Ficha 5 - DC

João Nunes (A82300)
Luís Braga (A82088)

13/03/2020

Conteúdo

1 Parte I

1.1 Quais as principais limitações de modelos de correlações?

Embora os modelos de correlação sejam extremamente úteis de modo a encontrar relações entre atributos de um dataset, não se deve tomar estas relações como *"absolutas"*, ao invés deve-se sempre procurar entender o resultado da correlação. Por exemplo, é possível que duas variáveis possuam alta correlação, contudo não podemos mesmo assim assumir que uma causa a outra uma vez que poderá haver uma terceira variável em jogo que influencia a correlação anterior. Em título de exemplo, assumindo que existe uma correlação positiva entre ver filmes violentos na televisão e possuir tendências violentas na adolescência. Não podemos assumir diretamente que estas duas variáveis estão ligadas uma vez que poderá existir uma terceira variável, como por exemplo crescer numa família violenta, que influencia o resultado anterior.

1.2 O que é um coeficiente de correlação e como é interpretado?

Um coeficiente de correlação pode ser positivo (varia entre 0 e 1) ou negativo (varia entre -1 e 0) e indica a maneira de como os valores se ligam. Ou seja, um coeficiente perto de 1 (com correlação positiva) indica que quando um atributo aumenta o outro também irá aumentar, e vice versa quando um desce o outro também desce. Por outro lado, um coeficiente perto de -1 indica que os atributos são *"inversos"* um do outro, assim quando um atributo cresce em valor o outro diminui, e vice versa. Resta referir que quanto mais pertos dos extremos (-1 ou +1) mais perfeita é a correlação, e caso dois atributos possuam um fator de correlação 0 então não existe qualquer tipo de ligação entre os dois atributos.

1.3 Qual a diferença entre uma correlação negativa e uma correlação positiva?

Tal como dito anteriormente, uma correlação positiva implica uma ligação linear na medida em que quando um atributo aumenta em valor ou outro também irá aumentar ou quando um atributo diminui em valor o outro também irá diminuir. Uma correlação negativa implica uma ligação inversa entre os atributos, ou seja, quando um atributo aumenta em valor o outro irá dimunuir e vice versa.

1.3.1 Se dois atributos diminuem essencialmente à mesma taxa é uma correlação positiva e negativa? Explique.

Se os dois atributos diminuem à mesma taxa então trata-se de uma correlação positiva, apenas seria uma correlação negativa se um aumenta-se e o outro diminui-se.

1.4 Como é medida a força de uma correlação? Quais os limites para essa força?

A força dos coeficientes de correlação é medida através do quão próxima está aos extremos da correlação positiva ou negativa, ou seja o -1 e o +1 respetivamente, quanto mais perto os coeficientes se encontram destes extremos mais forte é a força de correlação. Quanto mais próximo o coeficiente se encontra do zero então mais fraca é a força de correlação.

1.5 Consegue pensar em atributos que poderiam ser interessantes incluir no dataset estudado no exemplo da aula?

Atributos que seriam interessantes incluir no dataset estudado poderiam ser, por exemplo, o número de instrumentos que consomem óleo de aquecimento em cada casa ou até mesmo uma classificação energética de cada casa.

2 Parte II

2.1 Aceda ao ficheiro mpg_dataset.csv.

2.2 Execute a operação de Data Understanding

Na fase de visualização dos dados, o dataset pretendido foi aberto no *Weka*, onde foi possível verificar que 406 instâncias e 8 atributos, sendo que todos os atributos presentes no *dataset* são numéricos.

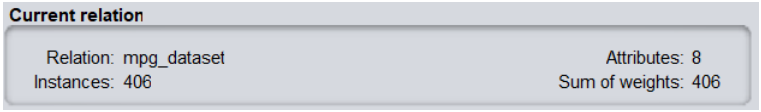

Figura 1: Dados gerais sobre o dataset.

Analisando cada um dos atributos foi possível identificar que o atributo *mpg* e o atributo *horsepower* possuem dados em falta.

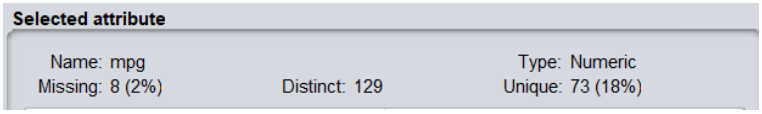

Figura 2: Dados em falta no atributo mpg.

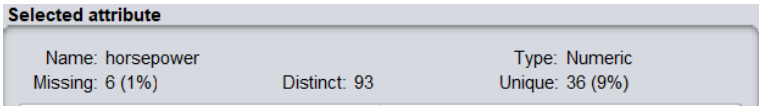

Figura 3: Dados em falta no atributo horsepower.

Para além disso, após verificar e analisar cada um dos atributos passou-se para a fase da visualização onde foi possível verificar a relação entre os atributos. Dentro de todas as relações destaca-se a relação quase linear entre o atributo *weight* e o atributo *displacement* na medida em que quando o *weight* aumenta o *displacement* também aumenta.

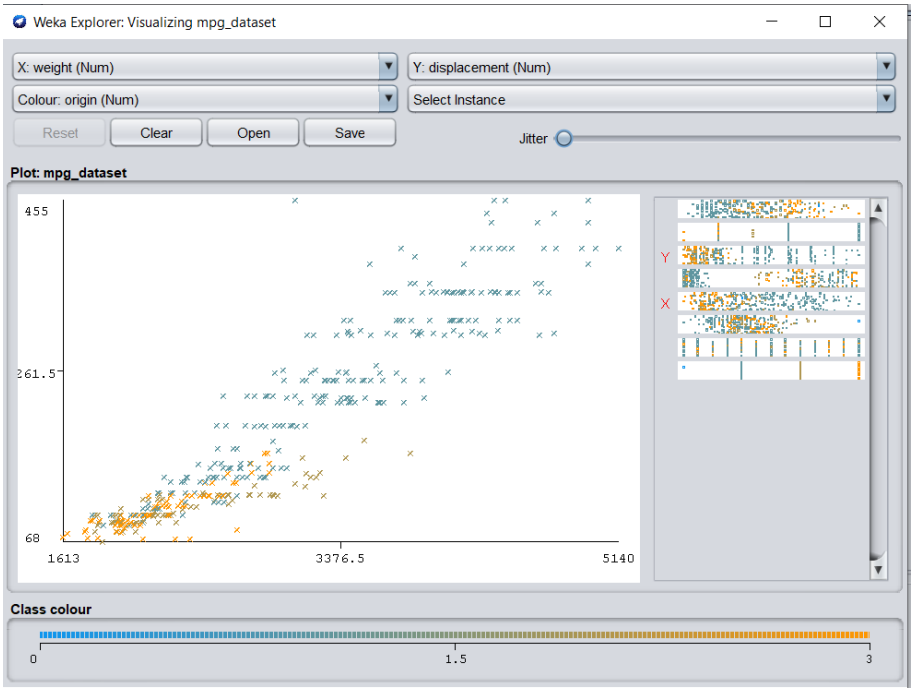

Figura 4: Relação entre o atributo weight e o displacement.

A partir da relação anterior poderá advinhar-se uma correlação forte e positiva entre ambos os atributos. Por sua vez também foi possível verificar que o atributo *weight* e o *mpg* também se relacionam de maneira inversa, na medida em que quando se aumenta o *weight* o *mpg* diminui, advinhando-se portanto uma correlação negativa entre ambos os atributos.

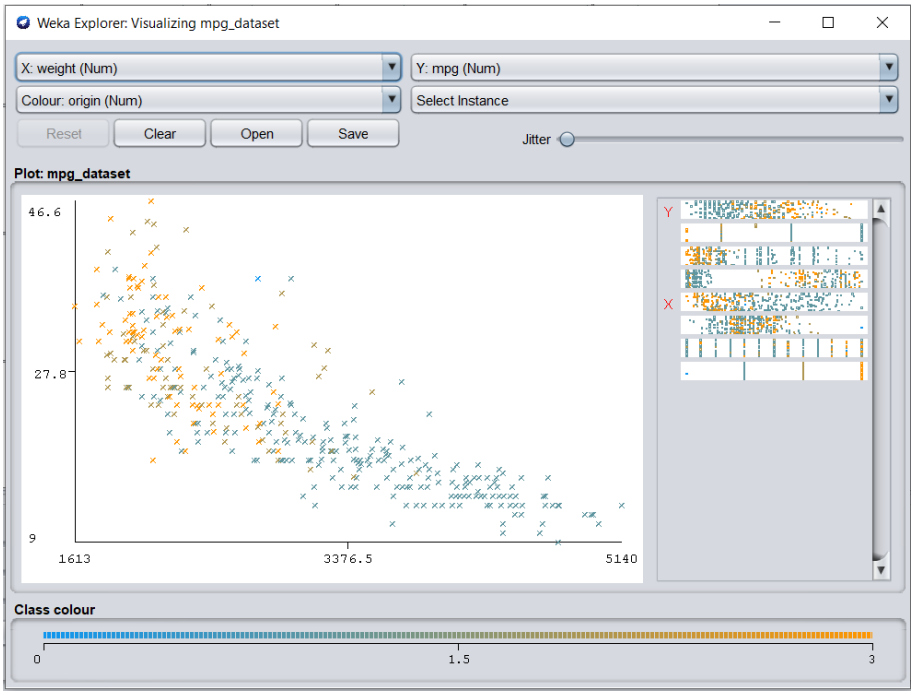

Figura 5: Relação entre o atributo weight e o mpg.

Portanto, atendendo à relação anterior é possível advinhar um fator de correlação negativo entre estes dois atributos.

2.3 Execute a etapa de Data Preparation no Weka. Não se esqueça de analisar a existência de 'outliers' e 'missing values'. Depois de devidamente processados, guarde os dados num ficheiro .csv que permita a execução no rapidminer do processo de correlação.

Ainda no *Weka* e atendendo à dica fornecida na pergunta foi aplicado um filtro *unsupervised* de *Replace Missing Values* que para os valores atributos que possuem dados em falta substitiu pela média (caso seja numérico o atributo) ou pela moda (caso seja nominal). Ou seja o algoritmo anterior foi utilizado para preencher os valores em falta para os atributos *mpg* e *horsepower* com o valor da média destes atributos.

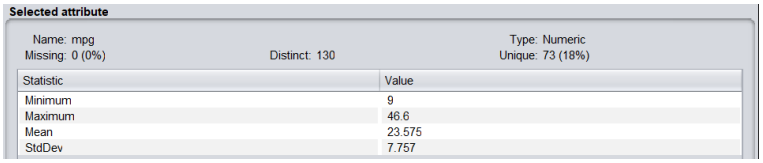

Figura 6: Atributo mpg sem missing values.

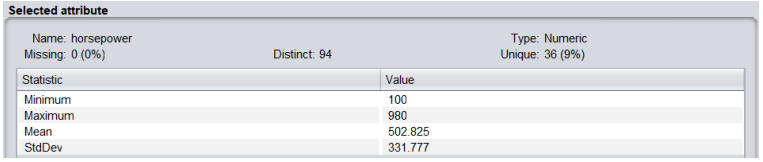

Figura 7: Atributo horsepower sem missing values.

Desta maneira foi possível obter o dataset sem valores em falta.

2.4 Documente quais os atributos que podem influenciar ou explicar o consumo/eficiência de combustível num determinado veículo (mpg).

Para esta fase e de modo a gerar a matriz de correlação, de modo a melhor entender a relação entre os atributos, foi necessário passar para um novo *software*, o *RapidMiner*.

No *RapidMiner* após importar o *dataset* sem *missing values* proveniente da questão anterior, foi possível gerar a seguinte matriz de correlação.

Attribut...	mpg	cylinders	displac...	horsep...	weight	acceler...	'model ...	origin
mpg	1	-0.749	-0.790	0.531	-0.820	0.414	0.566	0.546
cylinders	-0.749	1	0.946	-0.701	0.892	-0.482	-0.350	-0.570
displace...	-0.790	0.946	1	-0.651	0.933	-0.546	-0.380	-0.606
horsepo...	0.531	-0.701	-0.651	1	-0.639	0.375	0.227	0.355
weight	-0.820	0.892	0.933	-0.639	1	-0.418	-0.314	-0.580
accelerat...	0.414	-0.482	-0.546	0.375	-0.418	1	0.302	0.184
'model y...	0.566	-0.350	-0.380	0.227	-0.314	0.302	1	0.178
origin	0.546	-0.570	-0.606	0.355	-0.580	0.184	0.178	1

Figura 8: Matriz de correlação entre os atributos.

Partindo desta matriz é possível verificar os coeficientes de correlação de cada atributo, onde para o atributo *mpg* é possível verificar que existe uma correlação negativa do *mpg* com o *cylinders* e *displacement*, na medida em que aumentando o número de cilindros diminui-se o *mpg* e vice versa. A correlação positiva mais forte é com o atributo *model year* onde é possível verificar que quando um destes dois atributos aumenta o outro também irá aumentar (também se aplica se um diminuir), contudo em menor escala uma vez que o coeficiente de correlação não é muito elevado.

Olhando para o resto da matriz, é possível verificar que o coeficiente de correlação positivo maior é entre o atributo *cylinders* e *displacement* o que indica uma relação quase linear entre os dois atributos.

As relações identificadas anteriormente na pergunta 2.2 também se traduziram em coeficientes de correlação adequados tendo em conta o que foi escrito nessa questão. Ou seja a relação entre o *weight* e o *displacement* traduziu-se num fator de correlação de +0.933 tal como foi dito nessa questão (correlação forte positiva), e a relação entre o *weight* e o *mpg* também se traduziu num coeficiente de correlação de -0.820 tal como também foi previsto nessa questão (correlação negativa).