FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Teoria da Informação

Licenciatura em Engenharia Informática

2023

Relatório | Trabalho Prático – 1

Entropia,

Informação Mútua

e Codificação de Huffman

Trabalho realizado por:

Eduardo Marques - 2022231584

João Cardoso - 2022222301

Sérgio Marques - 2022222096

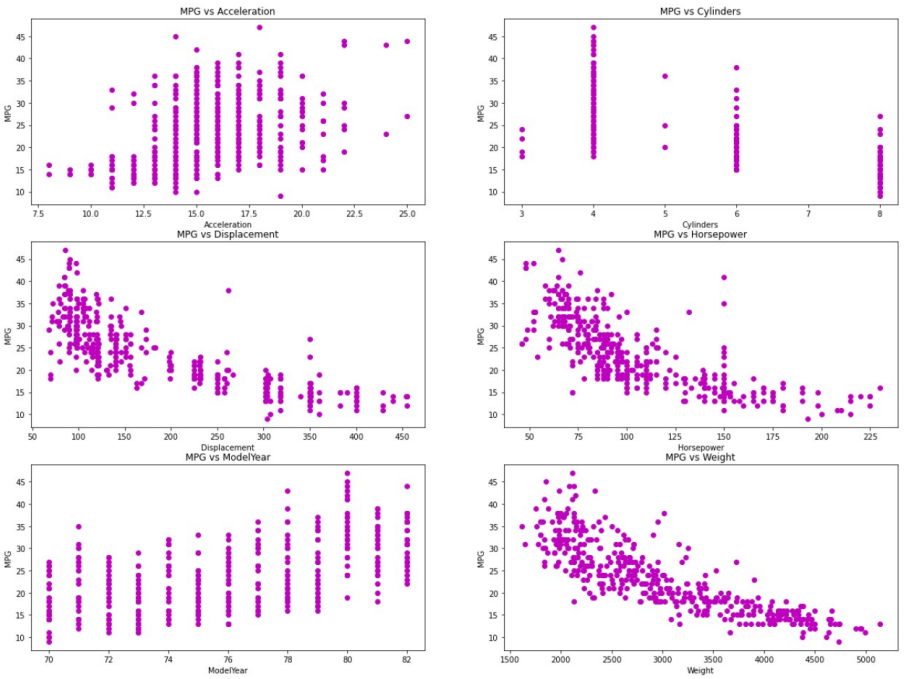
Índice

Exercício 1

Lemos o ficheiro Excel utilizando o módulo que nos foi fornecido e criamos duas funções “*matrizDados*” e “*listaVariaveis*” que devolvem, respetivamente uma matriz que contém todos os dados da tabela e uma lista com os nomes das variáveis contidas na mesma, que serão úteis à realização do trabalho.

Exercício 2

Criamos uma nova função “*compararMPG*” onde são criados os gráficos que relacionam os valores de cada variável com os respetivos valores de “MPG”, que se encontram na matriz de dados. Para cada gráfico, a variável que pretendemos estudar será definida no eixo x enquanto que os valores de “MPG” irão ocupar o eixo y. Os gráficos são apresentados todos na mesma figura, como podemos ver na imagem abaixo, o que nos permite comparar os gráficos de cada variável de uma forma mais acertada.



d) Relação de MPG com as restantes variáveis

Através dos gráficos, chegamos à conclusão de que os valores de MPG diminuem com o aumento dos valores do Displacement, do Horsepower e do Weight. As restantes variáveis apresentam gráficos bastante irregulares, sendo que os valores de MPG aumentam com a Acceleration e com o ModelYear. A relação entre MPG e Cylinders apresenta uma enorme irregularidade.

Exercício 3

É criado um alfabeto que contém todos os valores encontrados para cada variável na matriz, convertidos para números inteiros de 16 bits, garantindo que não existem valores duplicados e abrangem os valores entre o menor e o maior valor encontrados.

Exercício 4

Neste exercício pretende-se contar as ocorrências para cada símbolo de cada variável, ou seja, contar o número de vezes que cada símbolo aparece na coluna da matriz destinada a cada variável. É criado um dicionário cujas chaves correspondem ao nome de cada variável. Os valores associados a essas chaves, são outros dicionários que guardam os símbolos e o número de ocorrências associadas a esses mesmos símbolos.

Exercício 5

Para a realização deste exercício, é utilizado o dicionário que foi criado no exercício anterior. A partir desse dicionário, criamos um outro dicionário, “ocorrenciasNaoNulas”, que guarda somente os conjuntos ”simbolo-ocorrencias”, cujo valor das ocorrências é não nulo. Utilizando este novo dicionário, são criados gráficos de barras para cada variável, representando no eixo x todos os símbolos dessa mesma variável e, no eixo y, o número de ocorrências associado ao respetivo símbolo.

Exercício 6

Neste ponto, era-nos pedido que fizéssemos o agrupamento de símbolos para as variáveis Displacement, Horsepower e Weight, com o intuito de melhorar a visualização dos dados nos gráficos de barras e de tornar mais fiável o cálculo de informação mútua. Para isso, inicializamos um novo dicionário, “novasOcorrencias” onde serão guardados os novos valores de ocorrências de cada símbolo.

De modo a realizar o agrupamento de símbolos para as variáveis desejadas, colocamos as chaves (símbolos) e os valores (ocorrências) do dicionário “ocorrencias” em listas devidamente adequadas. De seguida, percorremos as duas listas de n em n elementos e, para cada conjunto de n elementos, encontramos o símbolo com o maior valor de ocorrências. Às ocorrências desse símbolo, somamos todos os valores de ocorrências dos outros símbolos presentes no intervalo e, adicionamos esse mesmo símbolo e o seu novo valor de ocorrências ao novo dicionário

Para as variáveis que não são sujeitas ao agrupamento, copiamos os dados que se encontram no dicionário “ocorrencias” para o novo dicionário, uma vez que os valores destas variáveis não sofrem qualquer alteração.

Exercício 7

De modo a calcular o número médio de bits por símbolo para cada variável, é calculada a probabilidade de um símbolo ocorrer para essa mesma variável. A partir das probabilidades obtidas, iremos utilizar o cálculo da entropia para calcular o número médio de bits. Posteriormente, será feito o mesmo para o conjunto completo de dados.

```
== CALCULO BITS ==
Valor médio de bits ( Acceleration ):  3.496423557860517
Valor médio de bits ( Cylinders ):  1.5904356903532713
Valor médio de bits ( Displacement ):  5.731068442748928
Valor médio de bits ( Horsepower ):  5.841551829284073
Valor médio de bits ( ModelYear ):  3.6906425111722605
Valor médio de bits ( Weight ):  8.394254916672104
Valor médio de bits ( MPG ):  4.835799622324453
Valor médio de bits (conjunto completo de dados):  7.604523003545578
```

c) Resultados valor médio de bits por símbolo

Como podemos observar pelos dados da imagem acima, os valores obtidos para o número médio de bits por símbolo são os expectáveis. Isto porque os valores mais baixos do número médio de bits por símbolo correspondem a variáveis cujos símbolos com ocorrências são valores pequenos. O mesmo acontece para as variáveis cujos símbolos com ocorrências são valores maiores - o seu valor médio de bits por símbolo é, também , maior.

Esta relação é percetível quando olhamos para as variáveis Cylinders e Weight. No caso da variável Cylinders, os valores dos símbolos com ocorrências são muito baixos (estão compreendidos entre 3 e 8). Logo, o seu valor médio de bits por símbolo é muito baixo, sendo, até, o menor valor de todas as variáveis. Os valores dos símbolos com ocorrências para a variável Weight são muito elevados (1613 a 5140), o que faz com que o número médio de bits por símbolo para esta variável seja o maior de todas as variáveis em análise.

Exercício 8

Neste exercício, a partir do módulo que nos foi fornecido no enunciado, foi calculado o valor médio de bits por símbolo utilizando codificação de Huffman. Primeiro, foram calculadas as probabilidades das ocorrências de cada símbolo para uma determinada variável. Posteriormente, calculamos a média dos comprimentos obtidos para cada símbolo, tendo, cada comprimento, o peso da probabilidade que lhe corresponde, calculada anteriormente.

b) Comparação com os valores obtidos no ponto 7

```
== HUFFMAN CODE ==
Valor medio de bits por simbolo ( Acceleration ):  3.535626535626536
Valor medio de bits por simbolo ( Cylinders ):  1.7297297297297298
Valor medio de bits por simbolo ( Displacement ):  5.764127764127764
Valor medio de bits por simbolo ( Horsepower ):  5.87223587223587
Valor medio de bits por simbolo ( ModelYear ):  3.7272727272727284
Valor medio de bits por simbolo ( Weight ):  8.464373464373462
Valor medio de bits por simbolo ( MPG ):  4.8697788697788695
```

```
== CALCULO BITS ==
Valor médio de bits ( Acceleration ):  3.496423557860517
Valor médio de bits ( Cylinders ):  1.5904356903532713
Valor médio de bits ( Displacement ):  5.731068442748928
Valor médio de bits ( Horsepower ):  5.841551829284073
Valor médio de bits ( ModelYear ):  3.6906425111722605
Valor médio de bits ( Weight ):  8.394254916672104
Valor médio de bits ( MPG ):  4.835799622324453
```

Como podemos perceber pelas imagens acima apresentadas, os valores obtidos para o número médio de bits por símbolo são muito idênticos (relativamente maiores) aos que foram obtidos no exercício anterior, onde foi usada a fórmula da entropia. Isto permite-nos perceber que o uso desta fórmula é fiável para o cálculo do número médio de bits por símbolo, uma vez que os valores obtidos utilizando esta fórmula pouco diferem daqueles que foram obtidos utilizando codificação de Huffman, um método mais exato.

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

Para reduzir a variância dos comprimentos dos códigos Huffman pode-se fazer o agrupamento dos valores que ocorrem com menos frequência (ou dos valores das variáveis com muitos valores diferentes), que é algo que já foi realizado. A função “binning”, que agrupa os símbolos de algumas variáveis, foi implementada com o intuito de melhorar a visualização dos gráficos e o cálculo da informação mútua. Porém, se tivessem sido usados, para o cálculo da variância dos comprimentos, os dados pós-agrupamento iríamos obter, certamente, valores inferiores. A redução da variância dos comprimentos dos códigos Huffman está relacionada com a eficiência da compressão. Isto é, quanto menor a variância, mais uniformes serão os comprimentos dos códigos. Além disso, reduzir a variância leva a uma melhor representação dos dados.

Exercício 9

Para a realização deste exercício, foi criada a função “corrPearson” que, utilizando a função corrcoef do Numpy indicada no enunciado, calcula o coeficiente de correlação de Pearson entre a variável MPG e as restantes variáveis.

Exercício 10

Neste exercício, pretendia-se calcular o valor de informação mútua entre a variável MPG e as restantes variáveis em estudo. Para isso, em cada cálculo de IM, calculamos os valores de entropia de MPG e da variável em questão, bem como o valor da entropia conjunta das duas variáveis.

Para calcular a entropia de MPG e das variáveis restantes, utilizamos o mesmo princípio que foi usado no exercício 7, quando, a partir da fórmula da entropia, calculamos o valor médio de bits por símbolo para cada variável.

Para o cálculo da entropia conjunta, fizemos um agrupamento dos valores das ocorrências da variável MPG e da variável em análise, criando um dicionário cujas chaves são os conjuntos obtidos e os valores são o número de vezes que esses conjuntos ocorrem. Por exemplo, se o valor da primeira ocorrência de MPG fosse igual a 1 e o valor da primeira ocorrência de Acceleration fosse igual a 2, o conjunto seria (1, 2) e o seu valor seria igual a 1. É, depois, a partir destes valores que são calculadas as probabilidades que são usadas no cálculo da entropia conjunta.

Por último, já com os valores da entropia das duas variáveis e da entropia conjunta calculados, efetuamos o cálculo do calor de informação mútua (entropia de MPG + entropia da variável - entropia conjunta).

Nota: Para as variáveis que sofreram agrupamento, deveríamos ter utilizado os dados após o agrupamento, o que não foi possível para o cálculo da entropia conjunta, uma vez que usamos os dados anteriores ao agrupamento. Isto acontece porque quando é feito o agrupamento, não alteramos os dados iniciais das variáveis em questão e criamos um dicionário, “novasOcorrencias”, que guarda os novos valores já com o valor das ocorrências atualizado (não realizando, primeiro, a substituição de todos os símbolos de um conjunto pelo símbolo desse conjunto cujas ocorrências têm o maior valor). Apesar de, para estas variáveis, o valor obtido ser relativamente diferente do valor real, sabemos que a ordem dos valores de informação mútua não se irá alterar.

```
== CORRELAÇÃO PEARSON ==
Coeficiente de correlação de Pearson entre MPG e Acceleration :  0.4135853380757749
Coeficiente de correlação de Pearson entre MPG e Cylinders :  -0.7760589899625312
Coeficiente de correlação de Pearson entre MPG e Displacement :  -0.8047025675071012
Coeficiente de correlação de Pearson entre MPG e Horsepower :  -0.7551351418256332
Coeficiente de correlação de Pearson entre MPG e ModelYear :  0.5872638852454327
Coeficiente de correlação de Pearson entre MPG e Weight :  -0.8321486403128854


== INFORMAÇÃO MUTUA ==
IM entre MPG e Acceleration :  0.8720358370364583
IM entre MPG e Cylinders :  0.9621786410869246
IM entre MPG e Displacement :  1.749206746964112
IM entre MPG e Horsepower :  1.2388172607813939
IM entre MPG e ModelYear :  1.029423662315513
IM entre MPG e Weight :  1.8044939328810763
```

b) Comparação com os resultados obtidos no ponto 9

Observando a imagem que se encontra na página anterior, é possível perceber que os resultados obtidos para a correlação de Pearson e para a informação mútua estão em sintonia e parecem estar corretos. Para a correlação de Pearson valores próximos de 1 e -1 indicam uma correlação forte, enquanto que valores próximos de 0 revelam que a correlação é fraca (ou praticamente inexistente). Na informação mútua, quanto maior é o valor, maior é a dependência que uma variável tem da outra.

Basta olharmos para o valor mais baixo e o valor mais alto (em módulo) dos coeficientes de relação de Pearson, para perceber que estes dois conceitos estão interligados. Para a variável Acceleration, o coeficiente de correlação de Pearson é aquele que está mais próximo de 0 de entre todas as variáveis. Ora, se revela uma correlação fraca com a variável MPG, a dependência de MPG da variável Acceleration deve ser baixa. Analisando os resultados de informação mútua podemos perceber que é isso que acontece, sendo o valor da variável Acceleration o maior valor de informação mútua. Da mesma maneira, na variável Weight, a forte correlação traduz-se numa alta dependência de MPG desta variável.

Exercício 11

Neste ponto, fizemos uma estimativa dos valores de MPG, a partir dos valores das restantes variáveis, utilizando uma equação que nos foi fornecida no enunciado. É isso que realizamos na função “preverMPG”. Primeiro, criamos um array onde serão armazenados os valores estimados de MPG e, de seguida, percorremos a matriz que contém os valores de todas as variáveis e aplicamos a equação aos devidos valores, obtendo, assim, uma estimativa para os valores de MPG.

Posteriormente, era-nos solicitado que realizássemos o mesmo procedimento tirando da equação, numa primeira abordagem, o menor valor de informação mútua e, de seguida, o maior valor de informação mútua. Uma vez que tínhamos os valores de informação mútua guardados no dicionário “valoresIM”, pudemos comparar os valores e perceber quais representavam o menor e o maior valor. Sabendo qual era o valor que deveria ser retirado (e retirando-o) da equação em cada caso, repetimos o procedimento anterior.

b) Comparação dos resultados com os valores verdadeiros de MPG

Comparando, simplesmente, os valores estimados de MPG com os seus valores reais, é possível perceber que os valores estimados estão relativamente próximos dos valores reais, sendo um pouco inferiores. Numa análise mais detalhada aos resultados obtidos, onde realizámos o cálculo do erro percentual dos valores estimados de MPG, obtivemos um valor de 11,7%. Apesar de não ser um valor assim tão elevado, permitiu-nos perceber que este procedimento tem as suas debilidades e não preza muito pela exatidão.

e) Comparação com os resultados obtidos nos pontos a, c e d

```
Valores estimados de MPG:  [15. 14. 15. 15. 15. 10. 10. 11. 10. 13. 18. 11. 12. 11.
13. 15. 15. 16.
 14. 18. 23. 19. 19. 21. 24. 25. 21. 22. 22. 24. 20.  8.  9.  9.  6. 25.
 24. 24. 24. 25. 21. 16. 17. 17. 17. 11. 10. 11. 12.  7.  8.  6. 19. 23.
 17. 18. 25. 25. 24. 25. 26. 27. 26. 25. 25. 25. 24. 23. 25. 12. 11. 12.
 12. 15. 10. 10. 10. 11. 25. 14. 12. 11. 13. 21. 23. 20. 25. 24. 24. 23.
```

```
Valores estimados de MPG (sem menor valor de IM):  [17. 16. 17. 17. 17. 12. 12. 12.
11. 15. 21. 13. 14. 13. 15. 16. 16. 17.
 15. 19. 25. 21. 22. 23. 26. 28. 23. 25. 25. 26. 23. 10. 11. 11.  9. 27.
 26. 26. 27. 28. 23. 19. 19. 20. 20. 13. 12. 14. 14.  9. 10.  8. 21. 26.
 20. 21. 27. 27. 27. 27. 29. 30. 29. 28. 27. 28. 27. 26. 27. 13. 13. 14.
 14. 17. 11. 12. 12. 13. 27. 16. 15. 13. 15. 23. 26. 23. 27. 26. 27. 26.
```

```
Valores estimados de MPG (sem maior valor de IM):  [36. 36. 36. 35. 36. 36. 36. 36.
36. 36. 36. 36. 36. 36. 36. 36. 36. 36.
 36. 36. 37. 36. 36. 36. 37. 36. 36. 37. 36. 37. 36. 35. 35. 35. 34. 37.
 37. 38. 37. 37. 37. 37. 37. 37. 37. 36. 36. 36. 36. 36. 36. 36. 37. 37.
 37. 37. 38. 38. 37. 37. 37. 37. 37. 37. 38. 38. 37. 37. 38. 37. 37. 37.
 37. 37. 37. 37. 37. 37. 39. 37. 37. 36. 37. 38. 38. 37. 38. 38. 38. 38.
```

Na primeira estimativa realizada aos valores de MPG é possível perceber, como já foi referido anteriormente, que os valores estimados não estão muito afastados dos valores reais.

Ao remover da equação a variável com menor valor de informação mútua, é expectável que os valores estimados estejam muito próximos dos valores reais, uma vez que os valores da variável MPG estão muito pouco dependentes dos valores desta variável. Observando a segunda imagem acima, conseguimos perceber que os valores obtidos são idênticos às primeiras estimativas de MPG e estão ainda mais próximos dos valores reais de MPG.

Quando removemos da equação a variável com o maior valor de informação mútua, os resultados da estimativa prevêem-se diferentes. Analisando os valores da última imagem, conseguimos perceber que os valores estimados são descabidos e estão muito afastados dos valores reais e dos restantes valores estimados anteriormente. Isto deve-se ao facto de os valores da variável MPG dependerem bastante dos valores desta variável, já que é a variável que tem o maior valor de informação mútua.

Podemos, então, concluir que os resultados obtidos nestas estimativas são os esperados, de acordo com aquilo que era pretendido analisar.