1290 FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE D COIMBRA

dei departamento de engenharia informática

RELATÓRIO

TP1 – Entropia, Informação Mútua e Codificação de Huffman

Unidade Curricular:

Teoria da Informação

Data:

03/11/2023

João Nuno Coelho
No 2021275030

Luana Carolina Reis
No 2022220606

Rodrigo Barata
No 2022212664

INTRODUÇÃO

Através do presente trabalho prático, pretende-se efetuar o estudo da relação de dependência entre determinadas características de um carro e seu respetivo rendimento de combustível (MPG ou *Miles Per Gallon*), para 407 modelos de carros diferentes.

Obtenção dos dados

Os dados necessários à elaboração deste trabalho são conseguidos através da utilização da biblioteca *pandas*, que facilita a análise de dados tabulares.

	A	B	C	D	E	F	G
1	Acceleration	Cylinders	Displacement	Horsepower	ModelYear	Weight	MPG
2	12	8	307	130	70	3504	18
3	12	8	350	165	70	3693	15
4	11	8	318	150	70	3436	18
5	12	8	304	150	70	3433	16
6	11	8	302	140	70	3449	17
7	10	8	429	198	70	4341	15
8	9	8	454	220	70	4354	14
9	9	8	440	215	70	4312	14
10	10	8	455	225	70	4425	14
11	9	8	390	190	70	3850	15
12	18	4	133	115	70	3090	19
13	12	8	350	165	70	4142	15
14	11	8	351	153	70	4034	15
15	11	8	383	175	70	4166	15

Fig. 1 – Dados fornecidos através do ficheiro Excel "CarDataset.xlsx"

Outras bibliotecas

Para além desta, é ainda fundamental a utilização da biblioteca *numpy*, capaz de suportar *arrays* multidimensionais e funções matemáticas que as manipulam de um modo eficiente. Utiliza-se também a biblioteca *matplotlib*, para efeitos de visualização de dados, com recurso a gráficos e *plots*.

RELAÇÃO MPG VS. VAR

2. d) Comente a relação de MPG com as restantes variáveis.

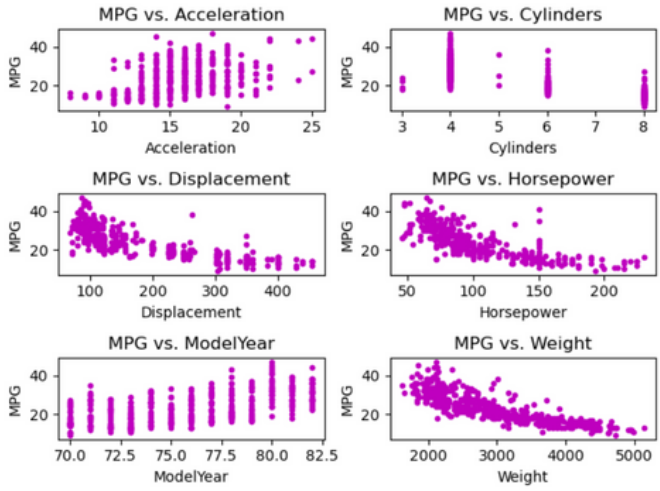


Fig. 2- Gráficos MPG vs. VAR

MPG vs. *Acceleration*

Há uma correlação positiva fraca entre ambas as variáveis. Isto sugere que quanto maior for a aceleração do carro (+*Acceleration*), menor será o consumo de combustível (+MPG). Isto pode ser explicado por uma maior resistência do ar ou maior potência do motor. Apesar desta relação existir, a dispersão de valores no gráfico é considerável, existindo alguns valores anormais, tornando esta relação inconsistente.

MPG vs. *Cylinders*

Há uma forte correlação negativa entre os cilindros do carro e o seu rendimento. Isto significa que carros com mais cilindros (+*Cylinders*) tendem a ter menores valores de rendimento (-MPG), pelo que consomem mais. Em geral, precisam de queimar mais combustível para manter o motor a funcionar, mesmo quando não precisam de toda a potência disponível.

MPG vs. *Displacement*

Há uma forte correlação negativa entre a cilindrada do carro e o seu rendimento. Logo, carros com maior tamanho de motor (+*Displacement*), apesar de terem uma maior capacidade de gerar força, tendem a ter um pior rendimento (-MPG) e, portanto, consomem mais combustível.

MPG vs. *Horsepower*

Há uma forte correlação negativa entre a potência de um carro e o seu rendimento. Quanto maiores forem os cavalos de potência (+*Horsepower*), menor será o rendimento (-MPG), devido à maior demanda de energia.

MPG vs. *ModelYear*

Há uma correlação positiva moderada entre o ano do modelo de um carro e o seu rendimento, dado que modelos mais recentes tendem a ter uma maior eficiência energética. Quanto mais recente for (+*ModelYear*), maior será o seu rendimento (+MPG) e menor o seu consumo. No entanto, esta relação não é muito forte.

MPG vs. *Weight*

Há uma forte correlação negativa entre o peso de um carro e o seu rendimento. Assim, carros mais pesados (+*Weight*) têm um menor rendimento (-MPG), logo, um maior consumo. Isto porque será necessária uma força maior para mover o carro, pois tem maior inércia.

VALOR MÉDIO DE BITS

7. c) Comente os resultados obtidos através do cálculo do valor médio (teórico) de bits por símbolo.

```
Valor médio de bits (teórico) por símbolo >> antes de binning

Acceleration:       3.496 bits/símbolo
   Cylinders:       1.590 bits/símbolo
Displacement:       5.731 bits/símbolo
  Horsepower:       5.842 bits/símbolo
   ModelYear:       3.691 bits/símbolo
      Weight:       8.394 bits/símbolo
         MPG:       4.836 bits/símbolo
         ALL:       7.212 bits/símbolo
```

```
Valor médio de bits (teórico) por símbolo >> depois de binning

Acceleration:       3.496 bits/símbolo
   Cylinders:       1.590 bits/símbolo
Displacement:       4.841 bits/símbolo
  Horsepower:       4.537 bits/símbolo
   ModelYear:       3.691 bits/símbolo
      Weight:       6.061 bits/símbolo
         MPG:       4.836 bits/símbolo
         ALL:       6.667 bits/símbolo
```

Fig. 3- Resultados obtidos no cálculo do valor médio (teórico) de bits por símbolo

O valor médio (teórico) de bits por símbolo representa uma medida da eficiência da codificação de um modelo. Depende do número de valores possíveis e da probabilidade de ocorrência de cada um desses valores.

Quanto maior for o número de valores possíveis, maior será o valor médio de bits necessários para representar o conjunto. Além disso, quanto maior a probabilidade de ocorrência de um certo valor, menor será o valor médio de bits necessários para o representar.

Este valor pode ainda ser reduzido através do processo de *binning*, uma técnica de pré-processamento de dados que consiste num agrupamento dos valores originais em intervalos menores e na substituição pelo valor mais representativo de cada um desses intervalos. Este processo é capaz de reduzir os efeitos de ruído nos dados, eliminando variações insignificantes. Apesar dos seus efeitos positivos, nalguns casos, pode reduzir a precisão, aumentando o erro padrão ou a variância dos comprimentos (regra geral, diminui a variância).

Variáveis	*Antes de binning*	Depois de *binning*
Weight	8.394 bits/símbolo	6.061 bits/símbolo
Displacement	5.731 bits/símbolo	4.841 bits/símbolo
Horsepower	5.842 bits/símbolo	4.537 bits/símbolo
ALL	7.212 bits/símbolo	6.667 bits/símbolo

Fig. 4- Valor médio (teórico) de bits por símbolo, antes e depois de *binning*

HUFFMAN VS. TEÓRICO

8. b) Compare os resultados obtidos através dos códigos de Huffman com os obtidos no ponto 7 (valor teórico), e comente as suas observações.

c) Como se pode reduzir a variância dos comprimentos, e qual é a importância disto?

```
Valor médio de bits (codificação de Huffman) por símbolo >> antes de binning

Acceleration:       3.536 bits/símbolo      0.814 variância dos comprimentos
   Cylinders:       1.730 bits/símbolo      0.713 variância dos comprimentos
Displacement:       5.764 bits/símbolo      1.758 variância dos comprimentos
  Horsepower:       5.872 bits/símbolo      2.190 variância dos comprimentos
   ModelYear:       3.727 bits/símbolo      0.198 variância dos comprimentos
      Weight:       8.464 bits/símbolo      0.440 variância dos comprimentos
         MPG:       4.870 bits/símbolo      0.885 variância dos comprimentos
         ALL:       7.247 bits/símbolo      4.929 variância dos comprimentos
```

```
Valor médio de bits (codificação de Huffman) por símbolo >> depois de binning

Acceleration:       3.536 bits/símbolo      0.814 variância dos comprimentos
   Cylinders:       1.730 bits/símbolo      0.713 variância dos comprimentos
Displacement:       4.872 bits/símbolo      1.040 variância dos comprimentos
  Horsepower:       4.560 bits/símbolo      1.264 variância dos comprimentos
   ModelYear:       3.727 bits/símbolo      0.198 variância dos comprimentos
      Weight:       6.088 bits/símbolo      0.783 variância dos comprimentos
         MPG:       4.870 bits/símbolo      0.885 variância dos comprimentos
         ALL:       6.697 bits/símbolo      2.378 variância dos comprimentos
```

Fig. 5- Resultados obtidos no cálculo do valor médio de bits por símbolo (codificação de Huffman)

Variáveis	*Antes de binning*	Depois de *binning*
Weight	Variância: 0.440	Variância: 0.783
Displacement	Variância: 1.758	Variância: 1.040
Horsepower	Variância: 2.190	Variância: 1.264
ALL	Variância: 4.929	Variância: 2.378

Fig. 6- Redução da variância dos comprimentos, após *binning* (exceto em *Weight*)

A codificação de Huffman é um dos métodos de compressão de dados existentes e utiliza a probabilidade de ocorrência de cada símbolo para determinar o tamanho do seu código. Por este motivo, os códigos têm um tamanho variável.

Face aos resultados obtidos em ambas as alíneas, verifica-se um claro aumento dos valores com códigos de Huffman, em comparação ao teorizado. Embora mínima, esta diferença deve-se ao facto dos valores teorizados representarem uma codificação ótima dos resultados. Há uma certa discrepância entre as probabilidades reais e as estimadas dos símbolos, que nem sempre são conhecidas com a máxima precisão. É esta incerteza que provoca um aumento do número de bits.

É possível reduzir a variância dos comprimentos, ao fazer com que a probabilidade de ocorrência de cada valor seja o mais equiprovável possível a todas as outras e ao ajustar o número de bits usados para cada valor, de forma a ser o menor possível, sem perder informação. Estas alterações permitem diminuir a incerteza em relação ao tamanho em bits dos valores seguintes, o que se traduz numa redução do tempo necessário para ler todos os dados. Para além disso, simplifica todas as operações realizadas sobre os mesmos.

PEARSON VS. MI

10.b) Comparar os resultados obtidos através do cálculo da Informação Mútua (MI) com os coeficientes de correlação de Pearson.

Variáveis	Coeficientes de Pearson	Informação Mútua (MI)
MPG e *Acceleration*	0.414	0.872
MPG e *Cylinders*	-0.776	0.962
MPG e *Displacement*	-0.805	2.036
MPG e *Horsepower*	-0.755	1.798
MPG e *ModelYear*	0.587	1.029
MPG e *Weight*	-0.832	2.631

Fig. 7- Resultados obtidos no cálculo dos coeficientes de correlação de Pearson e no cálculo da Informação Mútua (MI)

Correlação muito forte > 0.9 / < -0.9

Correlação forte 0.7 a 0.9 / -0.9 a -0.7

Correlação moderada 0.5 a 0.7 / -0.7 a -0.5

Correlação fraca 0.3 a 0.5 / -0.5 a -0.3

Correlação desprezível 0.0 a 0.3 / -0.3 a 0.0

Tal como analisado no ponto 2. d), podemos observar que os valores de MPG sugerem forte dependência em relação aos resultados de *Weight*, *Horsepower*, *Displacement* e *Cylinders*, um dependência moderada dos resultados de *ModelYear* e uma fraca dependência de *Acceleration*.

A informação mútua (MI) é uma medida da dependência entre duas variáveis aleatórias, útil para quantificar a informação que uma variável contém sobre outra. Segundo os dados, *Weight, Displacement e Horsepower* apresentam um maior valor de informação mútua (MI), em comparação com as restantes variáveis .

Assim, e de acordo com a análise gráfica inicial, podemos confirmar que estas são as variáveis que mais influenciam os valores de MPG.

Além disso, é razoável presumir que os maiores valores de MPG serão atingidos com valores baixos destas três, uma vez que todos os seus coeficientes são negativos.

MPG REAL VS. PREVISTO

11. b) Comparar o resultado obtido através do cálculo do 'MPG' estimado com os valores reais de 'MPG'.

e) Comparar os resultados das métricas de erro MAE ("Mean Absolute Error") e MSE ("Mean Squared Error"), antes da remoção de qualquer termo da expressão MPGpred, ao remover o termo da variável com o menor valor de MI e ao remover o termo da variável com maior valor de MI.

Amostra	MPG Real	MPG Previsto
Índice 000	18	15,411
Índice 023	21	21,144
Índice 402	44	31,458

Fig. 8- Resultados de MPG Real vs. MPG Previsto

	MAE	MSE
Sem remoções	2.571	12.034
Remoção de termo com -MI	3.076	15.010
Remoção de termo com +MI	17.153	332.065

Fig. 9- Resultados obtidos no cálculo das métricas de erro MAE e MSE

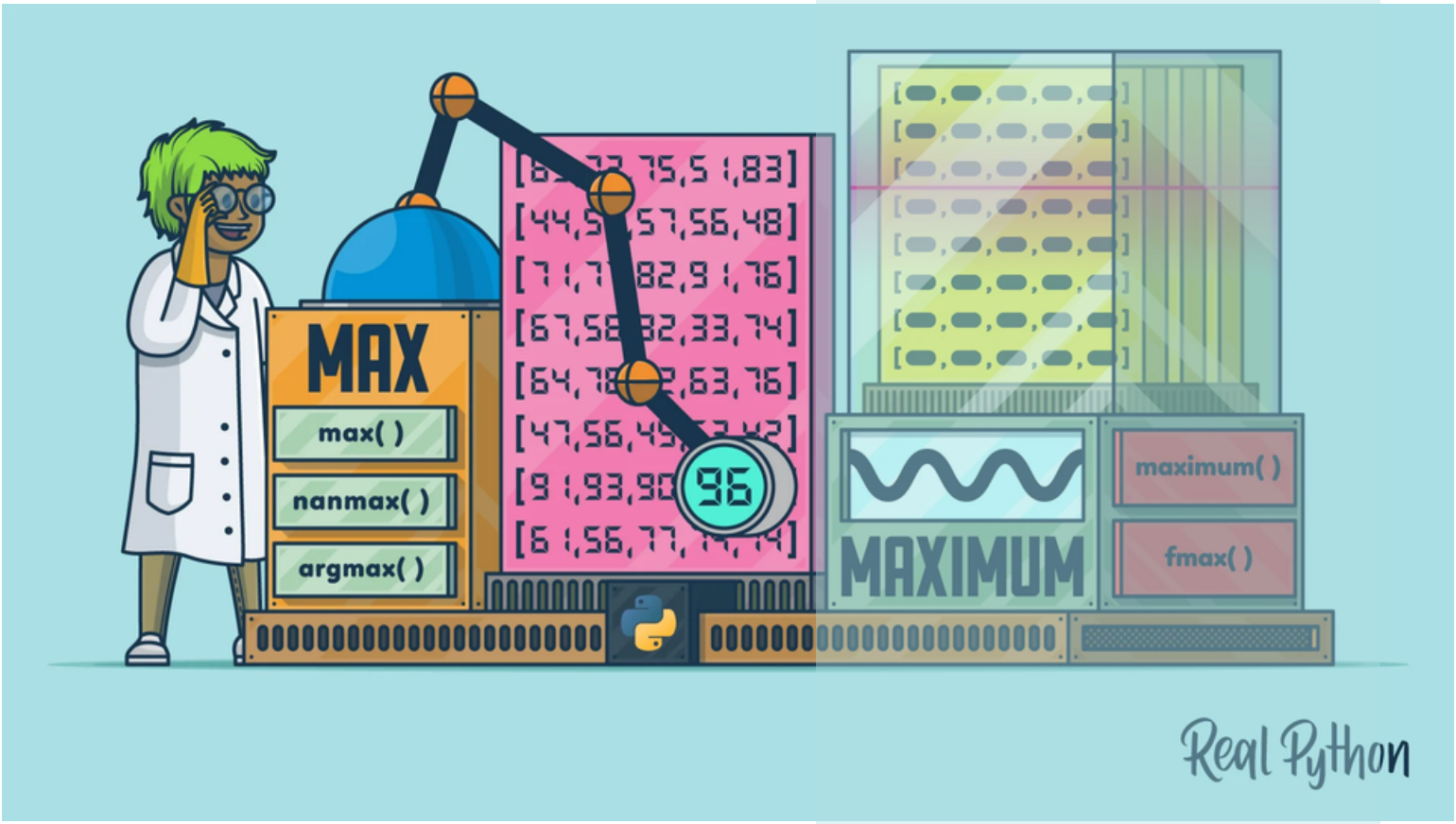
$$MPG_{pred} = -5.5241 - 0.146 * Acceleration - 0.4909 * Cylinders + 0.0026 * Distance - 0.0045 * Horsepower + 0.6725 * Model - 0.0059 * Weight$$

Fig. 10- Relação para a estimativa de MPG

As métricas de erro MAE (*Mean Absolute Error*) e MSE (*Mean Squared Error*) são utilizadas para avaliar o desempenho de modelos de regressão e medem a média da diferença entre os valores previstos e os observados. A utilidade da utilização do MSE é a sua sensibilidade a erros maiores, relativamente à métrica MAE. Um baixo valor destas métricas indica uma boa capacidade de previsão do modelo.

Retirar da equação o termo da variável que apresentou o menor valor de IM (*Acceleration*) reduziu ligeiramente o desempenho do modelo (aumentando as métricas de erro), uma vez que esta auxilia de algum modo esta previsão, ainda que a sua influência para o resultado final não seja tão grande quanto a das restantes variáveis.

Por outro lado, a remoção do termo da variável que apresentou o maior valor de IM (*Weight*) reduziu significativamente o desempenho do modelo (aumentando exageradamente as métricas de erro), pois esta tem grande influência na precisão da previsão do mesmo.



WEBGRAFIA

[1] Matplotlib, *Matplotlib 3.8.1 documentation*. [online]. Disponível em: https://matplotlib.org/stable/index.html. [Consultado em 02/10/2023].

[2] NumPy, *NumPy reference*. [online]. Disponível em: https://numpy.org/doc/stable/reference/. [Consultado em 02/10/2023].

[3] Pandas, *API reference*. [online]. Disponível em: https://pandas.pydata.org/docs/reference/index.html. [Consultado em 02/10/2023].

[4] Real Python, *NumPy's max() and maximum(): Find Extreme Values in Arrays*. [online]. Disponível em: https://realpython.com/numpy-max-maximum/. [Consultado em 03/11/2023].

[5] Wikipédia, *Coeficiente de correlação de Pearson*. [online]. Disponível em: https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson. [Consultado em 30/10/2023].