Trabalho de Matemática

Nome: *Pedro Jorge de Souza Colombrino*
RA: *0051352311015* ***Curso:*** *Ciência de Dados*

Tema: Entropia

Alexandre Garcia de Oliveira

Faculdade de Tecnologia da Baixada Santista Rubens Lara

Introdução:

Busco esclarecer neste arquivo questões como: "O que é entropia?", "Como fica o código?"e ilustrar por meio deste documento a resolução e explicar suas funções.
Vale comentar que a base de dados utilizada neste projeto reflete nos casos de covid no Brasil. Estes dados são de domínio publico e podem ser acessados clicando aqui.

O que é entropia?

A entropia é um conceito fundamental em teoria da informação e probabilidade. É uma medida da incerteza ou imprevisibilidade presente em um conjunto de dados, uma variável aleatória ou um sistema.

Em termos simples, a entropia quantifica a quantidade de informação necessária para descrever ou representar um evento, ou uma distribuição de probabilidade. Quanto maior a entropia, maior é a incerteza e a quantidade de informação necessária.

A entropia é calculada usando a probabilidade dos diferentes eventos em um conjunto de dados ou a distribuição de probabilidade de uma variável aleatória. Quanto mais uniforme for a distribuição de probabilidade, maior será a entropia, indicando maior incerteza. Por outro lado, quando a distribuição de probabilidade é altamente concentrada em um ou alguns poucos eventos, a entropia será menor, indicando menor incerteza.

A fórmula geral para calcular a entropia de um conjunto de dados discreto é:

$$H(X) = -(p(i) * log2(p(i))) \tag{1}$$

Onde H(X) é a entropia do conjunto de dados X, p(i) é a probabilidade do evento i ocorrer e a soma é realizada para todos os eventos possíveis.

A entropia é amplamente utilizada em diversos campos, incluindo ciência da computação, aprendizado de máquina, teoria da informação, estatística e física, sendo uma medida essencial para avaliar a informação e a aleatoriedade em sistemas e dados.

Entropia

Suponha que temos um conjunto de dados que representa o clima de um determinado local em dias diferentes. Os dados são os seguintes:

Dias: 1, 2, 3, 4, 5
Clima: Ensolarado, Chuvoso, Ensolarado, Nublado, Chuvoso

Para calcular a entropia desse conjunto de dados, precisamos determinar a probabilidade de ocorrência de cada categoria/classe. Neste caso, temos três categorias/classes: Ensolarado, Chuvoso e Nublado.

Ensolarado: 2 ocorrências
Chuvoso: 2 ocorrências
Nublado: 1 ocorrência
Agora, vamos calcular a probabilidade de ocorrência de cada categoria. Dividimos o número de ocorrências de cada categoria pelo total de dias (5).

Ensolarado: $2/5 = 0.4$
Chuvoso: $2/5 = 0.4$
Nublado: $1/5 = 0.2$
Com as probabilidades determinadas, podemos calcular a entropia usando a fórmula:

$$H(X) = -(0.4 * log2(0.4) + 0.4 * log2(0.4) + 0.2 * log2(0.2)) \tag{2}$$

Agora, vamos calcular o valor numérico da entropia:

$$H(X) = -(0.4 * (-1.3219) + 0.4 * (-1.3219) + 0.2 * (-2.3219)) \tag{3}$$

$$H(X) = -(-0.5288 - 0.5288 - 0.4644) \tag{4}$$

$$H(X) = -(-1.522) \tag{5}$$

$$H(X) = 1.522 \tag{6}$$

Portanto, a entropia desse conjunto de dados é aproximadamente **1.522**. Isso indica a medida de incerteza ou diversidade presente nas categorias/classe do conjunto, considerando a probabilidade de ocorrência de cada uma.

Entropia Máxima

A entropia máxima ocorre quando todas as categorias em um conjunto de dados têm a mesma probabilidade de ocorrência. Isso significa não haver preferência ou padrão na distribuição das categorias, resultando em máxima incerteza ou diversidade.
Vamos considerar um exemplo para ilustrar a entropia máxima. Suponha que temos um conjunto de dados com 8 elementos, divididos igualmente em 4 categorias:
Categoria A: 2 ocorrências
Categoria B: 2 ocorrências
Categoria C: 2 ocorrências
Categoria D: 2 ocorrências
Nesse caso, todas as categorias têm a mesma probabilidade de ocorrência, que é $2/8 = 0.25$.
Agora, vamos calcular a entropia máxima usando a fórmula:

$$H_max = -(0.25 * log2(0.25) + 0.25 * log2(0.25) + 0.25 * log2(0.25) + 0.25 * log2(0.25)) \tag{7}$$

Podemos simplificar a fórmula, pois todos os termos são iguais:

$$H_max = -4 * (0.25 * log2(0.25)) \tag{8}$$

Agora, vamos calcular o valor numérico da entropia máxima:

$$H_max = -4 * (0.25 * (-2)) \tag{9}$$

$$H_max = -4 * (-0.5) \tag{10}$$

$$H_max = 2 \tag{11}$$

Portanto, a entropia máxima para esse conjunto de dados é 2. Isso ocorre porque todas as categorias têm a mesma probabilidade de ocorrência, o que resulta em máxima incerteza ou diversidade possível nas categorias.

Explicação do código

O código importa as bibliotecas necessárias e lê um arquivo CSV chamado 'dados _covid.csv' usando pandas. Ele seleciona uma coluna de interesse e calcula a probabilidade de cada valor na coluna. A entropia dos dados é então calculada pela fórmula -np.sum(probabilidades * np.log2(probabilidades)).

A entropia máxima é calculada pela fórmula np.log2(len(probabilidades)). O algoritmo k-NN é então usado para calcular uma entropia aproximada. O algoritmo calcula a distância entre cada ponto de dados e seus k vizinhos mais próximos. A entropia de cada ponto de dados é então calculada usando a fórmula -np.sum(proporcoes * np.log2(proporcoes)), onde proporcoes é a proporção da contribuição de cada vizinho para a distância total.

Por fim, o código imprime a entropia calculada, a entropia máxima e a entropia aproximada no console.

Código em python

Este código está retirando os números para o seu funcionamento por meio de uma conexão com outro arquivo estabelecida pela biblioteca pandas. Perceba onde está escrito **"Casos"** e para explorar outras opções ou adicionar mais, basta alterar para satisfazer seus gostos.

```python
import pandas as pd
import numpy as np
from sklearn.neighbors import NearestNeighbors

dados = pd.read_csv('dados_covid.csv')

# Seleção das colunas desejadas
colunas_dados = ['Casos']  # professor, substitua pela coluna de dados que mais lhe interessar, para saber as possibilidades, abra o arquivo em csv pra ver as colunas.
dados = dados[colunas_dados]

#probabilidade
quantidade_elementos = dados.size
probabilidades = dados.groupby(colunas_dados).size() / quantidade_elementos

#entropia
entropia = -np.sum(probabilidades * np.log2(probabilidades))

#entropia máxima
entropia_maxima = np.log2(len(probabilidades))

# usando o k-NN para cálculo da entropia
matriz_dados = dados.values
num_linhas = len(matriz_dados)

k = min(5, num_linhas - 1)

knn = NearestNeighbors(n_neighbors=k + 1)
knn.fit(matriz_dados)

distancias, indices_vizinhos = knn.kneighbors(matriz_dados)
distancias = distancias[:, 1:]
indices_vizinhos = indices_vizinhos[:, 1:]

entropias = []
for i in range(num_linhas):
    vizinhos = indices_vizinhos[i]
    dists = distancias[i]
    dists[dists == 0] = 1e-10
    proporcoes = 1 / dists
    proporcoes /= np.sum(proporcoes)
    entropia_local = -np.sum(proporcoes * np.log2(proporcoes))
    entropias.append(entropia_local)

entropia_media = np.mean(entropias)

print("Entropia aproximada:", entropia)
print("Entropia máxima aproximada:", entropia_maxima)
print("Entropia média aproximada (k-NN):", entropia_media)
```

Console

Perceba que na saída do console, ele entrega a entropia, entropia máxima e uma entropia (k-nn).
Ressalta-se que todas elas se tratam de uma aproximação.
Meu ver lógico por ter essas três são que:
Todos os números que apresentam a mesma probabilidade de acontecerem retornarão resultados iguais à entropia na entropia máxima, ou seja:

$$Entropia = Entropia_Max \tag{12}$$

Que nos diz indiretamente que, "Tudo pode acontecer, tenha cuidado!".

Ao utilizar o KNN, percebe-se que essa regularidade se tornou irregular, porém dúvidas vinham e dúvidas iam então decide-se inserir ambas.

```
In [35]: runfile('D:/Python para Entropia/Entropia.py', wdir='D:/Python para Entropia')
Entropia aproximada: 2.321928094887362
Entropia máxima aproximada: 2.321928094887362
Entropia média aproximada (k-NN): 1.574766836144621
```

Csv

Um arquivo CSV (*Comma-Separated Values*) é um formato de arquivo utilizado para armazenar dados tabulares, como uma planilha, de forma simples e legível por máquinas. O nome "Comma-Separated Values"se deve ao fato de os valores dentro do arquivo serem separados por vírgulas.
Em um arquivo CSV, cada linha geralmente representa uma entrada de dados, e os valores são organizados em colunas separadas por vírgulas. Cada valor pode ser um texto ou um número, e as colunas podem ter um cabeçalho que descreve o conteúdo das colunas.

D: > Python para Entropia > dados_covid.csv

```
1  Região,Casos,Óbitos,Incidência/100mil hab.,Mortalidade/100mil hab.,Atualização
2  Centro-Oeste,4324932,66154,26538.1,405.9,05/05/2023 11:28
3  Sul,7980261,111208,26622.2,371,05/05/2023 11:28
4  Norte,2905231,51657,15762.8,280.3,05/05/2023 11:28
5  Nordeste,7359485,135046,12895.2,236.6,05/05/2023 11:28
6  Sudeste,14918062,337768,16881.1,382.2,05/05/2023 11:28
7
```

Excel

Neste projeto, foi feito uma planilha em Excel para melhor ilustração dos dados, perceba que este arquivo existe somente para fins ilustrativos e não possui usos no código.

Região	Casos	Óbito	Incidência/100mil hab	Mortalidade/100mil hab	Atualização
Centro-Oeste	4.324.932	66.154	26538,1	405,9	05/05/2023 11:28
Sul	7.980.261	111.208	26622,2	371	05/05/2023 11:28
Norte	2.905.231	51.657	15762,8	280,3	05/05/2023 11:28
Nordeste	7.359.485	135.046	12895,2	236,6	05/05/2023 11:28
Sudeste	14.918.062	337.768	16881,1	382,2	05/05/2023 11:28