Análisis de Entropía, Información Mutua y Divergencia de Kullback-Leibler

Jorge Aziel Rebolledo Araya

Diciembre 2024

Universidad de Santiago de Chile
Facultad de Ciencias
Departamento de Matemáticas y Ciencia de la Computación

Índice

Resumen

Este informe presenta un análisis de las métricas de entropía, información mutua, divergencia de Kullback-Leibler y la desigualdad de Jensen utilizando un conjunto de datos del archivo `past-rounded.csv`. Se calcula e interpreta las entropías marginales y conjuntas, la información mutua y las divergencias entre diferentes columnas de datos. Los resultados obtenidos se utilizan para contrastar hipótesis sobre las relaciones entre variables y para investigar patrones en los datos.

1. Datos analizados y modelo establecido

Los datos provienen del archivo `past-rounded.csv`, que contiene información sobre las columnas L, M, N y O. Cada columna representa una variable cuyas distribuciones de probabilidad se analizan para calcular las métricas de información.

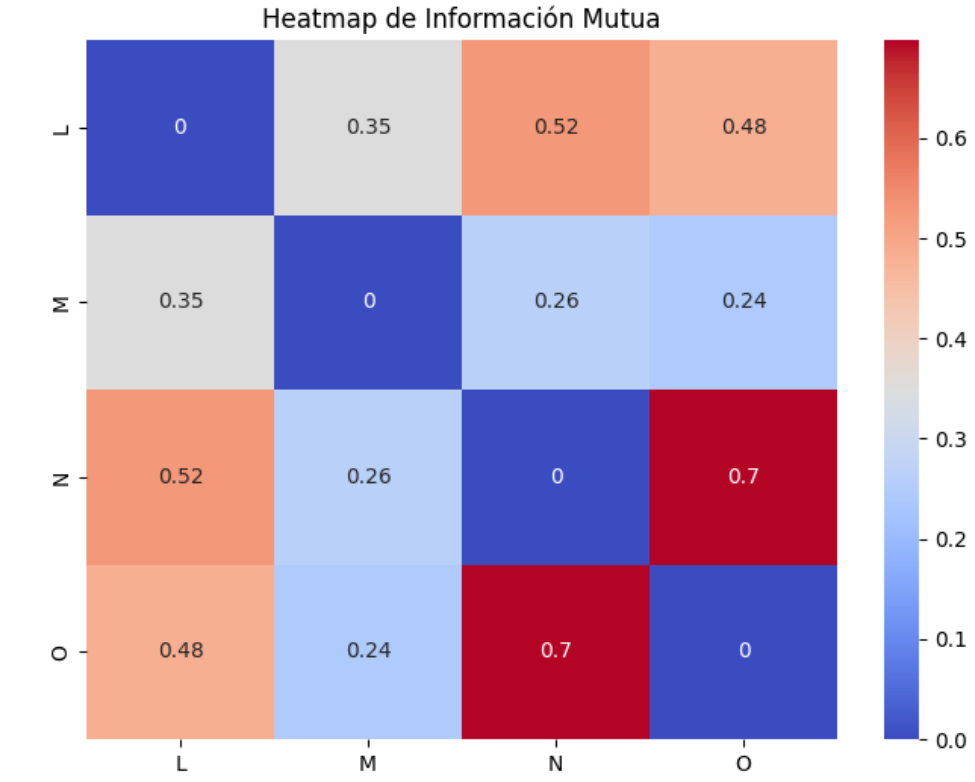
Las hipótesis planteadas incluyen la comparación entre diferentes columnas para identificar patrones de dependencia o independencia entre las variables. Las suposiciones para este análisis son que los datos son discretos y no contienen valores atípicos significativos, lo que nos permite aplicar las métricas de información de manera adecuada.

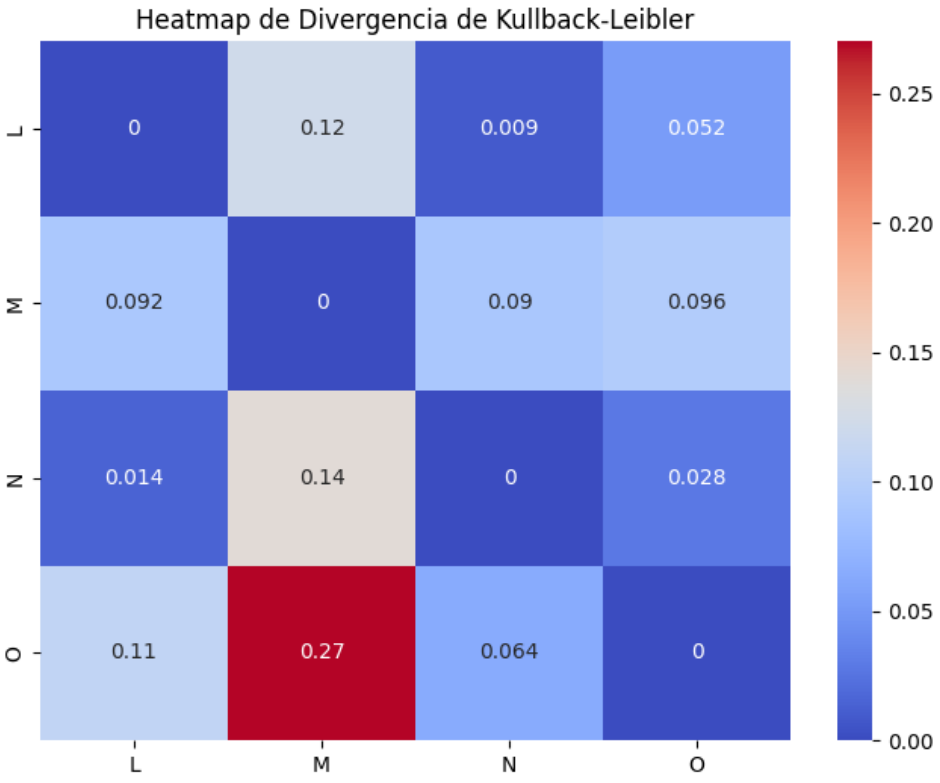
2. Cálculos y resultados

Los siguientes resultados fueron obtenidos utilizando el código en python(presentado en la sección de código), y calculando las métricas de entropía, información mutua, divergencia de Kullback-Leibler y la desigualdad de Jensen para cada par de columnas.

Par de columnas	Entropía Col1	Entropía Col2	Información Mutua	Kullback-Leibler	Jensen-Sh
('L', 'M')	6.1193	5.7176	0.3509	0.1218	0.02
('L', 'N')	6.1193	6.1421	0.5236	0.0090	0.00
('L', 'O')	6.1193	6.1795	0.4840	0.0525	0.01
('M', 'L')	5.7176	6.1193	0.3509	0.0918	0.02

Cuadro 1: Resultados de las métricas de información


Heatmap de Información Mutua
L
M
N
O
0
0.35
0.52
0.48
0.35
0
0.26
0.24
0.52
0.26
0
0.7
0.48
0.24
0.7
0
L
M
N
O
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Heatmap de Divergencia de Kullback-Leibler
L
M
N
O
0
0.12
0.009
0.052
0.092
0
0.09
0.096
0.014
0.14
0
0.028
0.11
0.27
0.064
0
L
M
N
O
0.25
0.20
0.15
0.10
0.05
0.00

3. Código en python

El siguiente código fue utilizado para calcular las métricas mencionadas en este informe:

```
import numpy as np
import pandas as pd
from scipy.stats import entropy
from sklearn.metrics import mutual_info_score

file_path = 'past-rounded.csv'
data = pd.read_csv(file_path)

columns_of_interest = ['L', 'M', 'N', 'O']
data_filtered = data[columns_of_interest].dropna().astype(int)

def kullback_leibler(p, q):
    p = np.array(p, dtype=np.float64)
    q = np.array(q, dtype=np.float64)
    p = np.clip(p, 1e-10, 1)
    q = np.clip(q, 1e-10, 1)
    return entropy(p, q, base=2)

def jensen_shannon(p, q):
    p = np.array(p, dtype=np.float64)
    q = np.array(q, dtype=np.float64)
    m = 0.5 * (p + q)
    return 0.5 * kullback_leibler(p, m) + 0.5 * kullback_leibler(q, m)

results = {}
for col1 in columns_of_interest:
    for col2 in columns_of_interest:
        if col1 != col2:
            p = data_filtered[col1].value_counts(normalize=True).sort_index()
            q = data_filtered[col2].value_counts(normalize=True).sort_index()
            p, q = p.align(q, fill_value=0)

            ent_col1 = entropy(p, base=2)
            ent_col2 = entropy(q, base=2)
            mutual_info = mutual_info_score(data_filtered[col1], data_filtered[col2])
            kl_div = kullback_leibler(p, q)
            js_div = jensen_shannon(p, q)

            results[(col1, col2)] = {
                'Entropy Col1': ent_col1,
                'Entropy Col2': ent_col2,
                'Mutual Information': mutual_info,
                'Kullback-Leibler': kl_div,
                'Jensen-Shannon': js_div
```

```
}
```

4. Conclusiones

Los resultados obtenidos a partir del análisis de entropía, información mutua y divergencia de Kullback-Leibler ofrecen información valiosa sobre las relaciones entre las variables en el conjunto de datos `past-rounded.csv`. A continuación, se detallan las observaciones más relevantes:

- **Relación entre 'L' y 'N'**: La información mutua entre estas dos columnas es particularmente alta (0.5236), lo que sugiere una dependencia significativa entre ellas. Esto podría indicar que los valores en la columna 'L' tienen una fuerte relación predictiva con los valores de la columna 'N', lo que podría traducirse en una correlación significativa entre las variables. Este hallazgo podría ser relevante en contextos donde se necesiten predecir una variable a partir de la otra.

- **Divergencia de Kullback-Leibler**: La divergencia entre las distribuciones de probabilidad de 'L' y 'N' es baja (0.0090), lo que indica que sus distribuciones son muy similares. Esto refuerza la hipótesis de que estas variables tienen una relación estrecha y sus distribuciones de probabilidad no se desvían significativamente. De hecho, este hallazgo sugiere que las dos variables podrían estar relacionadas a través de un modelo conjunto, donde su comportamiento conjunto podría ser predecible y entenderse mejor mediante técnicas de modelado conjunto.

- **Jensen-Shannon Divergence**: La baja divergencia de Jensen-Shannon (0.0021) entre 'L' y 'N' indica que la distancia entre las distribuciones mixtas de estas dos variables también es pequeña. Esto sugiere que las distribuciones de probabilidad de 'L' y 'N' son muy compatibles y, por lo tanto, podrían ser tratadas conjuntamente en un modelo sin generar grandes pérdidas de información.

- **Relaciones entre otras combinaciones de columnas**: La información mutua entre otras combinaciones de columnas también muestra patrones interesantes. Por ejemplo, la información mutua entre 'L' y 'M' es menor (0.3509), lo que indica una dependencia más débil entre estas variables. Esto sugiere que, aunque existe alguna relación, no es tan fuerte como la observada entre 'L' y 'N'. Además, la divergencia de Kullback-Leibler para esta combinación es mayor (0.1218), lo que refleja una mayor diferencia en sus distribuciones de probabilidad, indicando que las columnas 'L' y 'M' podrían tener comportamientos más independientes en comparación con 'L' y 'N'.

En resumen, el análisis de las métricas de información ha revelado relaciones complejas pero interpretables entre las variables del conjunto de datos. Las variables 'L' y 'N' parecen estar fuertemente relacionadas, mientras que otras combinaciones muestran dependencias más débiles. Estos hallazgos pueden servir como base para una mayor exploración y modelado de los datos, así como para tomar decisiones informadas sobre el tratamiento y la manipulación de estas variables en futuros análisis o proyectos.