Comp 2610/6261

Tutorial 4 : Relative Entropy & Mutual Information

Decomposability of Entropy:

$$
\begin{aligned}
H(\mathbf{p}) = & H\left(\sum_{i=1}^{m} p_i, \sum_{i=m+1}^{|\mathcal{X}|} p_i\right) \\
& + \left(\sum_{i=1}^{m} p_i\right) H\left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right) \\
& + \left(\sum_{i=m+1}^{|\mathcal{X}|} p_i\right) H\left(\frac{p_{m+1}}{\sum_{i=m+1}^{|\mathcal{X}|} p_i}, \ldots, \frac{p_{|\mathcal{X}|}}{\sum_{i=m+1}^{|\mathcal{X}|} p_i}\right)
\end{aligned}
$$

KL Divergence:

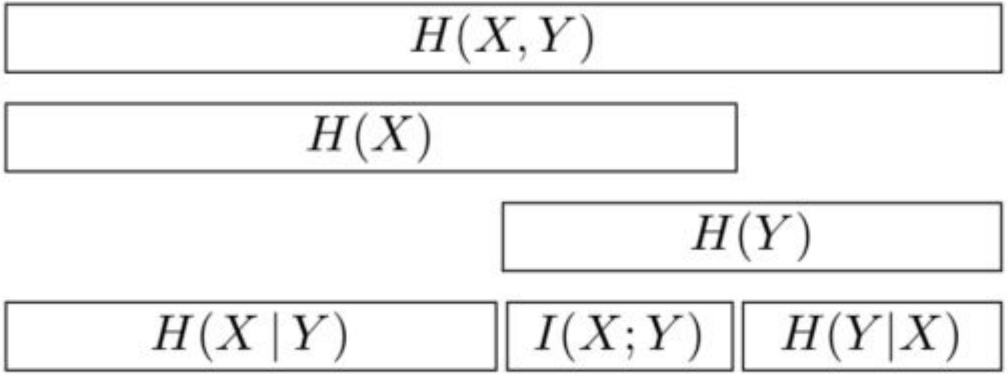
$$
\begin{aligned}
D_{\mathrm{KL}}(p \| q) &= \sum_{x \in \mathcal{X}} p(x) \left(\log \frac{1}{q(x)} - \log \frac{1}{p(x)}\right) \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right].
\end{aligned}
$$

Mutual Information:

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

$$
\begin{aligned}
I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x \in \mathcal{X}} \log p(x) \sum_{y \in \mathcal{Y}} p(x, y) - \left(-\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y)\right) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

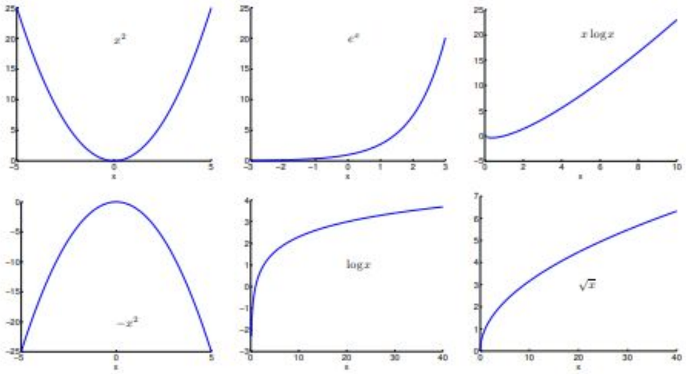
Breakdown of Joint Entropy:



Conditional Mutual Information:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \mathbb{E}_{p(X,Y,Z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

Convex and Concave functions:



Jensen’s Inequality:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Data-Processing Inequality:

$$\text{if } X \to Y \to Z \text{ then: } I(X;Y) \geq I(X;Z)$$

1. Suppose Y is a geometric random variable, $Y \sim Geom(y)$. i.e., Y has probability function,

$$P(Y = y) = p\,(1-p)^{y-1},\quad y = 1, 2, \ldots$$

Determine the mean and variance of the geometric random variable.

Solution: The expectation of the geometric random variable can be calculated as,

$$\begin{aligned} E[Y] &= \sum_{y=1}^{\infty} y \cdot P(Y=y) \\ &= \sum_{y=1}^{\infty} y \cdot p(1-p)^{y-1} \\ &= p\sum_{y=1}^{\infty} y(1-p)^{y-1} \end{aligned}$$

$$E[Y] = p[1 + 2(1-p) + 3(1-p)^2 + \ldots] \tag{1}$$

$$(1-p)\,E[Y] = [(1-p) + 2(1-p)^2 + 3(1-p)^3 + \ldots] \tag{2}$$

$$E[Y] \cdot (1-(1-p)) = p\,[1 + (1-p) + (1-p)^2 + \ldots] \qquad (1) - (2)$$

$$E[Y] \cdot p = p \cdot \frac{1}{(1-(1-p))} \qquad (*)$$

$$E[Y] = \frac{1}{p} \tag{3}$$

(*) Here we use the sum to infinity of geometric series, where $|p| < 1$,

$$\sum_{i=1}^{\infty} p^i = \frac{1}{1-p} \tag{4}$$

To calculate the variance, we need to calculate $E[Y^2]$:

$$\begin{aligned} E[Y^2] &= \sum_{y=1}^{\infty} y^2 \cdot P(Y=y) \\ &= \sum_{y=1}^{\infty} y^2 \cdot p(1-p)^{y-1} \\ &= \sum_{y=1}^{\infty} (y-1+1)^2 \cdot p(1-p)^{y-1} \\ &= \sum_{y=1}^{\infty} ((y-1)^2 + 2(y-1) + 1) \cdot p \cdot r^{y-1} && \text{let } r = 1-p \\ &= \sum_{z=0}^{\infty} z^2 p r^z + 2\sum_{z=0}^{\infty} z p r^z + \sum_{z=0}^{\infty} p r^z && \text{let } z = y-1 \\ &= r \cdot \sum_{z=0}^{\infty} z^2 p r^{z-1} + 2r \cdot \sum_{z=0}^{\infty} z p r^{z-1} + p\sum_{z=0}^{\infty} r^z \\ &= r \cdot \sum_{z=1}^{\infty} z^2 p r^{z-1} + 2r \cdot \sum_{z=1}^{\infty} z p r^{z-1} + p \cdot \frac{1}{1-(1-p)} && \text{using (4)} \end{aligned}$$

$$\begin{aligned} E[Y^2] &= r \cdot E[Y^2] + 2r \cdot E[Y] + 1 \\ E[Y^2] &= \frac{1+r}{p^2} \end{aligned} \tag{5}$$

Therefore, the Variance can be calculated as:

$$\begin{aligned} Var[Y] &= E[Y^2] - (E[Y])^2 \\ &= \frac{1+r}{p^2} - \left(\frac{1}{p}\right)^2 && \text{using (5)} \\ &= \frac{r}{p^2} \\ &= \frac{1-p}{p^2} \end{aligned} \tag{6}$$

2. The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate

a) H(X)

b) H (Y)

c) H (Y |X)

d) H (X|Y)

Solution: According to question, two teams play until one of them has won 4 games.

There are **2** (AAAA, BBBB) - World Series with **4** games. Each happens with probability **$(1/2)^4$** .

⇨ The probability of a **4** game series (Y = 4) is $2(1/2)^4$ = **1/8.**

There are **8** = 2 * $^4C_3$ - World Series with **5** games. Each happens with probability **$(1/2)^5$** .

⇨ The probability of a **5** game series (Y = 5) is $8(1/2)^5$ = **1/4.**

There are **20** = 2 * $^5C_3$ - World Series with **6** games. Each happens with probability **$(1/2)^6$** .

⇨ The probability of a **6** game series (Y = 6) is $20(1/2)^6$ = **5/16.**

There are **40** = 2 * $^6C_3$ - World Series with **7** games. Each happens with probability **$(1/2)^7$**.

⇨ The probability of a **7** game series (Y = 7) is $40(1/2)^7$ = **5/16.**

$$\boldsymbol{H(X)} = \sum p(x) log \frac{1}{p(x)} = 2\left(\frac{1}{16}\right) log16 + 8\left(\frac{1}{32}\right) log32 + 20\left(\frac{1}{64}\right) log64 + 40\left(\frac{1}{128}\right) log128 = \mathbf{5.8125}$$

$$\boldsymbol{H(Y)} = \sum p(y) log \frac{1}{p(y)} = \left(\frac{1}{8}\right) log8 + \left(\frac{1}{4}\right) log4 + \left(\frac{5}{16}\right) log\left(\frac{16}{5}\right) + \left(\frac{5}{16}\right) log\left(\frac{16}{5}\right) = \mathbf{1.924}$$

Since, Y is a deterministic function of X, so if you know X there is no randomness in Y. Or, ***H (Y |X)* = 0.**

Also, ***H(X) + H (Y |X) = H (X, Y) = H (Y) + H (X|Y)***

⇨ ***H (X|Y)*** = *H(X) + H (Y |X) - H (Y)* = ***3.889***

3. Recall that for a random variable X, its variance is $\mathrm{Var}[X] = E[X^2] - (E[X])^2$. Using Jensen's inequality, show that the variance must always be nonnegative.

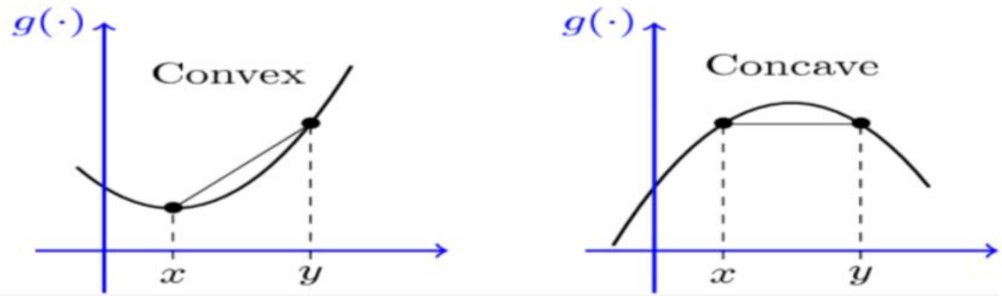
Solution: This is a direct application of Jensen's inequality to the convex function $g(x) = x^2$.

For a random variable X, variance is always positive i.e., $\mathrm{Var}[X] = E[X^2] - (E[X])^2 \geq 0$ (1)

⇨ $E[X^2] \geq (E[X])^2$ (2)

If we define a **convex function** $g(x) = x^2$, then from eqn. (2) $E[g(X)] \geq g(E[X])$ (3)

According to **Jensen's inequality**, for any convex function g, we have $E[g(X)] \geq g(E[X])$. Here **Convex function** may be defined as a function for which if we pick any two points on the graph and draw a line segment between the two points then the entire segment will always lie above the graph.



To use Jensen's inequality, we need to determine if a function 'g' is Convex. A useful method to check weather a function is Convex or not is to find its second derivative. If $g''(x) \geq 0$, then the function will be Convex otherwise Concave. For e.g., if $g(x) = x^2$ then $g''(x) = 2 \geq 0$, thus Convex function. Hence according to Jensen's inequality, for any Convex function g(x), $E[g(X)] \geq g(E[X]) \Rightarrow E[x^2] \geq (E[X])^2$

⇨ $E[x^2] - (E[X])^2$ = Variance (V(x)) will always be non-negative.

4. Let X and Y be independent random variables with possible outcomes $\{0, 1\}$, each having a Bernoulli distribution with parameter ½, i.e.

$$p(X=0) = p(X=1) = \frac{1}{2}$$
$$p(Y=0) = p(Y=1) = \frac{1}{2}.$$

a) Compute $I(X; Y)$.

b) Let $Z = X + Y$. Compute $I(X; Y | Z)$.

c) Do the above quantities contradict the data-processing inequality? Explain your answer.

/1

Solution:

(a) I (X;Y) = H (X) + H (Y) – H(X,Y) = H (X) + H (Y) – [H(X) + H(Y)] (because X and Y are independent)

=> I (X;Y) = 0

(b) To compute $I(X;Y|Z)$ we apply the definition of conditional mutual information:

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

Now, X is fully determined by Y and Z. In other words, given Y and Z there is only one state of X that is possible, i.e it has probability 1. Therefore the entropy $H(X|Y,Z) = 0$. We have that:

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

To determine this value we look at the distribution $p(X|Z)$, which is computed by considering the following possibilities:

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	2

Therefore:

$$\mathrm{p}(X|Z=0) = (1,0)$$
$$\mathrm{p}(X|Z=1) = (1/2,1/2)$$
$$\mathrm{p}(X|Z=2) = (0,1)$$

From this, we obtain: $H(X|Z=0) = 0$, $H(X|Z=2) = 0$, $H(X|Z=1) = 1$ bit. Therefore:

$$I(X;Y|Z) = p(Z=1)H(X|Z=1) = (1/2)(1) = 0.5 \text{ bits.}$$

(c) This does not contradict the data-processing inequality (or more specifically the "conditioning on a down-stream variable" corollary): the random variables in this example do not form a Markov chain. In fact, Z depends on both X and Y.