ENTROPIA

Atividade para P2

Gustavo Costa Arakaki

Sumário

1 Entropia

1.1 Conceito e fórmula

Para a área de Teoria da Informação, a entropia é definida como sendo uma forma de medir o grau médio de incerteza a respeito de fontes de informação, o que consequentemente permite a quantificação da informação presente que flui no sistema. Em termos simples, o conceito de entropia se associa à ideia de que, quanto mais incerto é o resultado de um experimento aleatório, maior é a informação que se obtém ao observar a sua ocorrência.

Uma das formas de se calcular a entropia é utilizando a seguinte fórmula:

$$\mathrm{H(x)} = -\sum_{i=1}^{n} P(xi) \cdot log_2 P(xi)$$

Onde $P(xi)$ é a probabilidade do i-ésimo resultado para a variável x.

1.2 Python

A entropia pode ser implementada em Python da seguinte forma:

```python
import math as m

#Probabilidade das classes
def probs(dados):
    return [dados.count(x)/len(dados) for x in set(dados)]

#Entropia
def entropia(probs):
    return -sum([x * m.log2(x) for x in probs])

#Entropia Maxima
def MAXentropia(dados):
    return m.log2(len(set(dados)))
```

No trecho acima, primeiro temos o comando que importa a biblioteca math do python para ser usada em cálculos logaritmos. Em seguida temos três funções: probs, entropia e MAXentropia

A função probs recebe um conjunto de dados e retorna uma lista contendo as probabilidades de cada classe existente no conjunto. A função entropia recebe uma lista com as probabilidades e realiza o cálculo da entropia, primeiro ela cria uma lista com os valores aplicados na fórmula e depois soma os valores com a função sum e em seguida multiplica o valor gerado por -1, ambas funções usam o list comprehension.

A última função (MAXentropia) realiza o cálculo da entropia máxima partindo de uma base de dados.

2 Cidades brasileiras

2.1 Base de dados

A base de dados escolhida para o trabalho contém informações sobre todas as cidades brasileiras e é uma ferramenta valiosa para diversos fins, desde análises demográficas e econômicas até desenvolvimento de aplicativos e sistemas de geolocalização. Essa base de dados abrange um vasto conjunto de informações sobre as cidades do Brasil, como nomes, população, localização geográfica, aspectos socioeconômicos, indicadores de desenvolvimento, entre outros dados relevantes.

Para fins de simplicidade e demonstração de conceitos apresentados anteriormente, a base foi reduzida para três campos, sendo eles: nome, unidade federativa (estado) e classificação rural ou urbana.

O campo de unidades federativas, representadas por siglas, possuí as seguintes 27 classes:
Acre - AC
Alagoas - AL
Amapá - AP
Amazonas - AM
Bahia - BA
Ceará - CE
Distrito Federal - DF
Espírito Santo - ES
Goiás - GO
Maranhão - MA
Mato Grosso - MT
Mato Grosso do Sul - MS
Minas Gerais - MG
Pará - PA
Paraíba - PB
Paraná - PR
Pernambuco - PE
Piauí - PI
Rio de Janeiro - RJ
Rio Grande do Norte - RN

Rio Grande do Sul - RS
Rondônia - RO
Roraima - RR
Santa Catarina - SC
São Paulo - SP
Sergipe - SE
Tocantins - TO

A classificação de uma cidade como rural ou urbana depende de diferentes critérios e definições utilizadas por diferentes instituições e países.

No contexto brasileiro, o Instituto Brasileiro de Geografia e Estatística (IBGE) define critérios para classificar uma cidade como rural ou urbana. O critério mais comum é o critério populacional.

Segundo o IBGE, uma cidade é considerada urbana quando possui uma concentração populacional e infraestrutura associada, como saneamento básico, eletricidade, serviços públicos, comércio e indústria. Geralmente, as cidades com mais de 20 mil habitantes são classificadas como urbanas no Brasil.

As categorias "rural remota", "rural adjacente", "intermediário remoto"e "intermediário adjacente"são classificações adicionais utilizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para fornecer mais detalhes sobre as áreas rurais do Brasil.

Essas classificações levam em consideração critérios como distância dos centros urbanos, infraestrutura e acesso a serviços básicos. Vejamos o significado de cada uma delas:

Rural Remota: Refere-se a áreas rurais localizadas em regiões isoladas, distantes dos centros urbanos.

Essas áreas geralmente possuem pouca infraestrutura, acesso limitado a serviços básicos e podem estar situadas em regiões geograficamente desafiadoras, como florestas, montanhas ou regiões de difícil acesso.

Rural Adjacente: Essa categoria engloba áreas rurais que estão próximas a centros urbanos, mas ainda possuem características predominantemente rurais.

São áreas que mantêm uma forte ligação com a agricultura, pecuária ou atividades relacionadas ao meio rural, mesmo estando em proximidade com áreas urbanizadas.

Intermediário Remoto: Essa classificação se aplica a áreas rurais que estão em uma posição intermediária entre as categorias "rural remota"e "rural adjacente".

Essas áreas podem apresentar um nível razoável de infraestrutura e serviços em comparação com as áreas remotas, mas ainda possuem características predominantemente rurais e podem enfrentar desafios em termos de acesso e desenvolvimento.

Intermediário Adjacente: Essa categoria engloba áreas rurais que estão em uma posição intermediária entre as categorias "rural adjacente"e "urbana".

Essas áreas podem estar em transição entre o meio rural e o urbano, apresentando um

maior nível de infraestrutura, serviços e desenvolvimento em comparação com as áreas rurais adjacentes, mas ainda sem atingir o nível de urbanização completo.

Essas categorias adicionais são utilizadas para melhor compreender a diversidade e as características específicas das áreas rurais do Brasil, considerando fatores como localização geográfica, infraestrutura e acesso a serviços.

Essas classificações auxiliam no planejamento e desenvolvimento de políticas públicas direcionadas a cada tipo de área rural.

2.2 Prática da entropia em python

O cálculo da entropia dos dados, assim como sua entropia máxima, pode ser calculada e visualizada em python utilizando o script já apresentado da seguinte forma:

```
import pandas as pd
from Entropia_funcs import probs, entropia, MAXentropia
```

Primeiro temos que importar a biblioteca pandas do python para manipular a base de dados que está em xlsx(open excel) e importar também as funções criadas no arquivo anterior (os arquivos devem estar na mesma pasta).

```
cidades = pd.read_excel('cidades.xlsx')
ruralUrban = cidades.RURAL_URBAN.values.tolist()
estados = cidades.STATE.values.tolist()
```

Criamos variáveis para armazenar as informações retiradas da base de dados, na variável cidades usamos a biblioteca pandas para fazer a leitura do arquivo e manter suas informações. Na variável ruralUrban filtramos a coluna da classificação de rural ou urbana para pegar somente essa informação das cidades e realizamos o mesmo procedimento com os estatos.

```
print(f"Entropia dos estados: {entropia(probs(estados))}")
print(f"Entropia máxima dos estados: {MAXentropia(estados)}")


print(f"Entropia das cidades rurais e urbanas: {entropia(probs(ruralUrban))}")
print(f"Entropia máxima das cidades rurais e urbanas: {MAXentropia(ruralUrban)}")
```

Por fim, utilizamos a função print para mostrar na tela a entropia e a entropia máxima dos dados fazendo uso das funções antes apresentadas. Note que fazemos uma composição de função no cálculo da entropia de modo que a base de dados é primeiro transformada em uma lista de probabilidade das classes e em seguida essa lista é usada como parâmetro para ser calculada a entropia.

No caso da entropia máxima ela recebe a base de dados e consegue verificar o número de classes para ser feito o cálculo, os resultados podem ser vistos a seguir:

```
Entropia dos estados: 4.173646396737657
Entropia máxima dos estados: 4.754887502163468
Entropia das cidades rurais e urbanas: 1.6629794610837205
Entropia máxima das cidades rurais e urbanas: 2.321928094887362
```

3 Conclusão

Após realizar o cálculo da entropia dos dados das cidades brasileiras, foi possível chegar a uma conclusão significativa sobre a distribuição da informação presente nessa base de dados.

O valor da entropia dos estados foi alto, isso indica uma maior diversidade e variabilidade nos dados das cidades brasileiras. Isso significa que as cidades apresentam uma ampla gama de características em relação a sua distribuição na área territorial do país.

Por outro lado, o valor da entropia da classificação de rural urbano foi baixo, isso indica uma menor diversidade e uma maior homogeneidade nos dados das cidades brasileiras. Isso sugere que as cidades apresentam características semelhantes. Nesse caso, podemos concluir que as cidades brasileiras possuem uma maior uniformidade em termos de suas características.

Em suma, a análise da entropia dos dados das cidades brasileiras permite tirar conclusões sobre a diversidade ou homogeneidade das características das cidades, bem como identificar os atributos que mais contribuem para essa variação. Essas conclusões podem ser úteis para o planejamento urbano, desenvolvimento regional, tomada de decisões políticas e outras áreas que dependem de um melhor entendimento da diversidade das cidades brasileiras.