ITCT Midterm

- R07944007 網媒碩一 林良翰

1.

a. Chain Rule for Relative Entropy

- By the definition of relative entropy

$$
\begin{aligned}
D\left(p\left(x,y\right)\parallel q\left(x,y\right)\right) &= \sum_{x\in X}\sum_{y\in Y}\left[p\left(x,y\right)\log\frac{p\left(x,y\right)}{q\left(x,y\right)}\right] \\
&= \sum_{x\in X}\sum_{y\in Y}\left[p\left(x,y\right)\log p\left(x,y\right)-p\left(x,y\right)\log q\left(x,y\right)\right] \\
&= \sum_{x\in X}\sum_{y\in Y}\left[p\left(x,y\right)\log\left(p\left(x\right)p\left(y|x\right)\right)-p\left(x,y\right)\log\left(q\left(x\right)q\left(y|x\right)\right)\right] \\
&= \sum_{x\in X}\sum_{y\in Y}\left[p\left(x,y\right)\log\frac{p\left(x\right)}{q\left(x\right)}+p\left(x,y\right)\log\frac{p\left(y|x\right)}{q\left(y|x\right)}\right] \\
&= \sum_{x\in X}p\left(x\right)\log\frac{p\left(x\right)}{q\left(x\right)}+\sum_{x\in X}p\left(x\right)\sum_{y\in Y}p\left(y|x\right)\log\frac{p\left(y|x\right)}{q\left(y|x\right)} \\
&= D\left(p\left(x\right)\parallel q\left(y\right)\right)+D\left(p\left(y|x\right)\parallel q\left(y|x\right)\right)
\end{aligned}
$$

b. Jensen's Inequality

- By the definition of the concave function, $p_1+p_2=1$

  $p_1f\left(x_1\right)+p_2f\left(x_2\right)\leq f\left(p_1x_1+p_2x_2\right)$
- Suppose the theorem $E\left[f\left(X\right)\right]\leq f\left(E\left[X\right]\right)$ is true for distributions with $K-1$ mass points when f is a concave function

  Let $p_i'=\frac{p_i}{1-p_k}$

$$
\begin{aligned}
f\left(E\left[X\right]\right) &= f\left(\sum_{i=1}^{k} p_i x_i\right) \\
&= f\left(\sum_{i=1}^{k} \left(1-p_k\right) p_i' x_i\right) \\
&= f\left(\left(1-p_k\right) \sum_{i=1}^{k} p_i' x_i\right) \\
&= f\left(\left(1-p_k\right) p_k' x_k + \left(1-p_k\right) \sum_{i=1}^{k-1} p_i' x_i\right) \\
&= f\left(p_k x_k + \left(1-p_k\right) \sum_{i=1}^{k-1} p_i' x_i\right) \\
&\geq p_k f\left(x_k\right) + \left(1-p_k\right) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \\
&\geq p_k f\left(x_k\right) + \left(1-p_k\right) \sum_{i=1}^{k-1} p_i' f\left(x_i\right) \\
&= \sum_{i=1}^{k} p_i f\left(x_i\right)
\end{aligned}
$$

c. Convexity of Mutual Information

- We want to prove $I\left(X;Y\right)$ is a convex function of $p\left(Y|X\right)$ for fixed $p\left(X\right)$

  It is complicated to find approve directly from the definition of $I\left(X;Y\right)$
- We introduce an auxiliary random variable $\tilde{Y}$ with a mixing distribution

  $p\left(\tilde{y}|x\right) = \lambda p_1\left(y|x\right) + \left(1-\lambda\right) p_2\left(y|x\right)$
- To prove convexity, we need to prove

  $I\left(X;\tilde{Y}\right) \leq \lambda I_{p_1}\left(X;Y\right) + \left(1-\lambda\right) I_{p_2}\left(X;Y\right)$
- Since $I\left(X;\tilde{Y}\right) = D\left(p\left(x,\tilde{y}\right)||p\left(x\right)p\left(\tilde{y}\right)\right)$

$$
\begin{aligned}
p\left(\tilde{y}\right) &= \sum_{x\in X} p\left(x\right) p\left(\tilde{y}|x\right) \\
&= \sum_{x\in X} p\left(x\right) \lambda p_1\left(y|x\right) + p\left(x\right)\left(1-\lambda\right) p_2\left(y|x\right) \\
&\overset{p(x)fixed}{=} \lambda p_1\left(y\right) + \left(1-\lambda\right) p_2\left(y\right)
\end{aligned}
$$

$$
\begin{aligned}
p\left(x,\tilde{y}\right) &= p\left(x\right) p\left(\tilde{y}|x\right) \\
&= p\left(x\right) \lambda p_1\left(y|x\right) + p\left(x\right)\left(1-\lambda\right) p_2\left(y|x\right) \\
&= \lambda p_1\left(x,y\right) + \left(1-\lambda\right) p_2\left(x,y\right)
\end{aligned}
$$

- Finally, by the convexity of $D\left(p||q\right)$

$$
\begin{aligned}
D\left(p\left(x,\tilde{y}\right)||p\left(x\right)p\left(\tilde{y}\right)\right) &= D\left(\lambda p_1\left(x,y\right) + \left(1-\lambda\right) p_2\left(x,y\right)||p\left(x\right)\lambda p_1\left(y\right) + p\left(x\right)\left(1-\lambda\right) p_2\left(y\right)\right) \\
&\leq \lambda D\left(p_1\left(x,y\right)||p\left(x\right)p_1\left(y\right)\right) + \left(1-\lambda\right) D\left(p_2\left(x,y\right)||p\left(x\right)p_2\left(y\right)\right) \\
&= \lambda I_{p_1}\left(X;Y\right) + \left(1-\lambda\right) I_{p_2}\left(X;Y\right)
\end{aligned}
$$

$$\Rightarrow I\left(X;\tilde{Y}\right) \leq \lambda I_{p_1}\left(X;Y\right) + \left(1-\lambda\right) I_{p_2}\left(X;Y\right)$$

d. Fano's Inequality

- Define a random variable E with following properties

  $$E = \begin{cases} 0, & X = \hat{X} \\ 1, & X \neq \hat{X} \end{cases}$$

- By the chain rule for entropies

  $$\begin{aligned} H\left(E, X|\hat{X}\right) &= H\left(X|\hat{X}\right) + H\left(E|X,\hat{X}\right) \\ &= H\left(E|\hat{X}\right) + H\left(X|E,\hat{X}\right) \end{aligned}$$

  - By the data processing inequality

    $$H\left(X|\hat{X}\right) \geq H\left(X|Y\right)$$

  - E is the function of X, $\hat{X}$

    $$H\left(E|X,\hat{X}\right) = 0$$

  - By the property of conditional entropy

    $$H\left(E|\hat{X}\right) \leq H\left(E\right) = H\left(P_e\right)$$

  - If $\hat{X}$ is known and $E = 0$, which means the prediction is correct

    $$H\left(X|E=0,\hat{X}\right) = 0$$

    If $\hat{X}$ is known and $E = 1$, which means the prediction is not correct, but we can assure that $X \neq \hat{X} \Rightarrow X$ has $|X| - 1$ possible ground truth

    $$H\left(X|E=1,\hat{X}\right) = \log\left(|X| - 1\right)$$

    Therefore

    $$\begin{aligned} H\left(X|E,\hat{X}\right) &= P\left(E=0\right) H\left(X|E=0,\hat{X}\right) + P\left(E=1\right) H\left(X|E=1,\hat{X}\right) \\ &= \left(1-P_e\right) 0 + P_e \log\left(|X| - 1\right) \end{aligned}$$

- Combining all the inequations above

  $$H\left(X|Y\right) \leq H\left(P_e\right) + P_e \log\left(|X| - 1\right)$$

  Weaken the boundary

  $$H\left(X|Y\right) \leq 1 + P_e \log|X|$$

  $$P_e \geq \frac{H(X|Y) - 1}{\log|X|}$$

2.

a. Uniform distribution of finite discrete sources gives Maximal Entropy

- Let p be the probability function of random variable X
- Let U be uniform distributed random variable, $u(x) = \frac{1}{|X|}$
- By the fact that relative entropy is always larger or equals to 0

  $$\begin{aligned} 0 &\leq D(p||u) \\ &= \sum_{x \in X} p(x) \log \frac{p(x)}{u(x)} \\ &= \sum_{x \in X} p(x) \log |X| + \sum_{x \in X} p(x) \log p(x) \\ &= \log |X| - H(X) \end{aligned}$$

- $\Rightarrow H(X) \leq \log |X|$

  If $H(X)$ reaches maximum

  $$\begin{aligned} H(X) &= \log |X| \\ -\sum_{x \in X} p(x) \log p(x) &= \sum_{x \in X} p(x) \log |X| \end{aligned}$$

  $\Rightarrow p(x) = \frac{1}{|X|}$
- p must be a uniform distribution to maximize it's entropy

b. Normal distribution of infinite continuous sources gives Maximal Differential Entropy

- Let X be a random variable with a probability density function f, the differential entropy of X is defined as

  $h(X) = \int_{-\infty}^{\infty} f(x) \log f(x) dx$
- Let $g(x)$ be a Gaussian PDF with mean μ and variance σ^2 and $f(x)$ and arbitrary PDF with same variance

  $h(g) = \frac{1}{2} \log \left(2 \pi e \sigma^2\right)$
- Apply the same method $\rightarrow$ the relative entropy is always larger or equals to 0

  $$\begin{aligned} 0 &\leq D(f||g) \\ &= \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx \\ &= -h(f) - \int_{-\infty}^{\infty} f(x) \log (g(x)) dx \end{aligned}$$

  Note that

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\log(g(x))dx &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx \\ &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)dx + \log(e)\int_{-\infty}^{\infty} f(x)\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \log(e)\frac{\sigma^2}{2\sigma^2} \\ &= -\frac{1}{2}\log\left(2\pi e\sigma^2\right) \\ &= -h(g)\end{aligned}$$

- Therefore

  $h(f) \leq h(g)$

  and the equality holds when $f = g$

3.

a.

- Show that a suffix code is uniquely decodable
- Suppose we has a code X satisfying prefix condition.
- If X doesn't satisfy unique decodable, then we can find a code $A = (a_1, a_2, \ldots, a_N)$ from X such that this code can be constructed from other codes.
- We can extract a code $B = (b_1, b_2, \ldots, b_K)$, $K \leq N$ from A, then

  $a_1 = b_1$

  $a_2 = b_2$

  $\ldots$

  $a_K = b_K$

  Thus B we become the prefix of A, and this contradict with the fact that A satisfies prefix condition $\Rightarrow$ B not exists, A is uniquely decodable.
- Therefore, when a code satisfies suffix condition, it must also be uniquely decodable.

b.

- Show that the minimum average length over all codes satisfying the suffix condition is the same as the average length of the Huffman code for that random variable.
- If the average code length of the shortest code which its suffix code A is shorter than Huffman code, then we reverse this code A'. The suffix code of A' will be come prefix code of A, thus we get a set of prefix code which has average code length shorter than Huffman code. This violates the fact that Huffman code is the optimal prefix code. Therefore, the suffix code which has average code length shorter than Huffman code does not exist.

- If the average code length of the shortest code which its suffix code is longer than Huffman code, where the Huffman code is prefix code, we can reverse all the codes in Huffman code, and these will become suffix code. These suffix code has average code length shorter than the suffix code of the code with shotest average code length we found in the beginning $\Rightarrow$ contradict.
- By the two cases above, we can conclude that the suffix code of the code with shortest average length has the same average code length of Huffman code.

4.

a.

- Step 1 - always divide the largest node (largest probability)
- Step 2 - divide until having 2^l nodes

b.

- $P_A = 0.7, P_B = 0.2, P_C = 0.1$
- Example input - $ABCCBAAAC$
- The code tree

$$\begin{cases} A - 0.7 \begin{cases} AA - 0.49 \begin{cases} AAA - 0.343 \begin{cases} AAAA \\ AAAB \\ AAAC \end{cases} \\ AAB - 0.098 \\ AAC - 0.049 \end{cases} \\ AB - 0.14 \\ AC - 0.07 \end{cases} \\ B - 0.2 \\ C - 0.1 \end{cases}$$

c.

- "variable source symbol"-to-"fixed length codeword"
  - Pros - certain strings of symbols are frequently repeated, strings can be assigned code words that represent the "entire string of symbols".
  - Cons - usually efficient only for long files or messages.
- Huffman "variable codeword length"-to-"fixed source symbol"
  - Pros - asymptotically approach the source entropy for long messages, the receiver does not require prior knowledge of the coding table constructed by the transmitter, allow the receiver to construct its own decoding table "on the fly", the information required to do this is transmitted early in the coded messages.
  - Cons - initially "expand" rather than "compress" the data, less and less information need be sent to aid the receiver as time progresses.

5.

a.

- 這堂課到目前為止給了我最大的收穫是重新構建了我在機率方面的知識，尤其是 Entropy 這一部分。其實在大二學機率的時候，看到 KL-Divergence, Mutual Information,... 之類的名詞，往往都只能背公式，而無法深入去理解它背後的含義，但是上過這門課之後，我對他們的定義至少有完整的基礎概念，而且還多學到了很多不等式，其中的 Jensen's Inequality 我覺得是我印象最深刻的，因為他的公式非常直覺，而且證明起來也不會用到太難的數學基礎。

b.

- 這個學期中有介紹 JPEG 編碼技術，因此我也想到如果可以的話，希望也能聽到關於 PNG 方面的技術，因爲他還能夠記錄透明度，我個人就比較好奇如果是透明度的話，那該如何去壓縮？然後他和 JPEG 比較的話，又有哪些優缺點？