ENTROPIA, INFORMAÇÃO MÚTUA E CODIFICAÇÃO DE HUFFMAN

Teoria da informação

Miguel Castela
uc2022212972
Engenharia Informática, FCTUC

Trabalho Prático Nº1
Novembro de 2023

ÍNDICE

1. Introdução e Objetivos

O objetivo deste relatório é apresentar, relacionar e interpretar alguns dos resultados obtidos no código realizado no âmbito do trabalho prático nº. 1 - Entropia, Informação Mútua e Codificação de Huffman em código na Linguagem de Programação Python.

2. Exercícios

2d)

Na tabela 1 são apresentadas as correlações de Pearson (valores aproximados) para as variaveis [Acceleration, Cylinders, Displacement, Horsepower, ModelYear, Weight e MPG]

Tabela 1

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
H(X)	3.496423557	1.590435690	4.874068785	4.583748555	3.690642511	6.040364750	4.835799622

Como o coeficiente de correlação de Pearson indica o grau da correlação (e a direção dessa correlação - se positiva ou negativa) entre duas variáveis, podemos ver que este valor difere significativamente entre os graficos que comparam MPG e as outras variaveis presentes do dataset (figuras 1 a 6).

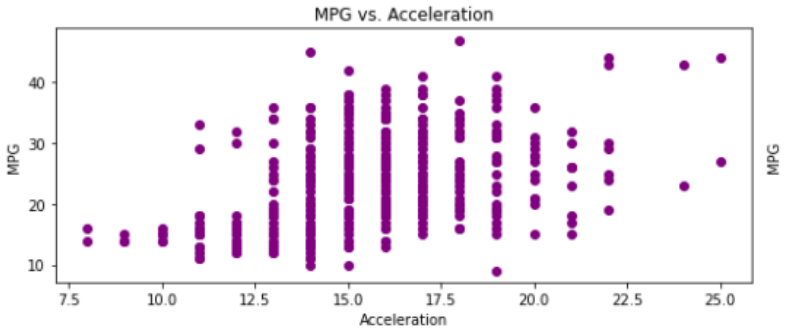

Figura 1 - A aceleração aumenta com MPG, dados dispersos

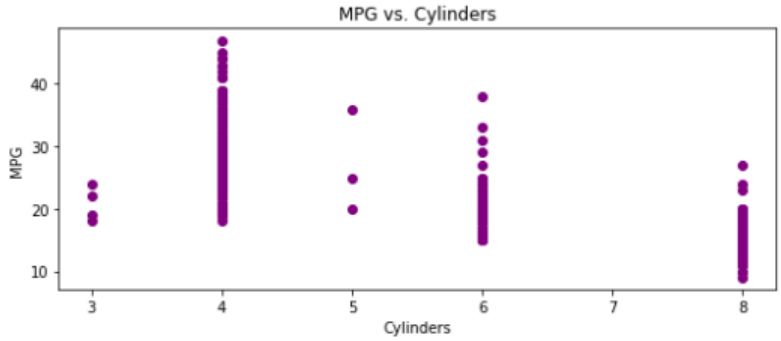

Figura 2 - Cilindros diminuem com MPG, dados dispersos

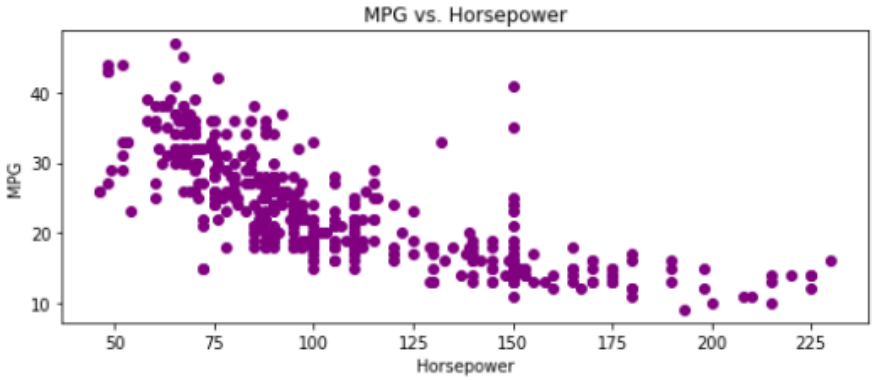

Figura 3 - Horsepower diminuem com MPG, dados menos dispersos

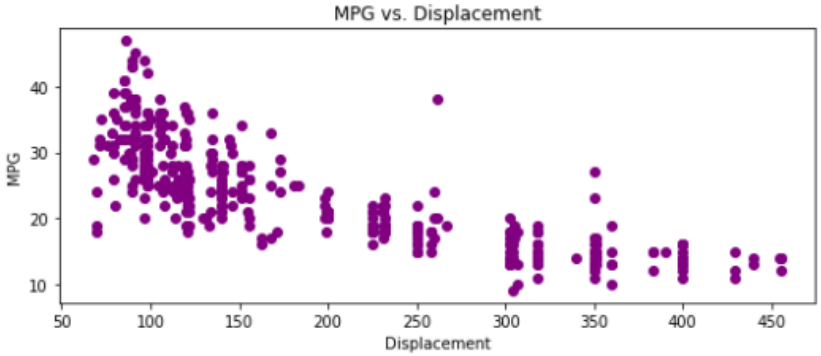

Figura 4 - Displacement diminui com MPG, dados dispersam com o aumento de displacement

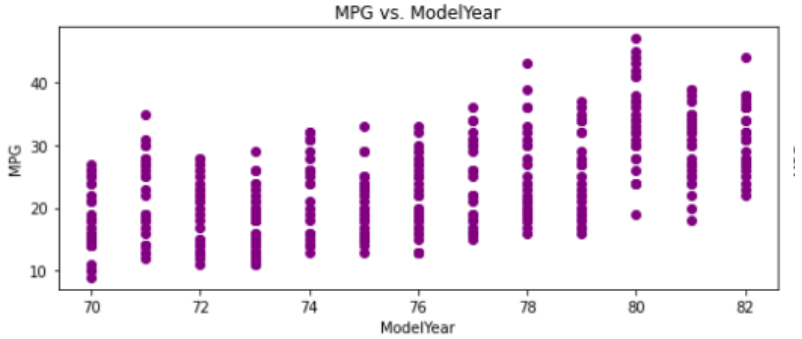

Figura 5 - ModelYear aumenta com MPG, dados dispersos

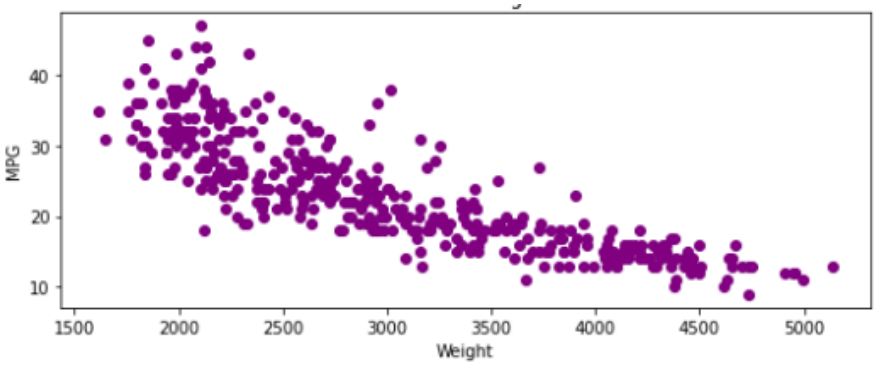

Figura 6 - Peso diminui com MPG, dados menos dispersos

Assim, os resultados indicam que quanto maior for o valor de **MPG**, menor é o valor de **weight, displacement, horsepower e cylinders**. Pelo contrário, aumentando o **MPG** verifica-se um aumento de **acceleration** e **model year**.

Nota: É de notar que os dicionários das ocorrências das variáveis wieght, displacement. Horsepower, a partir de agora, foram alvo de uma função de *binning*, como pedido no exercício 6.

7c)

Para as variaveis [Acceleration, Cylinders, Displacement, Horsepower, ModelYear, Weight e MPG] as entropias são (aproximadamente) as apresentadas na tabela 2.

Tabela 2

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
H(X)	3.496423557	1.590435690	4.874068785	4.583748555	3.690642511	6.040364750	4.835799622

A entropia de cada uma das variáveis é calculada através da fórmula:

$$H(X) = -\sum P(X = xi) \log 2P(X = x_i)$$

A entropia é dada pela medida média de bits por símbolo. Quanto maior o valor da entropia, mais imprevisível é o conjunto de símbolos, o que significa que cada símbolo tem associado mais informação.

Valores baixos de entropia sugerem que o conjunto de símbolos é mais previsível e contém menos informações únicas por símbolo.

Verifica-se que **quanto mais dispersas estão as distribuições de probabilidades, representado nos gráficos, maior é o número médio teórico de bits por símbolo,** o que está de acordo com a definição de entropia.

8b)

A entropia definida usando a codificação de huffmam, para as variaveis [Accelaration, Cylinders, Displacement, Horsepower, ModelYear, Weight e MPG] é a apresentada na tabela 3, que representa a entropia após a codificação de Huffman.

A entropia após a codificação de Huffman, no caso em estudo, apresenta valores superiores à entropia teórica correspondente. Tal como se constata nas tabelas 3 e 4.

Tabela 3

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
L (C,X)	3.535626535	1.729729729	4.911547115	4.614250614	3.727272727	6.076167076	4.86977886

Tabela 4

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
H(X)	3.496423557	1.590435690	4.874068785	4.583748555	3.690642511	6.040364750	4.835799622

Desta forma, verifica-se o Teorema da Codificação de Fonte de Shannon:

$$H(X) \leq L(C, X)$$

A observação de que a codificação de Huffman aumenta ligeiramente a entropia média em vez de a reduzir pode ser devido à natureza da codificação, uma vez que esta introduz códigos diferentes de tamanho variável. Embora isso possa aumentar a entropia média, a vantagem está na redução do tamanho dos dados, economizando espaço de armazenamento.

Os **comprimentos** dos códigos de Huffman para cada um dos símbolos nas respetivas variáveis são os apresentados na tabela 5.

Tabela 5

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
L (X)	0.813841315	0.713194767	1.456549692	1.47036203	0.198347107	0.80746638	0.884762358

Esses comprimentos representam o número médio de bits necessário para representar cada símbolo usando a codificação de Huffman.

Os comprimentos dos códigos de Huffman revelam como os símbolos individuais são representados de maneira eficiente ou ineficiente na codificação. Esses valores são influenciados pela distribuição de probabilidade dos símbolos em cada variável. A eficiência da codificação de Huffman depende de quão bem ela reflete a frequência dos símbolos nas variáveis específicas.

8c)

A redução da variância dos comprimentos dos códigos de Huffman é fundamental para otimizar a eficiência de compressão e o uso de recursos em sistemas de comunicação e armazenamento de dados, evitando que um buffer de comunicação seja maior do que o necessário.

A codificação de Huffman bem projetada deve refletir de forma precisa a distribuição de probabilidade dos símbolos para minimizar a variância dos comprimentos dos códigos.

Para reduzir a variância, pode usar-se uma modificação no algoritmo de construção da árvore de Huffman que prioriza os nós que têm menos folhas. Isto porque códigos com menor variância resultam em uma melhor eficiência de compressão, pois os símbolos mais frequentes são representados com menos bits. Este facto reduz o tamanho total dos dados codificados, economizando espaço de armazenamento e largura de banda de transmissão.

10b)

Os valores dos coeficientes da correlação de Pearson entre a variável MPG e as restantes variáveis são apresentados na tabela 6.

Tabela 6

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
P (X)	0.413585338	-0.77605898	-0.805470113	-0.755171723	0.587263885	-0.831248894	1.0

Os valores da informação mútua entre MPG e as restantes variáveis são apresentados na tabela 7.

Tabela 7

	Acceleration	Cylinders	Displacement	Horsepower	Model Year	Weight	MPG
MI(X)	0.872035837	0.962178641	2.112231080	1.837167729	1.029423662	2.61468365	4.83579962

A informação mútua entre MPG e as outras variáveis é dada pela seguinte expressão:

$$I(Var; MPG) = H(Var) - H(Var|MPG)$$

$$H(Var|MPG)$$

Quanto maior for o valor absoluto do coeficiente de correlação de Pearson, menor será e mais próxima estará a informação mútua de:

$$H(Var)$$

Isso verifica-se nos dados apresentados, pois para valores maiores absolutos da correlação de pearson mais próximo está o valor da informação mútua com a entropia.

11b)

Quando comparado com o valor real de MPG, o erro é de 2.5721872235872243, este é um erro pequeno pois usamos todas as variáveis na formula que estima o MPG

11e)

Quando retiramos da equação que estima o valor de MPG o termo envolvendo a variável que apresenta o maior valor de informação mútua, o erro entre este valor estimado e o real é de 17.154. Porém, quando retiramos o de menor valor de informação mútua, o erro é de 3.0999. Essa discrepância ocorre devido à influência da variável que tem a informação mútua mais alta com MPG, a qual impacta mais significativamente o resultado da estimativa de MPG em comparação com a variável que possui uma informação mútua mais baixa. Isso ocorre porque a informação mútua indica a correlação entre as duas variáveis, de forma que um valor mais alto denota uma correlação mais forte. Assim, ao remover a variável com maior correlação, a estimativa de MPG é mais afetada.

Por sua vez, verifica-se que se obtém um erro menor quando todas as variáveis são usadas. O que, ao utilizar todas as informações mútuas disponíveis, é possível realizar uma estimativa mais eficiente do valor de MPG. Essa abordagem abrange todas as correlações e dependências relevantes entre as variáveis, resultando em uma estimativa mais precisa (tabela 8).

Tabela 8

Com todas as variáveis	Tendo em conta todas as variáveis, exceto aquela com maior informação mútua	Tendo em conta todas as variáveis, exceto aquela com maior informação mútua	Valor Real de MPG
Valor de MPG previsto	Valor de MPG previsto	Valor de MPG previsto	
15.4060 (15)	36.0796 (36)	17.1580 (17)	18
14.2504 (14)	36.0391 (36)	16.0024(16)	15
16.0505 (16)	36.1695 (36)	17.6564(18)	18
15.8628 (16)	35.9818 (36)	17.6148(18)	16
15.8474 (16)	36.1729 (36)	17.4533(17)	17
10.8705 (11)	36.3880 (36)	12.3305(12)	15
10.9825 (11)	36.5000 (37)	12.2965(12)	14
11.1515 (11)	36.4862 (36)	12.4655(12)	14
10.2443 (10)	36.3341 (36)	11.7043(12)	14
13.7536 (14)	36.4687 (36)	15.0676(15)	15

Quando se retira da equação o termo envolvendo a variável que apresentou o menor valor de MI, o erro de precisão aumenta ligeiramente, para 3.0999169533169537. Por outro lado, ao retirar o membro envolvendo a variável com maior MI, o erro de precisão aumenta significativamente para 17.154129975429978

3. Conclusões

Após concluir os objetivos deste projeto, chegou-se à conclusão de que o valor médio teórico de bits por símbolo desempenha um papel fundamental na compressão de dados. Além disso, constatou-se que a codificação de Huffman se aproxima significativamente do limite teórico da média de bits por símbolo, o que é altamente relevante no contexto da compressão de dados. Identificou-se também uma forte correlação entre os coeficientes de Pearson e a informação mútua de variáveis, o que destaca a importância desses coeficientes na análise estatística. Por fim, durante o processo de estimativa de MPG, observou-se que algumas variáveis possuem mais informação mútua do que outras, o que pode influenciar significativamente a qualidade das estimativas.

Neste projeto executaram-se conteúdos lecionados nas aulas praticas, como por exemplo a função de binning, o que melhorou a compreensão dos conceitos e formulas de teoria de informação