Teoria da Informação (#02)

Entropia, Entropia Conjunta, Entropia Condicional, Informação Mútua, Divergência de Kullback-Leibler

Miguel Barão

ESCOLA DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA

Surpresa de uma mensagem

$\boxed{\text{Fonte}} \xrightarrow{X}$

- Alfabeto $\mathcal{X}$
- Símbolos com distribuição $X \sim p(x)$

Se num determinado momento t a fonte gera o símbolo x_t, então a surpresa (ou *self-information*) dessa observação é

$$\log_2 \frac{1}{p(x_t)}.$$

$\boxed{\text{Fonte}} \xrightarrow{X}$

- Alfabeto $\mathcal{X}$
- Símbolos com distribuição $X \sim p(x)$

Se num determinado momento t a fonte gera o símbolo x_t, então a surpresa (ou *self-information*) dessa observação é

$$\log_2 \frac{1}{p(x_t)}.$$

Exemplo

Sabendo que probabilidade de chover num certo dia é 0.3, qual surpresa da mensagem *"Está a chover"*?

Resposta: Sendo X uma variável aleatória com alfabeto

$$\mathcal{X} = \{\text{“}\textit{Está a chover}\text{”}, \text{“}\textit{Não está a chover}\text{”}\}$$

e sabendo que os símbolos têm probabilidades 0.3 e 0.7, respectivamente, então a surpresa da mensagem *"Está a chover"* é de

$$\log_2 \frac{1}{0.3} \approx 1.737 \text{ bits.}$$

Porque motivo é usada a fórmula $\log_2 \frac{1}{p(x)}$?

A fórmula deve satisfazer os requisitos seguintes:

1. A surpresa deve ser grande quando a probabilidade é baixa.
   Deve portanto variar inversamente à probabilidade.
2. A surpresa em observações de variáveis aleatórias independentes deve ser a soma das surpresas obtidas em cada uma separadamente. Diz-se que a surpresa deve ser aditiva.
3. A surpresa deve ser uma função contínua das probabilidades.

A fórmula $\log_2 \frac{1}{p(x)}$ satisfaz os axiomas anteriores!

Exemplo

Dois indivíduos A e B fazem lançamentos de moedas obtendo resultados X_A e X_B, respectivamente. O indivíduo A tem uma moeda perfeita, enquanto o indivíduo B tem uma moeda defeituosa em que $\Pr\{X_B = \text{“cara”}\} = 0.6$. Saiu “cara” em ambas as moedas. Qual a surpresa neste resultado?

Resposta:

1. Para a moeda A temos

$$\log_2 \frac{1}{\Pr\{X_A = \text{“cara”}\}} = \log_2 \frac{1}{0.5} = 1 \text{ bit.}$$

2. Para a moeda B temos

$$\log_2 \frac{1}{\Pr\{X_B = \text{“cara”}\}} = \log_2 \frac{1}{0.6} \approx 0.737 \text{ bit.}$$

3. A surpresa total é

$$\log_2 \frac{1}{\Pr\{X_A = \text{“cara”}, X_B = \text{“cara”}\}} = \log_2 \frac{1}{0.5 \cdot 0.6} \approx 1.737 \text{ bit}$$

i.e., a soma dos resultados anteriores pois são lançamentos independentes.

A surpresa calcula-se após uma observação ser feita e diz respeito a essa observação particular.

> Como se pode lidar com a surpresa gerada por uma fonte de informação em geral?

A surpresa calcula-se após uma observação ser feita e diz respeito a essa observação particular.

Como se pode lidar com a surpresa gerada por uma fonte de informação em geral?

Ideia: Usa-se a surpresa esperada.

Exemplo

$$\text{Fonte binária} \longrightarrow \ldots, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, \ldots$$

1. $p(0) = 0.3$, a informação contida em cada 0 é: $\log_2 \frac{1}{0.3} \approx 1.737$ bits.
2. $p(1) = 0.7$, a informação contida em cada 1 é: $\log_2 \frac{1}{0.7} \approx 0.515$ bits.
3. Em média obtêm-se

$$\approx 0.3 \times 1.737 + 0.7 \times 0.515 \text{ bits}$$

por cada símbolo gerado.

Definição (Entropia)

A entropia é o valor esperado da surpresa gerada por uma fonte:

$$H(X) = E\left[\log_2 \frac{1}{p(x)}\right] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x).$$

Definição (Entropia)

A entropia é o valor esperado da surpresa gerada por uma fonte:

$$H(X) = E\left[\log_2 \frac{1}{p(x)}\right] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x).$$

- Podem ser usadas bases diferentes de 2 no logaritmo, embora a base binária seja a mais comum.
- A unidade da entropia é o
  - bit quando é usado o logaritmo na base 2 (binary unit).
  - nat quando o logaritmo está na base natural $e \approx 2.7183$ (natural unit).
  - Hartley quando o logaritmo está na base 10 (base decimal).

Considere uma fonte binária com alfabeto $\mathcal{X} = \{0, 1\}$. A entropia $H(X)$ desta fonte depende das probabilidades dos símbolos.

1. Se $p(0) = p(1) = 0.5$, então

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -(0.5 \log 0.5 + 0.5 \log 0.5) = 1 \text{ bit}$$

Considere uma fonte binária com alfabeto $\mathcal{X} = \{0, 1\}$. A entropia $H(X)$ desta fonte depende das probabilidades dos símbolos.

1. Se $p(0) = p(1) = 0.5$, então

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -(0.5 \log 0.5 + 0.5 \log 0.5) = 1 \text{ bit}$$

2. Se $p(0) = 0.1$ e $p(1) = 0.9$, então

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -(0.9 \log 0.9 + 0.1 \log 0.1) \approx 0.469 \text{ bit}$$

Entropia de uma fonte binária (processo de Bernoulli)

Considere uma fonte binária com alfabeto $\mathcal{X} = \{0, 1\}$. A entropia $H(X)$ desta fonte depende das probabilidades dos símbolos.

1. Se $p(0) = p(1) = 0.5$, então

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -(0.5 \log 0.5 + 0.5 \log 0.5) = 1 \text{ bit}$$

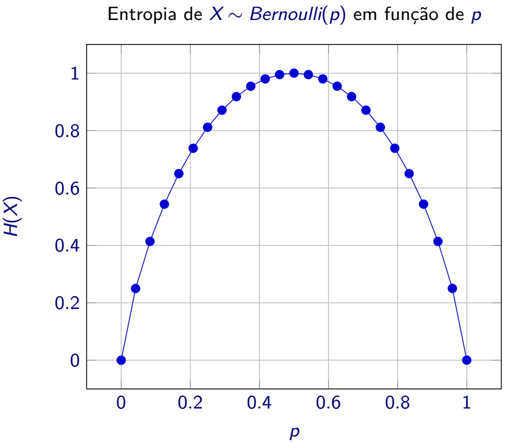
2. Se $p(0) = 0.1$ e $p(1) = 0.9$, então

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -(0.9 \log 0.9 + 0.1 \log 0.1) \approx 0.469 \text{ bit}$$

3. Se $p(0) = 0$ e $p(1) = 1$, então

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -(0 \log 0 + 1 \log 1) = 0 \text{ bit}$$

(Nota: Foi usada a convenção $0 \log 0 = 0$. A justificação para esta convenção é de que $\lim_{p \to 0} p \log p = 0$.)



A entropia é máxima quando $p = 0.5$ e aproxima-se de zero quando um dos símbolos é altamente provável. Assim, a entropia fornece uma medida de incerteza sobre uma variável aleatória.

A entropia goza das seguintes propriedades:

1. é não negativa:

$$H(X) \geq 0.$$

2. em alfabetos finitos, é majorada superiormente:

$$H(X) \leq \log_2 |\mathcal{X}|,$$

   onde $|\mathcal{X}|$ representa a cardinalidade do alfabeto $\mathcal{X}$.

3. em alfabetos finitos, a entropia é máxima quando X tem distribuição uniforme:

$$p(x) = 1/|\mathcal{X}|.$$

Exemplo

Considere uma fonte com alfabeto infinito $\mathcal{X} = \{1, 2, 3, \ldots\}$. Para gerar um destes símbolos é efectuada uma sequência de lançamentos de uma moeda até que se obtenha uma face diferente da obtida no lançamento anterior, contando-se o número total de faces iguais consecutivas. Por exemplo, (cara,cara,cara,coroa) gera o símbolo $x = 3$.
Com este processo obtém-se uma distribuição de probabilidade $p(x) = 2^{-x}$, uma vez que

$$\begin{aligned}
\Pr\{X = 1\} &= \Pr\{\text{“cara,coroa”}\} + \Pr\{\text{“coroa,cara”}\} \\
&= 0.5 \times 0.5 + 0.5 \times 0.5 = 0.5 \\
\Pr\{X = 2\} &= \Pr\{\text{“cara,cara,coroa”}\} + \Pr\{\text{“coroa,coroa,cara”}\} \\
&= 0.5 \times 0.5 \times 0.5 + 0.5 \times 0.5 \times 0.5 = 0.25 \\
&\vdots
\end{aligned}$$

A entropia da fonte é

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = -\sum_{x=1}^{+\infty} 2^{-x}\log_2 2^{-x} = \sum_{x=1}^{+\infty} x2^{-x} = 2 \text{ bits.}$$

A entropia generaliza-se facilmente para um par de variáveis aleatórias (X, Y) usando a distribuição conjunta $p(x, y)$.

Definição (Entropia conjunta)

$$H(X, Y) = E\big[-\log_2 p(x, y)\big] = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x, y)\log_2 p(x, y)$$

Na realidade a entropia conjunta é simplesmente a entropia do par (x, y) considerado como símbolo do alfabeto composto pelo produto cartesiano[1] $\mathcal{X} \times \mathcal{Y}$.

Do mesmo modo, a entropia conjunta de N variáveis aleatórias $\{X_1, X_2, \dots, X_N\}$ é

$$H(X_1, X_2, \dots, X_N) = -\sum_{x_1,\dots,x_N} p(x_1, x_2, \dots, x_N)\log_2 p(x_1, x_2, \dots, x_N).$$

[1] O produto cartesiano de dois conjuntos $\mathcal{X}$ e $\mathcal{Y}$ é o conjunto formado por todos os pares (x, y) de elementos $x \in \mathcal{X}$ e $y \in \mathcal{Y}$.

Exemplo

Duas variáveis aleatórias têm a distribuição conjunta $p(x, y)$ indicada na matriz seguinte:

$$\begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}$$

onde as variáveis x e y indexam as linhas e colunas respectivamente.
A entropia conjunta é

$$\begin{aligned} H(X, Y) &= -0.1 \log_2 0.1 - 0.2 \log_2 0.2 - 0.3 \log_2 0.3 - 0.4 \log_2 0.4 \\ &\approx 1.846 \text{ bits.} \end{aligned}$$

As variáveis X e Y podem ter alfabetos $\mathcal{X}$ e $\mathcal{Y}$ diferentes.

Problema

Considere duas variáveis aleatórias com as probabilidades conjuntas indicadas na matriz

$$\begin{bmatrix} 0.1 & 0.05 & 0.05 \\ 0.3 & 0.25 & 0.25 \end{bmatrix}$$

onde os alfabetos são $\mathcal{X} = \{1, 2\}$ *e* $\mathcal{Y} = \{1, 2, 3\}$. *Calcule a entropia conjunta* $H(X, Y)$. *Calcule também* $H(X)$ *e* $H(Y)$.

- A entropia conjunta de variáveis aleatórias independentes é a soma das entropias de cada uma separadamente:

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

- Se duas variáveis aleatórias são iguais, $X = Y$, a entropia conjunta é

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

  isto é, as duas variáveis não contêm mais informação do que uma só.

- Se duas variáveis estão relacionadas deterministicamente por $Y = f(X)$, onde a função $f\colon \mathcal{X} \to \mathcal{Y}$ pode ser *não invertível*, então a entropia conjunta é

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

  e a entropia de Y satisfaz

$$H(Y) \leq H(X)$$

  com igualdade se e só se f é invertível.

Entropia condicional

Consideremos agora duas variáveis aleatórias dependentes X e Y para as quais é conhecida a distribuição condicional $p(y|x)$.

O que acontece à entropia de Y quando X é observado?

Consideremos agora duas variáveis aleatórias dependentes X e Y para as quais é conhecida a distribuição condicional $p(y|x)$.

O que acontece à entropia de Y quando X é observado?

Esta questão é simples, pois quando se observa $X = x$ ficamos apenas com uma distribuição $p(y|x)$ sobre a variável Y, com x fixo, em vez da distribuição condicional original com a variável x livre.

Assim, a entropia de Y com $X = x$ fixo é

$$H(Y|X = x) = -\sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x).$$

Esta conta pode ser repetida para cada valor x no alfabeto $\mathcal{X}$ obtendo-se entropias $H(Y|X = x)$ diferentes.

Conhecendo $p(x)$ podemos calcular o valor esperado destas entropias

$$\begin{aligned} E\big[H(Y|X = x)\big] &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x), \end{aligned}$$

o que leva à definição de entropia condicional.

Definição (Entropia condicional)

Dadas duas variáveis aleatórias X e Y, a entropia condicional é definida por

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x)$$

Propriedades:

- O condicionamento reduz a entropia:

$$H(Y|X) \leq H(Y).$$

- A entropia condicional satisfaz as seguintes igualdades:

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

estes resultados são análogos à regra da cadeia $p(x,y) = p(x)p(y|x)$.

A informação mútua mede o grau de dependência entre duas variáveis aleatórias X e Y.

Definição (Informação mútua)

Conhecendo-se a distribuição conjunta $p(x, y)$ e as distribuições marginais $p(x)$ e $p(y)$, define-se a informação mútua por

$$I(X; Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}.$$

1. Quando as variáveis são independentes, $p(x, y) = p(x)p(y)$, então a informação mútua anula-se:

$$I(X; Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log_2 \underbrace{\frac{p(x, y)}{p(x)p(y)}}_{=1} = 0.$$

2. A informação mútua é não negativa:

$$I(X; Y) \geq 0$$

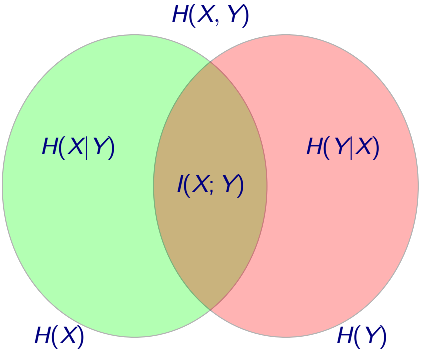
3. É limitada superiormente:

$$I(X; Y) \leq \min \left(H(X), H(Y)\right)$$

Podem provar-se as seguintes igualdades:

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

As igualdades que relacionam os conceitos de entropia, entropia conjunta, entropia condicional e informação mútua podem ser ilustrados no diagrama seguinte:



Entropia relativa (divergencia de Kullback-Leibler)

Definição (Entropia relativa)

A entropia relativa (ou divergência de Kullback-Leibler) entre duas distribuições $p(x)$ e $q(x)$ é definida por

$$D(p(x)\|q(x)) = \sum_{x\in\mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$$

A entropia relativa mede o "desvio" entre duas distribuições de probabilidade, mas não é uma distância:

- Satisfaz dois dos axiomas de uma distância

$$D(p\|p) = 0$$
$$D(p\|q) > 0, \text{ se } p \neq q$$

- Mas não satisfaz os axiomas de simetria e desigualdade triangular

$$D(p\|q) \neq D(q\|p)$$
$$D(p\|q) \nleq D(p\|r) + D(r\|q)$$

pelo que não é uma distância.

Das definições de entropia relativa e informação mútua

$$D\big(p(x)\|q(x)\big) = \sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)} \qquad \text{(Entropia relativa)}$$

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \qquad \text{(Informação mútua)}$$

observa-se que a informação mútua não é mais que a entropia relativa entre a distribuição conjunta $p(x,y)$ e o produto das marginais $p(x)q(x)$:

$$I(X;Y) = D\Big(p(x,y)\,\big\|\,p(x)p(y)\Big).$$

Conclui-se que

- se X e Y são variáveis aleatórias independentes, então

$$p(x,y) = p(x)p(y) \quad \Rightarrow \quad I(X;Y) = 0$$

- se X e Y são variáveis aleatórias dependentes, então

$$p(x,y) \neq p(x)p(y) \quad \Rightarrow \quad I(X;Y) > 0$$

Pretende-se inferir X a partir de observações de uma variável Y.
Supõem-se conhecidos:

- o *modelo* $p(y|x)$, que descreve como Y depende de X;
- o *prior* $p(x)$, que descreve a informação conhecida acerca de X.

Aplicando o teorema de Bayes obtém-se a distribuição *posterior* $p(x|y)$:

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

O processo de inferência é interpretado do seguinte modo:

1. tendo uma distribuição inicial $p(x)$, que contém toda a informação conhecida acerca de X antes de efectuar qualquer observação,
2. e conhecendo o modelo das observações $p(y|x)$,
3. é possível, fazendo uma observação de Y, corrigir a distribuição de X de modo a incorporar a nova informação observada, obtendo-se a distribuição *a posteriori* $p(x|y)$.

Pretende-se inferir X a partir de observações de uma variável Y.
Supõem-se conhecidos:

- o *modelo* $p(y|x)$, que descreve como Y depende de X;
- o *prior* $p(x)$, que descreve a informação conhecida acerca de X.

Aplicando o teorema de Bayes obtém-se a distribuição *posterior* $p(x|y)$:

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

O processo de inferência é interpretado do seguinte modo:

1. tendo uma distribuição inicial $p(x)$, que contém toda a informação conhecida acerca de X antes de efectuar qualquer observação,
2. e conhecendo o modelo das observações $p(y|x)$,
3. é possível, fazendo uma observação de Y, corrigir a distribuição de X de modo a incorporar a nova informação observada, obtendo-se a distribuição *a posteriori* $p(x|y)$.

A observação de Y aumenta ou reduz a incerteza acerca de X?

Resposta: Se calcularmos a entropia do prior $H(X)$ e a entropia $H(X|Y=y)$, observa-se que esta última pode ser maior ou menor que a do prior, dependendo do valor de y. Como é isto possível?

Exemplo

Sabe-se que uma doença, chamada zorgzifismo, afecta 1 em cada 1000 pessoas. Para verificar se uma pessoa tem zorgzifismo é feito um teste que infelizmente não é muito fiável e só acerta 90% das vezes.
Qual o valor da entropia $H(X)$ e de $H(X|Y=y)$?

1. A entropia $H(X)$ é

$$H(X) = -(0.999 \log_2 0.999 + 0.001 \log_2 0.001) \approx 0.0114 \text{ bits.}$$

2. A entropia $H(X|Y=y)$ é

$$H(X|Y=y) = -\sum_{x\in\mathcal{X}} p(x|y) \log_2 p(x|y).$$

onde a distribuição $p(x|y)$ pode ser calculada pela lei de Bayes:

$$p(x|y) = \frac{p(y|x)}{\sum_{x\in\mathcal{X}} p(y|x)p(x)} p(x).$$

Exemplo (continuação...)

Supondo que o teste deu positivo, $Y = P$, obtém-se

$$\Pr\{X = P | Y = P\} = \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.1 \times 0.999} \approx 0.00893$$
$$\Pr\{X = N | Y = P\} = \frac{0.1 \times 0.999}{0.9 \times 0.001 + 0.1 \times 0.999} \approx 0.99107$$

$$H(X|Y = P) \approx 0.0736 \text{ bits}$$

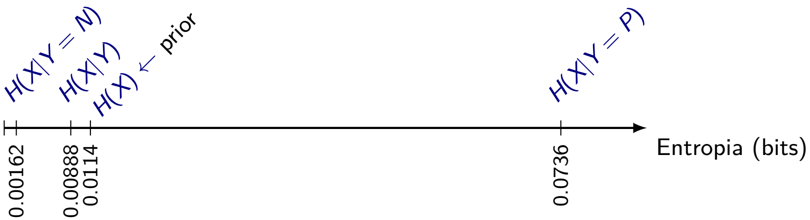
Supondo que o teste deu negativo, $Y = N$, obtém-se

$$\Pr\{X = P | Y = N\} = \frac{0.1 \times 0.001}{0.9 \times 0.999 + 0.1 \times 0.001} \approx 0.00011$$
$$\Pr\{X = N | Y = N\} = \frac{0.9 \times 0.999}{0.9 \times 0.999 + 0.1 \times 0.001} \approx 0.99989$$

$$H(X|Y = N) \approx 0.00162 \text{ bits}$$

Conclui-se que observar um teste positivo aumenta a entropia enquanto um teste negativo reduz a entropia. No entanto a entropia condicional reduz-se:

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2 p(x|y) = 0.00888 \text{ bits.}$$



À entropia relativa entre a distribuição *a posteriori* e o *prior*, $D\big(p(x|y) \,\|\, p(x)\big)$, da-se o nome de ganho de informação.

O seu valor esperado

$$E_Y\left[D\big(p(x|y) \,\|\, p(x)\big)\right]$$

coincide com a informação mútua $I(X;Y)$, ou de outro modo, coincide com a redução esperada na entropia $H(X) - H(X|Y)$.