkel - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11245361

Relative entropy (aka. Kullback-Leibler divergence)

- **Relative entropy** or **Kullback-Leibler divergence**: "*quantifies the distance between two probability distributions*"
- Let $P(x)$ and $Q(x)$ are two probability distributions of a discrete random variable X. That is, both $P(x)$ and $Q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0$. For any $x \in R_X$, the KL divergence from P to Q is defined as

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x \in R_X} P(x) \log \frac{P(x)}{Q(x)} = E\left[\log \frac{P(x)}{Q(x)}\right]. \tag{6}$$

- Note 1: $0 \log \frac{0}{Q} = 0$ and $P \log \frac{P}{0} = \infty$.
- A measure of distance:
  - $D_{\mathrm{KL}}(P\|Q) \geq 0$
  - $D_{\mathrm{KL}}(P\|Q) = 0$ iff $P(x) = Q(x)$.
- However, it is not a true distance: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$

K-L divergence: Example

For a binary random variable X with range $R_X = \{0, 1\}$, we assume two distributions P and Q with $P(0) = 1 - r$, $P(1) = r$ and $Q(0) = 1 - s$, $Q(1) = s$. The KL Divergences are:

$$D_{\mathrm{KL}}(P \| Q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$
$$D_{\mathrm{KL}}(Q \| P) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}$$

If $r = s$, then $D_{\mathrm{KL}}(P \| Q) = D_{\mathrm{KL}}(Q \| P) = 0$
If $r = \frac{1}{2}$ and $s = \frac{1}{4}$, then

$$D_{\mathrm{KL}}(P \| Q) = 0.2075$$
$$D_{\mathrm{KL}}(Q \| P) = 0.1887$$

Other distance measures

- Cross Entropy: For two discrete distributions P and Q, cross entropy is defined as

$$H(P,Q) = -\sum_{x\in R_X} P(x)\log Q(x) = H(P) + D_{\mathrm{KL}}(P\|Q)$$

- Jensen–Shannon divergence (JSD): a symmetrized and smoothed version of the Kullback–Leibler divergence $D(P \parallel Q)$. It is defined by

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2}D_{\mathrm{KL}}(Q \parallel M)$$

  where $M = \frac{1}{2}(P+Q)$

- Wasserstein Distance (aka. Earth Mover's distance) [1]: the minimum energy cost of moving and transforming a pile of dirt in the shape of one probability distribution to the shape of the other distribution.

[1] If you want to learn more about these two distance measures and how to use them to train Generative Adversarial Networks, please read this paper from OpenAI: From GAN to WGAN

Mutual Information

- **Mutual information**: "*measures the information that X and Y share*"
- Intuitively, it measures how much knowing one of these variables reduces uncertainty about the other
- Formally, the mutual information of two discrete random variables X and Y is defined as:

$$I(X;Y) = \sum_{x \in R_X} \sum_{y \in R_Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right).$$

MI and KL divergence

- The mutual information $I(X;Y)$ is also defined as the KL divergence between the joint distribution and the product of marginal distributions

$$\begin{aligned} I(X;Y) &= \sum_{x \in R_X} \sum_{y \in R_Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) \\ &= D_{\mathrm{KL}}(P(X,Y) \| P(X)P(Y)) \end{aligned}$$

- Interpretation: the mutual information essentially measure the distance (error) of using $P(X)P(Y)$ to model the joint probability $P(X,Y)$. When X and Y are independent of each other, i.e., $p(x,y) = p(x)p(y)$, we have

$$I(X;Y) = D_{\mathrm{KL}}(P(X,Y) \| P(X)P(Y)) = 0$$

MI and KL divergence

Because

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

we have

$$\begin{aligned} I(X;Y) &= \sum_{x \in R_X} \sum_{y \in R_Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) \\ &= \sum_{y \in R_Y} p(y) \sum_{x \in R_X} p(x|y) \log \frac{p(x|y)}{p(x)} \\ &= \sum_{y \in R_Y} p(y)\, D_{\mathrm{KL}}(P(X|Y) \| P(X)) \\ &= E_Y \{ D_{\mathrm{KL}}(P(X|Y) \| P(X)) \}. \end{aligned}$$

Similarly, we have

$$I(X;Y) = E_X \{ D_{\mathrm{KL}}(P(X|Y) \| P(Y)) \}.$$

Mutual information and Entropy

We can also express mutual information in terms of joint and conditional entropies:

$$\begin{aligned} I(X;Y) &\equiv H(X) - H(X|Y) && (7) \\ &\equiv H(Y) - H(Y|X) && (8) \\ &\equiv H(X) + H(Y) - H(Y,X) && (9) \\ &\equiv H(X,Y) - H(Y|X) - H(X|Y) && (10) \end{aligned}$$

Visualising mutual information

The above equations can be visualised by this figure:

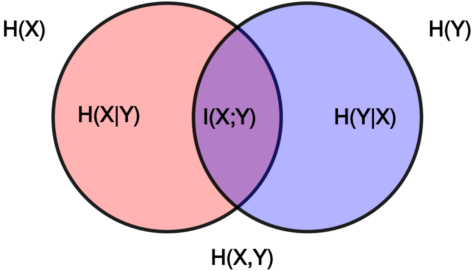

Figure: By KonradVoelkel - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11245361

Properties of Mutual Information

- Non-negativity: $I(X;Y) \geq 0$
- Symmetric
- Measure of dependence between X and Y:
  - $I(X;Y) = 0$ iff $X \perp Y$
  - $I(X;Y)$ not only increases with the dependence of X and Y but also with $H(X)$ and $H(Y)$
- $I(X;X) = H(X) - H(X|X) = 0$

Exercise question

Suppose we have a simplified language which has only three vowels: ‘a’, ‘i’ and ‘u’, and three consonants ‘p’, ‘t’ and ‘k’. Let’s denote a discrete random variable X which maps the events that consonants ‘p’, ‘t’ and ‘k’ occurring to the range $R_X = \{1, 2, 3\}$. Similarly, we define the a discrete random variable Y with range $R_Y = \{1, 2, 3\}$ for the occurring of vowels in the language. We estimated the joint PMF of a vowel and a consonant occurring together in the same syllable:

$p(x, y)$	$x = 1$	$x = 2$	$x = 3$
$y = 1$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
$y = 2$	$\frac{1}{16}$	$\frac{3}{16}$	0
$y = 3$	0	$\frac{3}{16}$	$\frac{1}{16}$

Questions: obtain the following

- Entropies $H(X)$ and $H(Y)$
- Conditional entropies $H(X|Y)$ and $H(Y|X)$
- Joint entropy $H(X, Y)$
- Mutual information $I(X; Y)$

Question and further reading

Question: Why should I lean the complex mathematics in this lecture?
Answer: to help you understand some advanced Generative AI techniques such as Stable Diffusion and GPT (at least GPT 1/2)

- Diffusion probabilistic models: What are Diffusion Models?
- GPT1: Improving Language Understanding by Generative Pre-Training
- GPT2: Language Models are Unsupervised Multitask Learners
- GPT3: Language Models are Few-Shot Learners

Solution for the Exercise question

$p(x,y)$	$x=1$	$x=2$	$x=3$	$p(y)$
$y=1$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2}$
$y=2$	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{4}$
$y=3$	0	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$p(x)$	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	1

Questions: obtain the following

- Entropies $H(X)$ and $H(Y)$

Answer:

$$H(X) = -\sum_{i}^{n} p(x_i)\log_2 p(x_i) = -(\frac{1}{8}\log\frac{1}{8}+\frac{3}{4}\log\frac{3}{4}+\frac{1}{8}\log\frac{1}{8}) = 1.061 \text{ bits}$$

$$H(Y) = -\sum_{i}^{n} p(y_i)\log_2 p(y_i) = -(\frac{1}{2}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{4}+\frac{1}{4}\log\frac{1}{4}) = 1.5 \text{ bits}$$

Solution for the Exercise question

Questions: obtain the following

- Conditional entropies $H(X|Y)$ and $H(Y|X)$
- Joint entropy $H(X,Y)$
- Mutual information $I(X;Y)$

Answer:

$$\begin{aligned}
H(X,Y) &= -\sum_{x_i \in R_X} \sum_{y_j \in R_Y} p(x_i, y_j) \log p(x_i, y_j) = 2.436 \text{ bits} \\
H(X|Y) &= H(X,Y) - H(Y) = 2.436 - 1.5 = 0.936 \text{ bits} \\
H(Y|X) &= H(X,Y) - H(X) = 2.436 - 1.061 = 1.375 \text{ bits} \\
I(X;Y) &= H(X) - H(X|Y) = 1.061 - 0.936 = 0.125 \text{ bits}
\end{aligned}$$