7600054 — Sistemas Complexos

Gonzalo Travieso

2020-06-15

Outline

Entropia

- Dada uma partição de Ω, $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ e usando $p_i = P(A_i)$, definimos:

$$H(\mathcal{A}) = -\sum_{i=1}^{m} p_i \log_2 p_i.$$

- Esta função é denominada a entropia da partição $\mathcal{A}$.
- A base do logaritmo usada na definição pode ser qualquer, resultando apenas em um fator multiplicativo de diferença, que indica a unidade usada. No caso da base 2, a unidade da entropia é bits.
- Por essa definição, quando um evento é de probabilidade 0, precisamos avaliar $\log_2 p$ que envolve uma divergência, mas como ela aparece multiplicada por p, o resultado é que $p \log_2 p = 0$ (verifique por limite), e esse evento não contribui para a entropia.

Exemplos

- Considere uma moeda justa e a partição $\mathcal{A} = \{\{\text{CARA}\}, \{\text{COROA}\}\}$. Neste caso $p_{\text{CARA}} = p_{\text{COROA}} = \frac{1}{2}$ e portanto

$$H(\mathcal{A}) = -2\tfrac{1}{2}\log_2 \tfrac{1}{2} = 1,$$

  isto é, esta partição tem 1 bit de entropia.

- Considere um dado justo e a partição $\mathcal{A} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \}$. Neste caso, $p_i = \frac{1}{6}$ e

$$H(\mathcal{A}) = -\sum_{i=1}^{6} \tfrac{1}{6}\log_2 \tfrac{1}{6} = -6\tfrac{1}{6}\log_2 \tfrac{1}{6} = \log_2 6 \approx 2.58,$$

  isto é, esta partição tem pouco mais de 2 bits de entropia.

Exemplos (cont)

- Considere novamente um dado justo, mas com a partição $\mathcal{A} = \{\{1,3,5\},\{2,4,6\}\}$. Neste caso, $p_1 = p_2 = \frac{1}{2}$ e portanto

$$H(\mathcal{A}) = 1.$$

- Considere ainda um dado justo agora com a partição $\mathcal{A} = \{\{1\},\{2,3\},\{4,5,6\}\}$. Neste caso, $p_1 = \frac{1}{6}$, $p_2 = \frac{1}{3}$ e $p_3 = \frac{1}{2}$, resultando em

$$\begin{aligned} H(\mathcal{A}) &= -\tfrac{1}{6}\log_2\tfrac{1}{6} - \tfrac{1}{3}\log_2\tfrac{1}{3} - \tfrac{1}{2}\log_2\tfrac{1}{2} \\ &= \tfrac{1}{6}\log_2 6 + \tfrac{1}{3}\log_2 3 + \tfrac{1}{2} \\ &\approx 1.46. \end{aligned}$$

Máximo

- Como $0 \leq p_i \leq 1$, então $p_i \log_2 p_i < 0$ e portanto a entropia é sempre não-negativa $H(\mathcal{A}) \geq 0$.
- Se temos um evento A, ele conjuntamente com seu complemento A_c formam uma partição. Se p é a probabilidade de A, essa partição terá entropia

$$-p \log_2 p - (1-p) \log_2 (1-p).$$

- Essa é uma função de p que vale zero tanto em $p = 0$ quanto em $p = 1$ e é positiva para os outros valores de p. Isto significa que ela tem um máximo para algum valor de p.

Máximo (cont)

- Fazendo
$$\frac{d}{dp}\left[-p\log_2 p-(1-p)\log_2(1-p)\right]=0$$
encontramos
$$p_{\max}=\tfrac{1}{2},$$
isto é, o máximo da entropia acontece quando a probabilidade do evento e de seu complementos são iguais.
- Na verdade, isto pode ser generalizado: para uma partição $\mathcal{A}$ com m evento, a entropia tem valor máximo se todas as probabilidades p_i dos eventos na partição são idênticas,
$$p_i=\tfrac{1}{m}.$$
- Neste caso, a entropia vale
$$H(\mathcal{A})=-\sum_{i=1}^{m}\tfrac{1}{m}\log_2\tfrac{1}{m}=\log_2 m.$$

Incerteza

- Como vimos, quanto temos m eventos, a entropia é maximizada quando todos são equiprováveis, isto é, quanto temos o mínimo de conhecimento sobre qual dos m eventos pode ocorrer (se um deles fosse muito mais provável que os outros, teríamos boa segurança de que esse evento ocorreria).
- Por outro lado, no caso de m eventos equiprováveis, a entropia é dada por $\log_2 m$, isto é, ela cresce com o número de eventos, o que também aumenta a nossa incerteza sobre o evento a ser escolhido.
- Apesar de *incerteza* não ser precisamente definida aqui, isto é sempre válido: um aumento de entropia indica uma maior incerteza sobre o processo aleatório associado.

Variáveis contínuas

- Numa variável contínua, a probabilidade de eventos individuais é zero, então precisamos generalizar a entropia com base na densidade de probabilidade.
- Dada uma variável aleatória X com densidade de probabilidade $\rho(x)$, definimos a sua entropia como:

$$H(X) = -\int_{-\infty}^{\infty} \rho(x) \log_2 \rho(x)\, dx.$$

Exemplos

- Para uma variável exponencial:

$$\rho(x) = \alpha e^{-\alpha x}, \quad x \geq 0,$$

temos

$$\log_2 \rho(x) = \log_2 \alpha - \alpha \log_2(e) x$$

e portanto

$$\begin{aligned} H(X) &= -\int_0^\infty \alpha e^{-\alpha x} [\log_2 \alpha - \alpha \log_2(e) x]\, dx \\ &= -\alpha \log_2 \alpha \int_0^\infty e^{-\alpha x}\, dx + \alpha^2 \log_2 e \int_0^\infty x e^{-\alpha x}\, dx \\ &= -\log_2 \alpha + \log_2 e \\ &= \log_2 \frac{e}{\alpha}. \end{aligned}$$

Exemplos

- Para uma distribuição gaussiana:

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

temos

$$\log_2 \rho(x) = -\log_2 \sigma\sqrt{2\pi} - \frac{(x-\mu)^2}{2\sigma^2} \log_2 e.$$

Portanto:

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} \rho(x) \left[-\log_2 \sigma\sqrt{2\pi} - \frac{(x-\mu)^2}{2\sigma^2} \log_2 e\right] dx \\ &= \log_2 \sigma\sqrt{2\pi} \int_{-\infty}^{\infty} \rho(x)\, dx + log_2 e \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 \rho(x)\, dx. \end{aligned}$$

Entropia da distribuição gaussiana (cont)

- Na expressão anterior, a primeira integral é apenas a integral da densidade de probabilidade por todos os reais, o que vale 1.
- A segunda integral é a média de $(x-\mu)^2$ que como μ é a média, corresponde à variância σ^2.
- Portanto ficamos com:

$$H(X) = \log_2 \sigma\sqrt{2\pi} + \frac{1}{2}\log_2 e = \log_2 \sigma\sqrt{2\pi e}.$$

Entropia da distribuição uniforme

Considere uma distribuição uniforme entre 0 e c:

$$\rho(x) = \frac{1}{c}, \quad 0 < x < c.$$

Exercício

1. Encontre a entropia associada com essa densidade de probabilidade.
2. Dadas uma distribuição uniforme como a acima e uma distribuição gaussiana com a mesma variância, qual das duas tem entropia maior?

Entropia conjunta

- No caso de duas variáveis aleatórias, podemos definir a sua entropia conjunta.
- No caso de variáveis discretas com probabilidades conjuntas p_{ij}:

$$H(X,Y) = -\sum_{i,j} p_{ij} \log_2 p_{ij}.$$

- No caso de variáveis contínuas com densidade de probabilidade conjunta $\rho(x,y)$:

$$H(X,Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho(x,y) \log_2 \rho(x,y)\, dx\, dy.$$

- Quando as variáveis X e Y são independentes, $\rho(x,y) = \rho(x)\rho(y)$ e portanto temos:

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

isto é, as entropias se somam se as variáveis forem independentes.

Entropia condicional

- Considere o cálculo da entropia de X dado um certo valor de $Y = y$. Para isso, as probabilidade de X são as probabilidades condicionais a y:

$$H(X \mid y) = -\sum_x P(x \mid y) \log_2 P(x \mid y) = -\sum_x \frac{P(x,y)}{P(y)} \log_2 \frac{P(x,y)}{P(y)}.$$

- Se agora fazemos a média dessas entropias considerando todos os possíveis valores de y, temos a denominada entropia condicional de X dado Y:

$$H(X \mid Y) = -\sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(y)}.$$

- Esta entropia indica o quanto de incerteza resta em X se conhecemos o valor de Y:
  - Se X e Y são independentes, então $H(X \mid Y) = H(X)$.
  - Se X é totalmente determinado por Y, então $H(X \mid Y) = 0$.

Informação mútua

- Definimos a informação mútua entre as variáveis X e Y como

$$\begin{aligned} I(X,Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X \mid Y) - H(Y \mid X). \end{aligned}$$

- Esta medida quantifica a dependência mútua entre as variáveis X e Y:
  - Se X e Y são independentes, então $H(X) = H(X \mid Y)$ e $I(X,Y) = 0$.
  - Se X é totalmente determinada por Y, então $H(X \mid Y) = 0$ e $I(X,Y) = H(X) = H(Y)$, isto é, a informação mútua é idêntica à entropia dessas variáveis.

Exercício

Mostre que todas as expressões apresentadas para $I(X,Y)$ são equivalentes.

Informação mútua para análise de dependência

- A informação mútua é similar aos coeficientes de correlação, no sentido de medir relações entre variávies aleatórias.
- Ela tem a vantagem de não precisar dos valores das variáveis. O coeficiente de Pearson usa diretamente os valores no cálculo da covariância e dos desvios-padrão e o coeficiente de Spearman usa os valores durante a ordenação. Isso significa que eles não podem ser utilizados em casos onde não são associados valores aos eventos. Já a informação mútua pode ser usada, pois ela apenas se interessa nas probabilidades dos eventos.
- Ela tem a também vantagem de não depender de linearidade, mas tem a desvantagem de que, ao contrário dos coeficientes de correlação, que têm valores entre -1 e 1 e portanto fácil interpretação, os valores da informação mútua são maiores que zero, mas de resto arbitrários.

Informação mútua normalizada

- Para resolver esse problema, vários métodos de normalização do valor foram propostos.
- Um deles é a chamada incerteza simétrica

$$U(X,Y) = 2\frac{I(X,Y)}{H(X)+H(Y)},$$

que tem valores entre 0 e 1, e vale 0 quando as variáveis são independentes e vale 1 se as duas são totalmente relacionadas.