

Escola de Engenharia

Sistemas de Representação de Conhecimento e Raciocínio

**TRABALHO PRÁTICO Nº 3**

Mestrado Integrado em Engenharia Informática

Grupo 30

**78416 Francisco José Moreira Oliveira**

**79617 Raul Vilas Boas**

**79175 Vitor Emanuel Carvalho Peixoto**

**Ano letivo 2017/2018**

Braga, maio de 2018

# Resumo

Este trabalho prático foi realizado com o intuito de desenvolver e evoluir as competências adquiridas, na unidade curricular de Sistemas de Representação de Conhecimento e Raciocínio.

Este relatório serve então para explicar o processo de desenvolvimento e raciocínio, bem como as escolhas tomadas no decorrer deste.

Índice

[Resumo i](#_Toc514602153)

[1. Introdução 1](#_Toc514602154)

[2. Preliminares 2](#_Toc514602155)

[3. Descrição do trabalho e análise de resultados 3](#_Toc514602156)

[3.1 Normalização de dados 3](#_Toc514602157)

[3.2 Atributos mais significativos 4](#_Toc514602158)

[3.3 Fórmulas 6](#_Toc514602159)

[3.4 Análise de resultados 6](#_Toc514602160)

[4. Conclusões e sugestões 10](#_Toc514602161)

[5. Bibliografia 11](#_Toc514602162)

# Introdução

Neste terceiro trabalho o objetivo é, recorrendo a Redes Neuronais Artificias, explorar o uso de sistemas não simbólicos na representação de conhecimento e desenvolver mecanismos de raciocínio para resolver certos problemas.

O caso em estudo deste projeto está relacionado com dependência da qualidade do vinho conforme alguns dos seus parâmetros são alterados. Para isso, foi nos fornecido dois *datasets* com um conjunto de dados recolhidos, para assim pudermos tratar e analisar o conhecimento descrito pelos dados e com a ajuda de uma solução baseada em RNAs na linguagem de programação R tirar as conclusões relativas ao problema.

# Preliminares

Para se conseguir realizar tudo o que foi proposto no enunciado do trabalho prático houve a necessidade de compreender melhor os conceitos teóricos e métodos de aplicação dos mesmos para assim aplicarmos da melhor maneira possível o conhecimento adquirido.

As Redes Neuronais Artificiais são modelos parecidos ao sistema nervoso do ser humano com o objetivo de processar os dados de uma maneira parecida ao cérebro humano, isto é, uma estrutura extremamente interconectada de unidades computacionais designados por nodos ou neurónios.

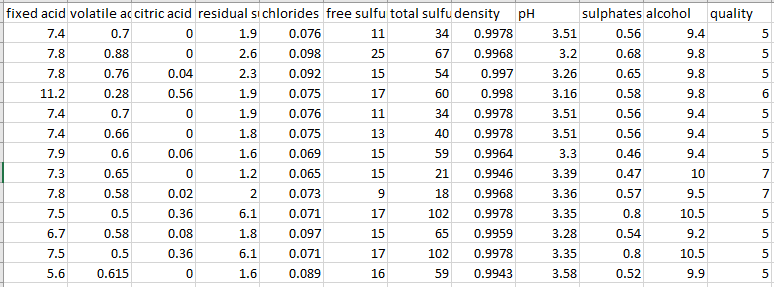
O objetivo das RNAs é processar a informação de forma a que sejam capazes de adquirir conhecimento, a partir de processos de aprendizagem, e de tomar decisões, e para além disso armazena esse mesmo conhecimento nas conexões entre os nodos, este comportamento é bastante semelhante ao comportamento do cérebro humano.

Os dois aspetos mais fundamentais do poder computacional de uma RNA é que numa topologia ela premeia o paralelismo e a sua capacidade de aprendizagem e generalização. Para além disto as RNAs apresentam características únicas [1], como por exemplo:

* **aprendizagem e generalização** - Conseguindo descrever o todo a partir de algumas partes, constituindo-se como formas eficientes de aprendizagem e armazenamento de conhecimento;
* **processamento maciçamente paralelo** - permitindo que tarefas complexas sejam realizadas num curto espaço de tempo;
* **transparência** - podendo ser vistas como uma caixa negra que transforma vetores de entrada em vetores de saída, via uma função desconhecida;
* **não linearidade** - atendendo a que muitos dos problemas reais a equacionar e resolver são de natureza não linear.

Um RNA é capaz de identificar padrões no input e produzir o output mais adequado à situação em causa, no entanto, para isso é necessário primeiro treinar a rede para ser possível identificar estes padrões. Após a rede neuronal estar treinada temos de a testar comparando o seu output com o output correto para assim calcular o RMSE (Root Mean Square Error) obtendo assim o valor do erro da rede.

# Descrição do trabalho e análise de resultados

Assim como instruído no enunciado do trabalho, os grupos que tinham um número par tinham de utilizar um certo *dataset.* No nosso caso foi utilizado um *dataset* obtido num repositório relativo à qualidade dos vinhos Este repositório continha dois *datasets* relativos a um estudo sobre o vinho tinto e vinho branco do norte de Portugal.

**Figura 1 -** Excerto do *dataset* do vinho tinto

## Normalização de dados

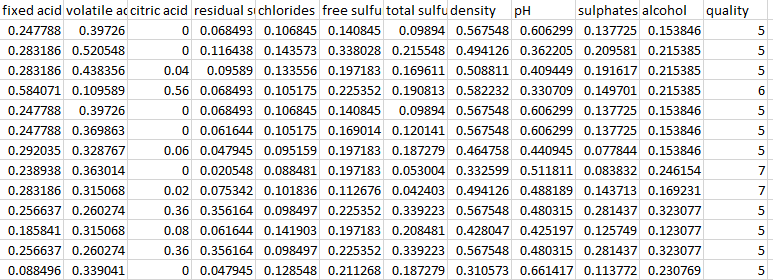
Visto que, os valores dos vários atributos estavam em diferentes intervalos de valores decidiu-se então normaliza-los para assim, estarem todos no mesmo intervalo, excetuando o atributo do output (quality) pois este é o valor que queremos obter e, por esse motivo, não se normalizou. O intervalo decidido foi entre [0,1], pois era o mais facilmente obtido a partir de um formula apresentada na figura 2.



**Figura 2 -** Formula utilizada para a normalização

Para normalizar os valores utilizou-se o Excel como ferramenta, pois já possuía funções que iriam auxiliar nos cálculos como, por exemplo, para calcular o valor máximo ou o valor mínimo das colunas, estes dois valores eram importantes pois estão presentes na formula como max(x) e min(x).

Por último, após aplicar a formula a todos os valores obteve-se assim os valores para todas a variáveis e passou-se estes novos valores para um novo ficheiro Excel.



**Figura 3 -** Excerto do *dataset* após a normalização

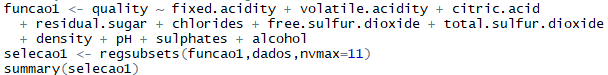
## Atributos mais significativos

Depois de normalizados podemos então ler o *dataset* no RStudio para futuramente o analisar. Para isso usou-se o comando ‘read’. Neste caso, ainda só estamos a analisar o *dataset* relativo ao vinho tinto.

**Figura 4** - Leitura do *dataset*

Para identificar os atributos mais significativos utilizou-se uma função existente nos pacotes disponíveis do R. A função ‘regsubsets’ permite descobrir quais são os atributos mais significativos para um determinado *dataset*, caso seja fornecido para além disto, uma formula e, por exemplo, um número máximo de objetos a analisar.

Para a formula, que se deu o nome de ‘funcao1’, deu-se o output do *dataset* que é a variável quality e também se deu todas as restantes variáveis. Visto que, vamos querer analisar as várias opções das variáveis significativas usou-se “nvmax=11” para pudermos ver todos os casos.



**Figura 5 -** Seleção das variáveis mais significativas

Assim, após obter os resultados no RStudio colocou-se o output na seguinte tabela. A tabela mostra que, por exemplo, para o caso de querermos as 3 variáveis mais significativas é só ver quais estão marcadas com uma cruz, isto é, as 3 variáveis mais significativas são o *Volatile Acidity*, *Sulphates* e o *Alcohol*.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fixed  Acidity | Volatile  Acidity | Citric  Acid | Residual  Sugar | Chlorides | Free Sulfur  Dioxide | Total Sulfur  Dioxide | Density | pH | Sulphates | Alcohol |
| 1 |  |  |  |  |  |  |  |  |  |  | x |
| 2 |  | x |  |  |  |  |  |  |  |  | x |
| 3 |  | x |  |  |  |  |  |  |  | x | x |
| 4 |  | x |  |  |  |  | x |  |  | x | x |
| 5 |  | x |  |  | x |  | x |  |  | x | x |
| 6 |  | x |  |  | x |  | x |  | x | x | x |
| 7 |  | x |  |  | x | x | x |  | x | x | x |
| 8 |  | x | x |  | x | x | x |  | x | x | x |
| 9 |  | x | x | x | x | x | x |  | x | x | x |
| 10 | x | x | x | x | x | x | x |  | x | x | x |
| 11 | x | x | x | x | x | x | x | x | x | x | x |

**Tabela 1** – Atributos significativos do *dataset* do vinho tinto

Para além de descobrir as variáveis mais significativas para o *dataset* do vinho tinto, também se fez para o *dataset* do vinho branco, obtendo o seguinte output apresentado seguinte na tabela.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fixed  Acidity | Volatile  Acidity | Citric  Acid | Residual  Sugar | Chlorides | Free Sulfur  Dioxide | Total Sulfur  Dioxide | Density | pH | Sulphates | Alcohol |
| 1 |  |  |  |  |  |  |  |  |  |  | x |
| 2 |  | x |  |  |  |  |  |  |  |  | x |
| 3 |  | x |  | x |  |  |  |  |  |  | x |
| 4 |  | x |  | x |  | x |  |  |  |  | x |
| 5 |  | x |  | x |  |  |  | x | x |  | x |
| 6 |  | x |  | x |  |  |  | x | x | x | x |
| 7 |  | x |  | x |  | x |  | x | x | x | x |
| 8 | x | x |  | x |  | x |  | x | x | x | x |
| 9 | x | x |  | x |  | x | x | x | x | x | x |
| 10 | x | x |  | x | x | x | x | x | x | x | x |
| 11 | x | x | x | x | x | x | x | x | x | x | x |

**Tabela 2** - Atributos mais significativos do *dataset* do vinho branco

## Fórmulas

Após descobrir os atributos mais significativos podemos passar à criação de fórmulas com estes mesmos atributos, depois estas fórmulas irão ser passadas à rede neuronal. Para comparar qual será a melhor formula criou-se assim onze fórmulas com um número de atributos significativos diferentes e testou-se para ver qual dava o melhor erro, mantendo os outros parâmetros todos iguais.



**Figura 6** - Exemplo de 4 fórmulas criadas para o *dataset* do vinho tinto

Visto que, os atributos mais significativos de ambos os *dataset* são diferentes foi necessário criar as respetivas fórmulas para o *dataset* do vinho branco.

## Análise de resultados

A ferramenta que utilizamos para treinar a RNA foi a função neuralnet em que os parâmetros utilizados foram: formula, data, hidden, threshold, algoritmo, lifesign e rep.

A fórmulas usada foram as que fórmulas com os atributos mais significativos mostradas anteriormente. A parâmetro data vai ser o *dataset* que queremos analisar. O hidden representa o número de nodos escondidos que queremos usar e o threshold é o valor de erro que vai parar a execução do programa. Por último, temos o lifesing que indica quanta informação vai ser mostrada no ecrã e o rep que indica quantas repetições queremos fazer.



**Figura 7** -Treino da rede neuronal

Assim, fixamos estes parâmetros, com os valores apresentados em cima. apenas alterando a formula, para assim, descobrir com qual se obtinha o menor erro.

Primeiro realizou-se isto para o *dataset* do vinho tinto e depois para o *dataset* do vinho branco.

|  |  |
| --- | --- |
| **Formula** | **rmse** |
| 01 | 0.8141076621 |
| 02 | 0.7354600618 |
| 03 | 0.7320472357 |
| 04 | 0.7112264561 |
| 05 | 0.7053338404 |
| 06 | 0.7041493998 |
| 07 | 0.6812538769 |
| 08 | 0.7135698774 |
| 09 | 0.7029629636 |
| 10 | 0.7511815957 |
| 11 | 0.7922833172 |

**Tabela 3** - Resultados das fórmulas do *dataset* do vinho tinto

Com base nos resultados obtidos, concluímos que a formula que possui um menor erro é a que contém 7 atributos e, por esse motivo, vamos utilizar esta fórmula enquanto alteramos os outros parâmetros para analisar quais as melhores opções que melhoram o resultado.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hidden** | **rep** | **threshold** | **Algoritmo** | **rmse** |
| c(4) | 1 | 0.1 | rprop+ | 0.6812538769 |
| c(4) | 1 | 0.5 | rprop+ | 0.7029629636 |
| c(4) | 3 | 0.05 | rprop+ | 0.7112264561 |
| c(6) | 1 | 0.01 | rprop+ | 0.7752429745 |
| c(10) | 1 | 0.01 | rprop+ | 0.8048268408 |
| c(4) | 1 | 0.1 | rprop- | 0.7170706501 |
| c(5) | 2 | 0.05 | rprop- | 0.7029629636 |
| c(4) | 1 | 0.03 | rprop- | 0.7456048177 |
| c(4) | 1 | 0.3 | sag | 0.7123991303 |
| c(4) | 1 | 0.2 | sag | 0.6970004898 |
| c(4) | 1 | 0.2 | sag | Não convergiu |
| c(8) | 2 | 0.3 | sag | 0.7354600618 |
| c(4) | 1 | 0.1 | slr | 0.7053338404 |
| c(4) | 1 | 0.05 | slr | 0.6946011695 |
| c(5) | 2 | 0.05 | slr | 0.7343242155 |

**Tabela 4** - Resultados obtidos do rmse do *dataset* do vinho tinto

As variações nos resultados dos erros não são muito significativas o que dificulta à análise para conseguirmos tirar conclusões, no entanto, o parâmetro que se destacou foi a alteração do algoritmo em que podemos concluir que o melhor algoritmo para o caso é o ‘prop+’, pois foi o que obteve o menor erro.

Após concluir o estudo do *dataset* realizou-se os mesmos passos para o *dataset* do vinho branco. Primeiro, fixamos todos os parâmetros alterando apenas as fórmulas com as variáveis significativas à procura da que apresenta o menor erro. Os resultados obtidos encontram-se na seguinte tabela.

|  |  |
| --- | --- |
| **Formula** | **rmse** |
| 01 | 0.8714731659 |
| 02 | 0.7476764304 |
| 03 | 0.7318707153 |
| 04 | 0.7172702522 |
| 05 | 0.7424452467 |
| 06 | 0.7616942784 |
| 07 | 0.7484207575 |
| 08 | 0.7288212407 |
| 09 | 0.7461855487 |
| 10 | 0.7550866851 |
| 11 | 0.7565600264 |

**Tabela 5** - Resultados das fórmulas do *dataset* do vinho branco

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hidden** | **rep** | **threshold** | **Algoritmo** | **rmse** |
| c(4) | 1 | 0.1 | rprop+ | 0.7172702522 |
| c(4) | 1 | 0.5 | rprop+ | 0.7439436241 |
| c(4) | 1 | 0.05 | rprop+ | 0.7219128266 |
| c(6) | 2 | 0.1 | rprop+ | 0.7157160353 |
| c(6) | 1 | 0.01 | rprop+ | 0.7055301755 |
| c(10) | 1 | 0.01 | rprop+ | Não convergiu |
| c(4) | 1 | 0.1 | rprop- | 0.7558237147 |
| c(5) | 2 | 0.05 | rprop- | 0.725758953 |
| c(4) | 1 | 0.03 | rprop- | 0.7195952834 |
| c(4) | 1 | 0.1 | sag | Não convergiu |
| c(4) | 1 | 0.5 | sag | 0.7180460991 |
| c(6) | 1 | 0.3 | sag | Não convergiu |
| c(4) | 1 | 0.1 | slr | Não convergiu |
| c(4) | 1 | 0.3 | slr | 0.7341495087 |
| c(8) | 1 | 0.03 | slr | 0.7280568763 |

**Tabela 6** - Resultados do rmse do *dataset* do vinho branco

A partir dos resultados obtidos, podemos concluir que para este *dataset* a fórmula que apresenta menor valor é a formula 4 que contém os 4 atributos mais significativos. Agora vamos repetir o que foi feito para o outro *dataset* para ver quais as opções que reduzem mais o erro.

Assim como foi concluímos no outro *dataset* o parâmetro onde se notou uma maior diferença nos resultados dos erros foi no algoritmo utilizado, pois em alguns deles nem chegava a convergir como é comprovado pelos resultados da tabela. A alteração do resto dos parâmetros não se notou numa grande diferença do valor do erro excetuando quando se diminui o valor do threshold em que houve uma diminuição do valor do erro em alguns dos casos.

# Conclusões e sugestões

Após concluir todas as funcionalidades propostas do enunciado, podemos passar a uma análise final do trabalho realizado.

Neste trabalho foram aplicadas as várias noções da linguagem R que foram aprendidas tanto nas aulas teóricas como nas aulas práticas que serviu para consolidar esses mesmo conceitos, assim sendo, podemos observar positivamente a realização deste trabalho.

Num projeto deste tipo, há sempre melhorias que podem ser implementadas, para permitir uma melhor manipulação do conhecimento armazenado e acrescentar mais funcionalidades. Esse é sem dúvida um trabalho a desenvolver futuramente, que seria capaz de melhorar este sistema.

# Bibliografia

|  |  |
| --- | --- |
| [1] | P. Cortez e J. Neves, “Redes Neuronais Artificiais,” Universidade do Minho, Braga, 2000. |