

# Entropy

FIAP

ENGENHARIA DE SOFTWARE

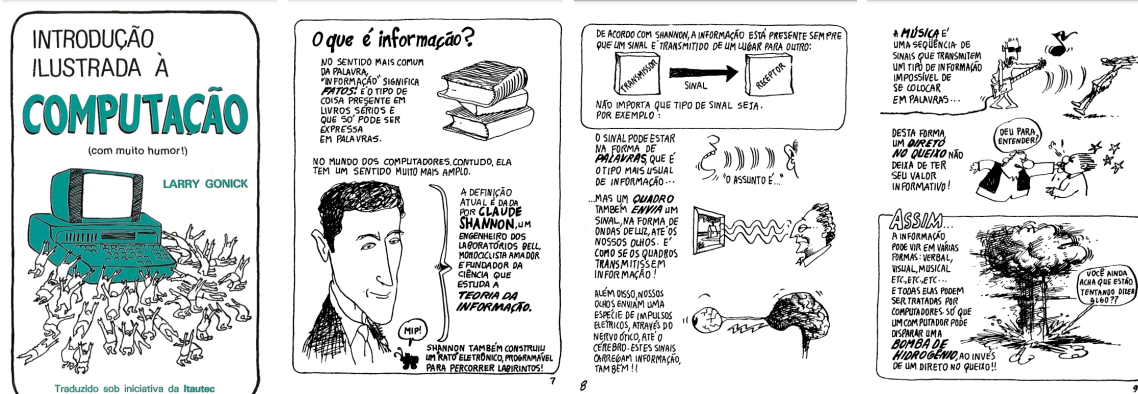
DATA SCIENCE & STATISTICAL COMPUTING

Prof. Daniel Carvalho - @danielscarvalho

Entropia e informação

Acompanhar online em [www.wolframcloud.com](http://www.wolframcloud.com)

## Introdução



O modelo de machine learning (ML - AI) árvore de decisão (Decision Tree) utilizados para classificação ou regressão, utiliza entropia e Gini para dividir/separar os dados na fase de aprendizagem.



Applied Mathematics &gt; Information Theory &gt;

## Entropy

In physics, the word entropy has important physical implications as the amount of "disorder" of a system. In mathematics, a more abstract definition is used. The (Shannon) entropy of a variable  $X$  is defined as

$$H(X) \equiv - \sum_x P(x) \log_2 [P(x)]$$

bits, where  $P(x)$  is the probability that  $X$  is in the state  $x$ , and  $P \log_2 P$  is defined as 0 if  $P = 0$ . The joint entropy of variables  $X_1, \dots, X_n$  is then defined by

$$H(X_1, \dots, X_n) \equiv - \sum_{x_1} \dots \sum_{x_n} P(x_1, \dots, x_n) \log_2 [P(x_1, \dots, x_n)].$$

<https://mathworld.wolfram.com/Entropy.html>

LLM GPT-4:



Definir informação



2/2

A informação é um conceito fundamental em ciência da informação e teoria da comunicação. Em termos gerais, informação refere-se a dados ou conhecimento que pode ser armazenado, recuperado, e transmitido. Na teoria da informação, a informação é frequentemente quantificada de acordo com a quantidade de incerteza que elimina.

Uma definição formal de informação foi introduzida por Claude Shannon em 1948, onde ele descreveu a informação como uma medida da incerteza em uma variável aleatória. Esta é conhecida como entropia de Shannon, e é calculada usando a fórmula:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

onde  $H(X)$  é a entropia da variável aleatória  $X$ ,  $p(x_i)$  é a probabilidade de um determinado resultado  $x_i$ , e a soma é feita sobre todos os possíveis resultados.

Em outras palavras, a informação está relacionada à redução da incerteza quando recebemos dados sobre o estado de um sistema ou evento. Quanto mais incerto um evento, maior a quantidade de informação que ele contém quando ocorre.

Se precisar de mais detalhes ou exemplos sobre como calcular a quantidade de informação usando a entropia, sinta-se à vontade para perguntar!



Em computação definir entropia



2/2

Na computação, a entropia é uma medida da incerteza ou aleatoriedade em um conjunto de dados. Ela é frequentemente usada para avaliar a quantidade de desordem ou imprevisibilidade em um sistema, especialmente em contextos como criptografia, compressão de dados e teoria da informação.

### Contextos Comuns de Entropia na Computação:

#### 1. Criptografia:

- Em criptografia, a entropia é crucial para garantir a segurança de chaves criptográficas. Uma chave com alta entropia é considerada mais segura, pois é mais difícil de prever ou adivinhar.

#### 2. Compressão de Dados:

- Na compressão de dados, a entropia ajuda a determinar o limite teórico de quão bem os dados podem ser comprimidos. Dados com baixa entropia têm padrões repetitivos e podem ser mais facilmente comprimidos, enquanto dados com alta entropia são mais aleatórios e menos compressíveis.

#### 3. Teoria da Informação:

- Como mencionado anteriormente com a entropia de Shannon, na teoria da informação, a entropia é usada para quantificar a quantidade média de informação produzida por uma fonte de dados.

### Cálculo da Entropia:

A fórmula clássica para calcular a entropia de Shannon de um conjunto de dados é:

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

onde:

- $H(X)$  é a entropia da variável aleatória  $X$ .
- $p(x_i)$  é a probabilidade de ocorrência do evento  $x_i$ .
- A soma é feita sobre todos os possíveis eventos.

Na computação, é importante que os sistemas que dependem de geração de números aleatórios, como sistemas criptográficos, tenham fontes de entropia suficientemente altas para garantir a segurança e a imprevisibilidade.

Se precisar de exemplos específicos ou de uma explicação mais detalhada sobre como a entropia é utilizada em algum desses contextos, sinta-se à vontade para perguntar!



Como a entropia é usada no modelo de machine learning árvore de decisão para calcular o gini



Na árvore de decisão em machine learning, tanto a entropia quanto o índice de Gini são usados para medir a impureza ou a desordem de um nó. Ambos são critérios para determinar como dividir os nós da árvore.

## Entropia

A entropia é usada para calcular o **ganho de informação**, que ajuda a decidir a melhor divisão em um nó. A fórmula é a mesma da entropia de Shannon:

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

Em uma árvore de decisão, escolhemos a divisão que resulta no maior ganho de informação, ou seja, a maior redução de entropia.

## Índice de Gini

O índice de Gini é uma medida alternativa da impureza de um nó. Ele é calculado como:

$$Gini = 1 - \sum_i (p(x_i))^2$$

O índice de Gini mede a probabilidade de um elemento ser classificado incorretamente se for aleatoriamente rotulado de acordo com a distribuição de classes no conjunto de dados.

## Uso em Árvores de Decisão

### 1. Divisão de Nós:

- **Entropia:** Calcula-se a entropia atual e a entropia após a divisão. O ganho de informação é a diferença, e escolhemos a divisão com o maior ganho.
- **Índice de Gini:** Calcula-se o índice de Gini antes e depois da divisão e escolhe-se a divisão que minimiza a impureza.

2. **Critério de Escolha:** Dependendo do algoritmo implementado (como ID3, C4.5, CART), pode-se usar um ou ambos os critérios para decidir as divisões.

Ambas as medidas ajudam a construir uma árvore que minimiza erros de classificação em novos dados, mas são usadas em contextos ligeiramente diferentes dependendo dos requisitos do modelo e dos dados.

## Exemplos

```
In[ ]:= Entropy[{10, 10, 10, 10, 10, 10, 10, 10}] // N
```

```
Out[ ]:= 0.
```

```
In[ ]:= Entropy[{1, 2, 3, 4, 5, 6, 7, 8}] // N
```

```
Out[ ]:= 2.07944
```

```
Entropy[Range[10000]] // N
```

```
Out[ ]:= 9.21034
```

```
In[ ]:= Entropy[RandomInteger[10, 10 000]] // N
```

```
Out[ ]:= 2.39777
```

```
In[ ]:= Entropy[Table[1, 10 000]]
```

```
Out[ ]:= 0
```

```
Entropy[RandomChoice[{1, 2}, 10 000]] // N
```

```
Out[ ]:= 0.693143
```

Imagens de 1000px x 1000px PNG (Figura) e JPEG (Foto)

```
In[ ]:= Entropy[

```

```
Out[ ]:= 2.61359
```

```
In[ ]:= Entropy[

```

```
Out[ ]:= 10.8693
```

```
In[16]:= SetDirectory[NotebookDirectory[]]
```

```
Out[16]=
```

```
/x4n/work/FIAP/FIAP-Python-Data-Science-Statistics/LAB001
```

```
In[22]:=
```





```
cd /x4n/work/FIAP/FIAP-Python-Data-Science-Statistics/LAB001
ls -l bear*
```

```
-rw-rw-r-- 1 fito fito 216850 out 17 10:06 bear.jpg
```

```
-rw-rw-r-- 1 fito fito 177083 out 17 10:06 bear.png
```

Out[22]=

Success[   Command: cd /x4n/work/FIAP/FIAP-Python-Data-Science-Statistics/LAB001  
ls -l bear\*  
ExitCode: 0 ]

Letra da música Faroeste Caboclo:

```
fc = Import["faroeste-caboclo.txt"];
```

```
In[4]:= StringLength[fc]
```

```
Out[4]= 6413
```

```
In[6]:= Entropy[fc] // N
```

```
Out[6]= 3.08041
```

```
In[13]:= ResourceFunction["SequenceGraph"][StringSplit[ToLowerCase[fc]]]
```

```
Out[13]=
```





Letra da música Single Ladies:

```
In[3]:= sl = Import["single-ladies.txt"];
```

```
In[7]:= StringLength[sl]
```

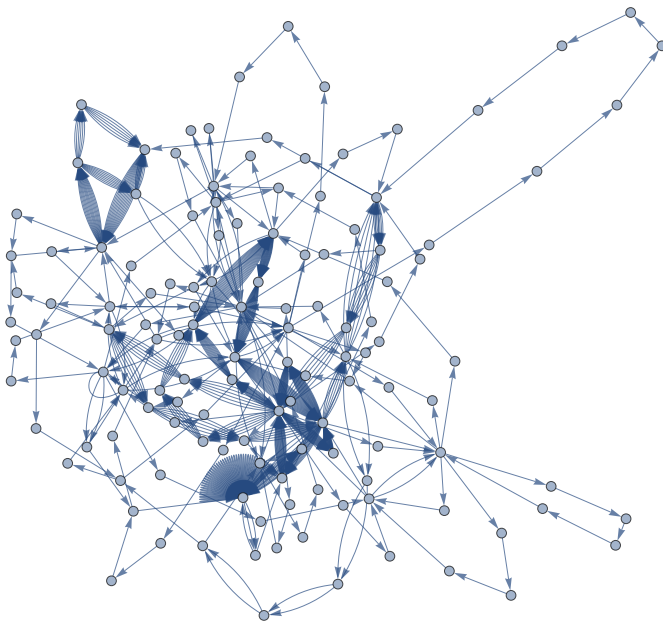
```
Out[7]= 2658
```

```
In[9]:= Entropy[sl] // N
```

```
Out[9]= 2.95203
```

```
In[14]:= ResourceFunction["SequenceGraph"][StringSplit[ToLowerCase[sl]]]
```

```
Out[14]=
```



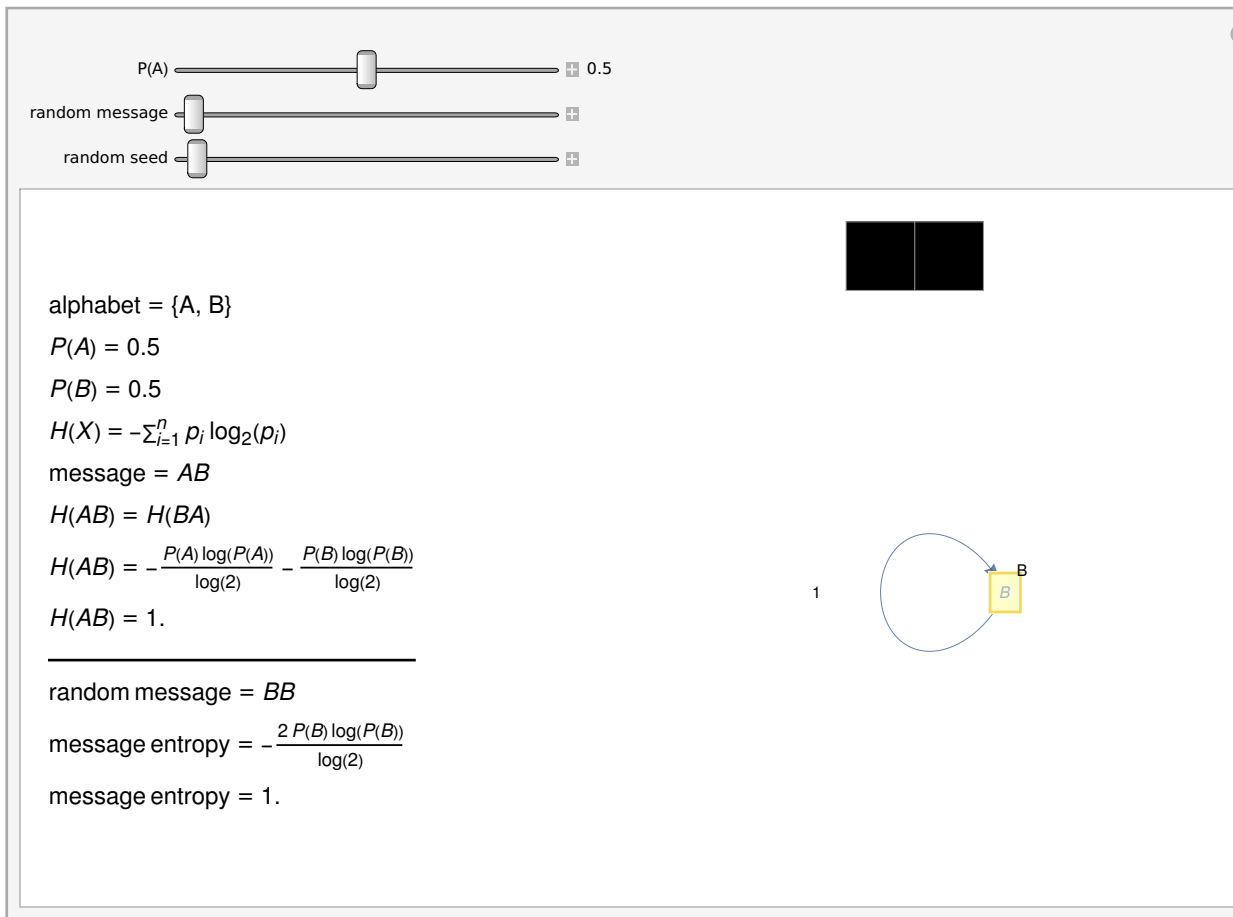
## Exemplos interativos

Wolfram Demonstration Project:

<https://demonstrations.wolfram.com/>

Entropy of a Message Using Random Variables

<https://demonstrations.wolfram.com/EntropyOfAMessageUsingRandomVariables/>



## Referências

- [1] C. E. Shannon, "Prediction and Entropy of Printed English," Bell Systems Technical Journal, 30, 1951 pp. 50–64.  
[www.ics.uci.edu/~fowlkes/class/cs177/shannon\\_51.pdf](http://www.ics.uci.edu/~fowlkes/class/cs177/shannon_51.pdf).