Teoria da Informação - Homework 03

Yuri Niitsuma

Review questions.

1) Answer formally the following questions:
   a) Define the Shannon information content $h(x)$ of the outcome x of a random experiment. Explain what the value $h(x)$ means.
   b) Define the entropy $H(X)$ of an ensemble X. Explain what the value $H(X)$ means.
   c) Define what is a convex function. Give at least two examples of functions that are convex, and at least two of functions that are not.
   d) State *Jensen's inequality*.
   e) What is the formula for the raw bit content of an ensemble X? What does it mean?
   f) Given an ensemble X, what is its smallest δ-sufficient subset S_δ?
   g) Given an ensemble X and a value $0 < \delta < 1$, what is the essential bit content $H_\delta(X)$? What does it mean?
   h) Shannon's source coding theorem can be stated as follows: If X is an ensemble with entropy $H(X) = H$ bits, then given any $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon$$

   Explain what it means for data compression.

Resposta

a) O índice de informação de Shannon é definido como:

$$h(x) = \log_2 \frac{1}{p(x)}$$

Representa em um número real o qual relevante é a informação para o contexto utilizado.

b) A entropia é:

$$H(X) = \sum_{x \in A_X} p(x) \log_2 \frac{1}{p(x)}$$
$$= - \sum_{x \in A_X} p(x) \log_2 p(x)$$

tem como objetivo em medir a diversidade em um conjunto de dados.

c) Uma função é convexa em um intervalo (a, b) quando $\forall x_1, x_2 \in (a, b)$ e $0 \leq \lambda \leq 1$, temos:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

Exemplo de duas funções que são convexas no domínio dos reais:

- $f(x) = x^2$
- $f(x) = e^x$

Exemplo de duas funções que não são convexas no domínio dos reais:

- $f(x) = x^3$
- $f(x) = \ln x$

d) Se f é uma função convexa e x é uma variável aleatória, então a desigualdade de Jensen é:

$$E\left[f(x)\right] \geq f\left(E[x]\right)$$

Sendo $E[X]$ é a esperança.

e) *raw bit content* de um conjunto é:

$$H_0(X) = \log_2 |A_X|$$

Representa a quantidade de bits necessário para identificar cada elemento do conjunto X.

f) *Principle of lossy data compression* é uma medida da máxima compressão com probabilidade de erro tolerado no máximp δ utilizando um subconjunto de palavras-chave de tamanho $\log_2 |S_\delta|$
TODO, não entendi direito o que a questão qur.

g) Se é aceito uma probabilidade de erro de δm $H_\delta(X)$ indica quantos bits será necessário pra codificar cada símbolo.

h) O teorema mostr que se você está codificando N símbolos do conjunto X e aceita um erro δ, a compressão irá usar aproximandamente $H(X) = H$ bits por símbolo se o N for muito grande. (Definição de limite do cálculo)

Problems (Chapter 2).

2) (Lower bound for Shannon entropy) [Easy] Show that for every ensemble $X = (x, A_X, P_X)$, it is the case that $H(X) \geq 0$.

Resposta

$\forall x \in A_X \Rightarrow 0 \leq p(x) \leq 1$. Então,

$$0 \leq p(x) \leq 1 \Rightarrow \frac{1}{p(x)} \geq 1$$
$$\log_2 \frac{1}{p(x)} \geq \log_2 1 = 0$$

O produto de dois valores não negativos é também não negativo.

$$p(x) \log_2 \frac{1}{p(x)} \geq 0$$

Como a soma $H(X)$ são de valores não negativos, então:

$$
\begin{aligned}
&p(x_1)\log_2\tfrac{1}{p(x_1)} \geq 0\\
&p(x_1)\log_2\tfrac{1}{p(x_1)} \geq 0\\
&\ldots\\
&p(x_n)\log_2\tfrac{1}{p(x_n)} \geq 0\\
&\underbrace{p(x_1)\log_2\frac{1}{p(x_1)} + \ldots + p(x_n)\log_2\frac{1}{p(x_n)}}_{H(X)} \geq 0
\end{aligned}
$$

3) (Upper bound for Shannon entropy) The following exercises are designed so you can prove an upper bound for Shannon entropy.
   a) (MacKay 2.21) [Easy] Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. Let $f(a) = 10$, $f(b) = 5$, and $f(c) = 10/7$. What is $\mathcal{E}[f(x)]$? What is $\mathcal{E}[1/P(x)]$?
   b) (MacKay 2.22) [Easy] For an arbitrary ensemble, what is $\mathcal{E}[1/f(x)]$?
   c) (MacKay 2.25) [Hard] Prove the assertion that $H(X) \leq \log(|A_X|)$ with equality iff $p_i = 1/|A_X|$ for all i. ($|A_X|$ denotes the number of elements in the set A_X.) [Hint: use Jensen's inequality (2.48); if your first attempt to use Jensen does not succeed, remember that Jensen involves both a random variable and a function, and you have quite a lot of freedom in choosing these; think about whether your chosen function f should be convex or concave.]

Resposta

a)

$$
\begin{aligned}
\mathcal{E}[f(x)] &= \sum_{k\in\{a,b,c\}} p_k f(k)\\
&= p_a f(a) + p_b f(b) + p_c f(c)\\
&= 0.1\cdot 10 + 0.2\cdot 5 + 0.7\cdot\frac{10}{7} = 3
\end{aligned}
$$

$$
\mathcal{E}\left[\frac{1}{P(x)}\right] = \mathcal{E}\left[f(x)\right] = 3
$$

b)

$$
\mathcal{E}\left[\frac{1}{P(x)}\right] = \sum_{x\in A_X}\frac{P(x)}{P(x)} = \sum_{x\in A_X} 1 = |A_X|
$$

c) TODO

4) (Thomas&Cover 2.1) (The entropy of a countably infinite probability distribution) [Medium] A fair coin is flipped until the first head occurs. Let X denote the number of flips required. Find the entropy $H(X)$ in bits. (The following expressions may be useful: $\sum_{n=0}^{\infty} r^n = 1/(1-r)$, and $\sum_{n=0}^{\infty} nr^n = 1/(1-r)^2$.

Resposta

A probilidade de atingir a primeira cara da moeda jogada em n tentativas é $P(X = n) = q^{n-1}p$, sendo que a moeda é justa. Assim, temos:

$$
p = q = \frac{1}{2} \Rightarrow P(X = n) = \frac{1}{2}\cdot\frac{1}{2^{n-1}} = \frac{1}{2^n}
$$

Calculando a entopia:

$$
\begin{aligned}
H(X) &= \sum_{n=1}^{\infty}\frac{1}{2^n}\cdot\log_2\frac{1}{1/2^n}\\
&= \sum_{n=1}^{\infty}\frac{1}{2^n}\cdot\underbrace{\log_2 2^n}_{=n} = \sum_{n=1}^{\infty}\frac{n}{2^n}\\
&= \sum_{n=1}^{\infty}\frac{n}{2^n} + 0 = \sum_{n=0}^{\infty} n\left(\frac{1}{2}\right)^n\\
&= \frac{1/2}{(1-1/2)^2} = \frac{1/2}{1/2^2} = \frac{1}{1/2} = 2 \text{ bits}
\end{aligned}
$$

Problems (Chapter 4).

5) (MacKay 4.2) [Easy] Show that, if x and y are independent, the entropy of the outcome x, y satisfies

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

In words, entropy is additive for independent variables.

Resposta

Dado que $p(x,y) = p(x)p(y)$.

$$
\begin{aligned}
H(X,Y) &= \sum_x\sum_y p(x,y)\log\tfrac{1}{p(x,y)}\\
&= \sum_x\sum_y p(x)p(y)\log\tfrac{1}{p(x)p(y)}\\
&= \sum_x\sum_y p(x)p(y)\log\left[p(x)p(y)\right]^{-1}\\
&= -\sum_x\sum_y p(x)p(y)\log\left[p(x)p(y)\right]\\
&= -\sum_x\sum_y p(x)p(y)\left[\log p(x) + \log p(y)\right]\\
&= -\sum_x\sum_y \left[p(x)p(y)\log p(x) + p(x)p(y)\log p(y)\right]
\end{aligned}
$$

Como a soma é finita podemos separar em duas somas.

$$
\begin{aligned}
&-\sum_x\sum_y p(x)p(y)\log p(x) - \sum_x\sum_y p(x)p(y)\log p(y)\\
&= -\sum_x p(x)\log p(x)\underbrace{\sum_y p(y)}_{=1} - \sum_y p(y)\log p(y)\underbrace{\sum_x p(x)}_{=1}\\
&= \sum_x p(x)\log p(x)^{-1} + \sum_y p(y)\log p(y)^{-1}\\
&= \underbrace{\sum_x p(x)\log\frac{1}{p(x)}}_{H(X)} + \underbrace{\sum_y p(y)\log\frac{1}{p(y)}}_{H(Y)}\\
&= H(X) + H(Y)
\end{aligned}
$$

6) (MacKay 4.5) [Medium] Could there be a compressor that maps an outcome x to a binary code $c(x)$, and a decompressor that maps c back to x, such that *every possible outcome* is compressed into a binary code of length *shorter* than $H_0(X)$ bits?

Resposta

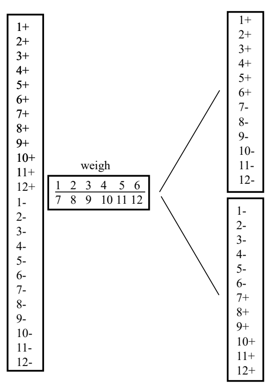
Não. Pelo princípio da casa dos pombos haverá dois números representados pela mesma string de bits se for tentar alocar com tamanho $H_0(X)$ bits.
Em outras palavras, como $c(x)$ é bijetiva e o conjunto é finito. Então o domínio e contradomínio possuem a

mesma cardinalidade. Se o domínio ou contradomínio tiver cardinalidades diferentes a função não é bijetiva.

7) (MacKay 4.9) [Easy] While some people, when they first encounter the weighing problem with 12 balls and the three-outcome balance (exercise 4.1 (p.66)), think that weighing six balls against six balls is a good first weighing, others say 'no, weighing six against six conveys no information at all'. Explain to the second group why they are both right and wrong. Compute the information gained about *which is the odd ball*, and the information gained about *which is the odd ball and whether it is heavy or light*.

Resposta

Porque intuitivamente as pessoas pensam que pode eliminar metade do conjunto das bolas. Mas o detalhe de uma bola é mais pesada ou mais leve que as outras da origem a uma ambiguidade que não é possível determinar se o conjunto mais pesado é o que tem a bola mais pesada ou o que subiu contém a mais leve.



A informação ganha da bola ser mais pesada ou leve é:

$$h(x) = \log_2 \frac{1}{p(x)} = \log_2 \frac{1}{1/24} \approx 4.58$$

A informação ganha de qual é a bola fora do padrão, os dois ao mesmo tempo, é:

$$h(x, y) = h(x) + h(y) \approx 9.16$$