Universidade do Minho
Mestrado Integrado em Engenharia Informática

Unidade Curricular de Sistemas de Representação de Conhecimento e Raciocínio

Ano Letivo de 2017/2018

3.º Exercício de Grupo

(Redes Neuronais Artificiais)

Diogo Emanuel da Silva Nogueira (a78957)

Fábio Quintas Gonçalves (a78793)

Sarah Tifany da Silva (a76867)

Grupo 8

Maio, 2018

Resumo

O presente relatório elaborado no âmbito da unidade curricular Sistemas de Representação de Conhecimento e Raciocínio é todo um resultado da elaboração do terceiro e último exercício proposto da componente prática desta unidade curricular. Abordando o tema *Conhecimento Não Simbólico (Redes Neuronais Artificiais)*, este é um exercício completamente distinto dos anteriores no sentido em que pretende motivar a utilização de sistemas não simbólicos na representação de conhecimento e no desenvolvimento de mecanismos de raciocínio (RNAs) para a resolução de problemas.

Tendo como base dois amplos conjunto de dados previamente fornecidos e após se ter recolhido/estudado os vários atributos mais significativos para representação do problema em análise, o grande e principal objetivo deste exercício prático passará por identificar as topologias de rede mais adequadas e selecionar as regras de aprendizagem de modo a treinar essas mesmas redes, chegando-se finalmente à resolução pretendida – identificação da qualidade do vinho com base nos diferentes testes químicos/atributos. Tudo isto, recorrendo ao ambiente de análise de dados R.

Índice

1. Introdução

Este terceiro exercício prático teve de ser pensado de uma forma mais metódica e sequencial de modo a abordar todos os aspetos pedidos e ao mesmo tempo ir de encontro com aquilo que realmente se pretende resolver.

Partindo-se de dois conjuntos de dados que aglomeram milhares de linhas de informação, todas elas relacionadas com amostras de vinho verde vermelho e branco, o que se ambiciona é identificar os diferentes níveis de qualidade de vinho com base nos diferentes testes químicos efetuados a cada uma destas amostras.

Assim, o relatório tratará de numa fase inicial de recolher alguma informação importante e esclarecedora relativamente aquilo que é uma rede neuronal artificial, tentando-se contextualizar com todo o objetivo do trabalho em si. Depois, uma análise dos dados atenta e uma explicação de como foi pensada e estabelecida a normalização das duas *datasets*. Por fim, uma análise de todos os resultados através da "separação" dos dois conjuntos de dados em termos de conteúdo, das várias fórmulas desenvolvidas e ainda dos testes criados para o efeito.

2. Preliminares

As Redes Neuronais Artificias, ou também designadas RNAs, acarretam inúmeras vantagens, uma vez que são inspiradas no sistema nervoso central humano.

Das inúmeras vantagens que apresentam podemos destacar:

- A sua capacidade de aprendizagem e generalização;
- O seu processamento maciçamente paralelo;
- A sua transparência relativamente ao utilizador e a sua não linearidade, salientando que, contrariamente ao cérebro humano, não possuem a propriedade de esquecimento.

Devido a estas inúmeras propriedades/vantagens que as RNAs apresentam, tornam-se num método bastante útil para solucionar problemas. É precisamente neste tema tão inovador em que todo este trabalho se foca: a realização de Redes Neuronais Artificias com o objetivo de modelar a **qualidade de vinho** tendo como núcleo o conjunto de dados fornecidos relativos aos testes físico-químicos realizados ao *Red Wine* e *White Wine* Portugueses.

3. Descrição do Trabalho e Análise dos Resultados

3.1. Análise dos Dados e Normalização

3.1.1. Análise de Dados

Antes mesmo de prosseguimos para o desenvolvimento do trabalho em si, é necessário produzir uma análise cuidada e pormenorizada a todas as identidades e domínios do conjunto de dados que nos foram fornecidos, visto estes mesmos servirão para averiguar a qualidade dos vinhos Portugueses. Estes dados são extremamente relevantes na medida em o seu tratamento é invisível para o utilizador, garantindo que os resultados obtidos ou os dados sejam imparciais.

Por conseguinte, relativamente ao conjunto de dados do *Red Wine* e *White Wine*, apresentamos os seguintes dados de entrada (input) das RNAs:

- Acidez volátil (*volatile acidity*);
- Acido cítrico (*citric acid*);
- Açúcar residual (*residual sugar*);
- Cloretos (*chlorides*);
- Dióxido de enxofre livre (*free sulfur dioxide*);
- Dióxido de enxofre total (*total sulfur dioxide*);
- Densidade (*density*);
- Ph;
- Sulfatos (*sulfates*);
- Álcool (*alcohol*)

Todos estes dados pertencem ao domínio dos números reais, contrariamente aos dados de saída (output), apresentados de seguida, que pertencem ao domínio dos números inteiros positivos:

- Qualidade (*quality)* – entre 0 a 10.

3.1.2. Normalização

Primeiramente, antes mesmo de prosseguirmos ao desenvolvimento das redes neuronais em si, começamos por normalizar os dados de entrada para que os valores sejam todos reduzidos ao mesmo limite: [0,1]. Este passo é necessário para que seja possível comparar valores com a mesma ordem de magnitudes de modo a que a rede não tome decisões erradas.

Logo, para as tabelas relativas ao *Red Wine* e *White Wine* foi necessário atualizar os valores dos dados de entrada, que referimos anteriormente, utilizando a seguinte fórmula:

$$\text{novo} = \frac{\text{atual} - \min}{\max - \min}$$

atual: valor a ser normalizado

min: valor mínimo relativo ao atributo o qual pertence o atual

max: valor máximo relativo ao atributo o qual pertence o atual

3.1.3. Redes Neuronais

A topologia geral utilizada para a resolução das Redes Neuronais, tanto para o caso de estudo relativo ao *Red Wine* como para o *White Wine*, pertence à categoria **Redes *Feedforward* de MultiCamada (RFMC)**.

A escolha deste tipo de rede baseou-se nos seguintes parâmetros:

- Ao aumentar o nº de camadas intermediárias está-se também a aumentar a capacidade da rede para **modelar** funções de maior complexidade, aumentando-se assim a capacidade de aprendizagem desta;
- Torna-se uma topologia mais adequada quando o número de nodos na camada de entrada é elevado;
- Adapta-se facilmente para problemas de classificação entre outros.

Com base em tudo isto, decidiu-se construir diferentes topologias de Redes *Feedforward* de Multicamada, fazendo variar os números de nodos por camada intermediária bem como o nº de camadas (até um máximo de duas).

Tal decisão pode ser suportada e comprovar pelo exemplo que iremos anexar a seguir que corresponde a uma rede com uma camada intermediária com quatro nodos escondidos. Este *plot* é facilmente obtido através do software *RStudio* e tendo como base todo o trabalho empregue no mesmo até então.

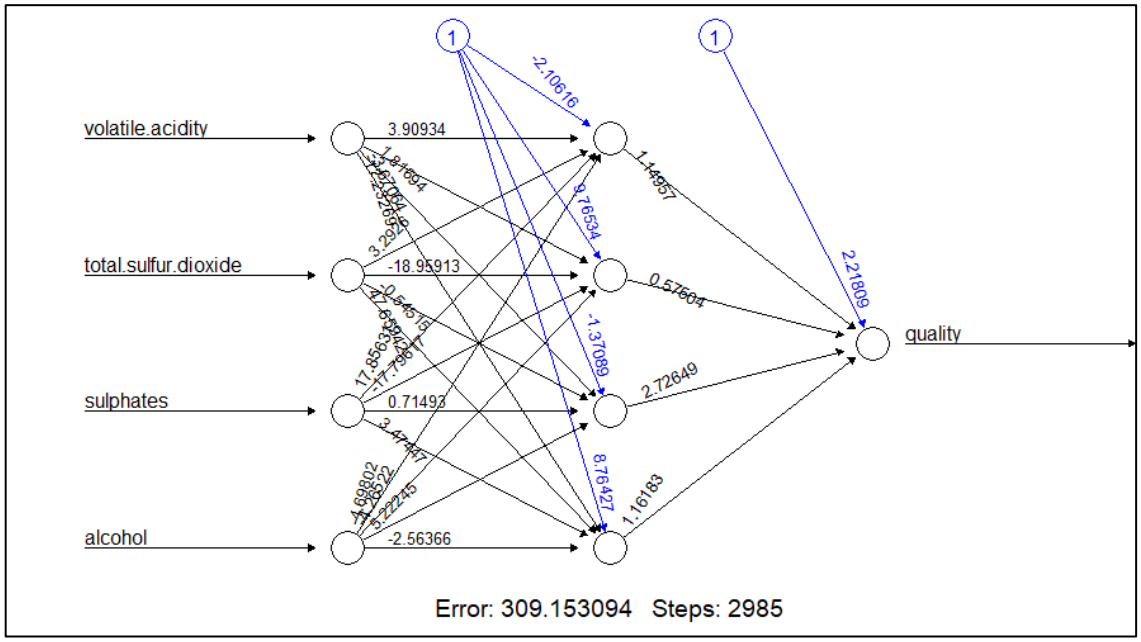


Figura 1: Topologia da RNA

Para a realização do treino das RNAs recorremos à função ***neuralnet*** apresentada genericamente de seguida**:**

$$neuralnet\ (formula, data, hidden, threshold, lifesign, threshold, algorithm, rep)$$

- **formula:** formula utilizada para treinar a rede;
- **data:** dataset que contem as variáveis da fórmula;
- **hidden:** define o nº de nodos escondidos (nº de camada intermediárias);
- **threshold:** valor de erro que irá parar a execução da função;
- **algorithm** : algoritmo responsável pela realização do treino;
- **rep**: número de repetições para o treino da rede.

O modelo de aprendizagem que aplicamos às redes neuronais baseia-se no Não Supervisionado e a sua aprendizagem é concretizada através da descoberta das características nos dados de entrada, possuindo uma capacidade de adaptação a regularidades estatísticas ou agrupamento de padrões dos dados de treino.

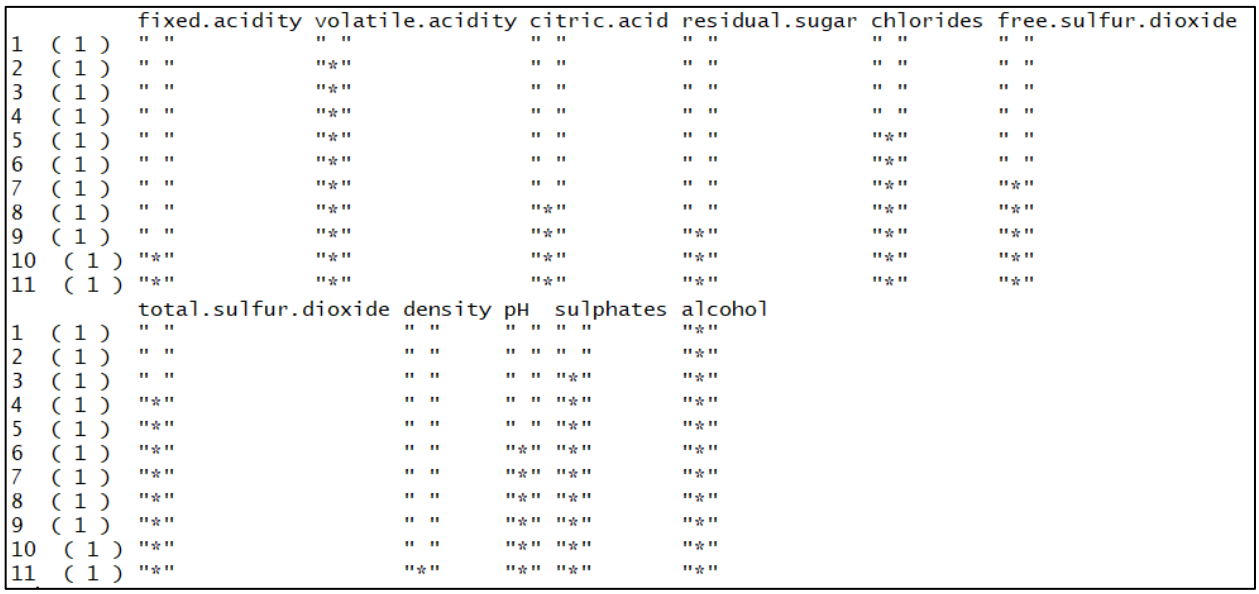
Assim, neste exercício prático utilizamos, essencialmente, vários algoritmos de aprendizagem como **RPROP(*rprop*),** o ***sag*** e o ***slr***. O algoritmo RPROP é uma variante do algoritmo *Back-Propagation* e mostra convergir mais rapidamente que os restantes algoritmos. O algoritmo ***sag*** e ***slr*** correspondem ao Algoritmo Globalmente Convergente. Este é baseado na retropropagação resiliente sem retrocesso de peso e modifica uma taxa de aprendizagem que no caso do ***sag*** a taxa a ser modificada corresponde à do menor gradiente absoluto, enquanto que no caso do ***slr,*** a taxa a ser modificada corresponde à de menor taxa de aprendizagem.

3.2. Identificação da Qualidade do Vinho

3.2.1. Conjunto de Dados *Red Wine*

3.2.1.1. Relevância dos Atributos

Partindo-se da observação das "tabelas" abaixo anexadas é possível visualizar os atributos mais relevantes para a qualidade do vinho (*quality*). Todos os resultados foram obtidos através do *software RStudio*.



```
          fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
1  ( 1 )  " "           " "              " "         " "            " "       " "
2  ( 1 )  " "           "*"              " "         " "            " "       " "
3  ( 1 )  " "           "*"              " "         " "            " "       " "
4  ( 1 )  " "           "*"              " "         " "            " "       " "
5  ( 1 )  " "           "*"              " "         " "            "*"       " "
6  ( 1 )  " "           "*"              " "         " "            "*"       " "
7  ( 1 )  " "           "*"              " "         " "            "*"       "*"
8  ( 1 )  " "           "*"              "*"         " "            "*"       "*"
9  ( 1 )  " "           "*"              "*"         "*"            "*"       "*"
10  ( 1 ) "*"           "*"              "*"         "*"            "*"       "*"
11  ( 1 ) "*"           "*"              "*"         "*"            "*"       "*"
          total.sulfur.dioxide density pH  sulphates alcohol
1  ( 1 )  " "                  " "     " " " "       "*"
2  ( 1 )  " "                  " "     " " " "       "*"
3  ( 1 )  " "                  " "     " " "*"       "*"
4  ( 1 )  "*"                  " "     " " "*"       "*"
5  ( 1 )  "*"                  " "     " " "*"       "*"
6  ( 1 )  "*"                  " "     "*" "*"       "*"
7  ( 1 )  "*"                  " "     "*" "*"       "*"
8  ( 1 )  "*"                  " "     "*" "*"       "*"
9  ( 1 )  "*"                  " "     "*" "*"       "*"
10  ( 1 ) "*"                  " "     "*" "*"       "*"
11  ( 1 ) "*"                  "*"     "*" "*"       "*"
```

Fazendo uma análise dos resultados obtidos, rapidamente concluímos que os 4 atributos mais relevantes são:

- volatile.acidity;
- total.sulfur.dioxide;
- sulphates;
- alcohol.

3.2.1.2. Fórmulas

Foram desenvolvidas várias formas de modo a se ir ao encontro com o pretendido.

- **Fórmula 1:** corresponde à fórmula que possui todas variáveis de entrada como atributos.

```
# Definição das camadas de entrada e saída da RNA
formula00 <- quality ~ fixed.acidity + volatile.acidity + citric.acid +
                       residual.sugar + chlorides + free.sulfur.dioxide +
                       total.sulfur.dioxide + density + pH + sulphates +
                       alcohol
```

- **Restantes fórmulas:** para além da fórmula principal, foram desenvolvidas mais 11 fórmulas, começando por aquela que possui apenas o primeiro atributo e terminando na que possui todos eles. Ou seja, de forma sequencial foi-se adicionando sempre mais um atributo à medida que se ia criando uma nova fórmula.

3.2.1.3. Escolha da melhor fórmula

Nº Teste	Fórmula	Hidden	Algoritmo	Threshold	Rep	Steps	Error	RMSE
1	formula.red01	c(4)	rprop+	0,1	1	248	156.25361	0.7790264238
2	formula.red02	c(4)	rprop+	0,1	1	495	139.72312	0.7413034205
3	formula.red03	c(4)	rprop+	0,1	1	948	129.5566	0.7037809404
4	formula.red04	c(4)	rprop+	0,1	1	7223	113.86392	0.7363636949
5	formula.red05	c(4)	rprop+	0,1	1	1527	122.76066	0.7213415841
6	formula.red06	c(4)	rprop+	0,1	1	10771	108.91959	0.7235066963
7	formula.red07	c(4)	rprop+	0,1	1	10403	103.28052	0.7607416391
8	formula.red08	c(4)	rprop+	0,1	1	5071	108.04638	0.7586827947
9	formula.red09	c(4)	rprop+	0,1	1	8390	102.18121	0.7162642046
10	formula.red10	c(4)	rprop+	0,1	1	4722	104.93137	0.7242269619
11	formula.red11	c(4)	rprop+	0,1	1	5208	107.5323	0.7566183478

Como podemos verificar através da análise da tabela anterior a fórmula que encontra a **menor raiz do erro médio quadrático (RMSE)** é a ***formula.red03*** que engloba na sua soma os atributos : *volatile.acidity*, *sulfates* e *alcohol*. Com esta descoberta, facilmente se faz uma comparação e se percebe que esta fórmula nada mais contém que os atributos anteriormente considerados significantes. É partindo da mesma, que vão ser feitos os vários testes e treinos para se poder chegar a uma conclusão final.

3.2.1.4. Discussão dos resultados

Nº Teste	Fórmula	Hidden	Algoritmo	Threshold	Rep	Steps	Error	RMSE
1	formula.red03	c(4)	rprop+	0,1	1	1488	129.78873	0.7242269619
2	formula.red03	c(4)	sag	0,1	1	15688	129.01441	0.7573071221
3	formula.red03	c(4)	slr	0,1	2	4231	129.93273	0.7155359236
4	formula.red03	c(4)	rprop+	0,05	1	37404	124.32993	0.7177185497
5	formula.red03	c(4,2)	rprop+	0,1	2	4472	125.12943	0.7299635027
6	formula.red03	c(4,2)	slr	0,1	1	11053	122.38476	0.7496958021
7	formula.red03	c(4,2)	sag	0,1	1	46	197.87814	0.8872871097
8	formula.red03	c(4,2)	rprop+	0,05	1	33940	117.14695	0.7384847661
9	formula.red03	c(5)	rprop+	0,1	1	3815	129.7807	0.7235066963
10	formula.red03	c(5)	sag	0,1	2	stepmax	-	-
11	formula.red03	c(5)	slr	0,1	1	1668	129.80222	0.7140771333
12	formula.red03	c(3,2)	rprop+	0,1	1	3406	122.64621	0.7206184344
13	formula.red03	c(4,2)	sag	0,05	3	stepmax	-	-
14	formula.red03	c(5,2)	slr	0,1	1	31	197.78668	0.8872871097
15	formula.red03	c(5,3)	rprop+	0,1	1	20489	104.36022	0.8143652889

Com base na análise da tabela apresentada anteriormente, podemos concluir que a rede mais adequada para a modularização da qualidade do *Red Wine* é a correspondente ao teste 11 já que apresenta o RSME menor.

Esta é uma rede *Feedforward* de MultiCamada com apenas uma camada intermediária e com 5 nodos escondidos. O algoritmo de aprendizagem utilizado para esta rede corresponde ao *sag*. – algoritmo globalmente convergente.

3.2.2. Conjunto de Dados *White Wine*

3.2.2.1. Relevância dos Atributos

Para a obtenção dos atributos mais relevantes relativos ao White Wine, segue-se o mesmo princípio do conjunto de dados anterior, no entanto, não esquecendo a necessidade de se gerar novas "tabelas" uma vez que estamos agora a lidar com o *dataset* correspondente às amostras de vinho branco.

Ainda assim, e sendo os testes físico-químicos efetuados nas várias amostras de vinho os mesmos, a fórmula principal para a obtenção dos atributos mais significativos permanece a mesma.

```
          fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
1   ( 1 ) " "           " "              " "         " "            " "       " "
2   ( 1 ) " "           "*"              " "         " "            " "       " "
3   ( 1 ) " "           "*"              " "         "*"            " "       " "
4   ( 1 ) " "           "*"              " "         "*"            " "       " "
5   ( 1 ) " "           "*"              " "         "*"            " "       " "
6   ( 1 ) " "           "*"              " "         "*"            " "       " "
7   ( 1 ) " "           "*"              " "         "*"            " "       "*"
8   ( 1 ) "*"           "*"              " "         "*"            " "       "*"
9   ( 1 ) "*"           "*"              " "         "*"            " "       "*"
10  ( 1 ) "*"           "*"              " "         "*"            "*"       "*"
11  ( 1 ) "*"           "*"              "*"         "*"            "*"       "*"
          total.sulfur.dioxide density pH  sulphates alcohol
1   ( 1 ) " "                  " "     " " " "       "*"
2   ( 1 ) " "                  " "     " " " "       "*"
3   ( 1 ) " "                  " "     " " " "       "*"
4   ( 1 ) " "                  "*"     " " " "       "*"
5   ( 1 ) " "                  "*"     "*" " "       "*"
6   ( 1 ) " "                  "*"     "*" "*"       "*"
7   ( 1 ) " "                  "*"     "*" "*"       "*"
8   ( 1 ) " "                  "*"     "*" "*"       "*"
9   ( 1 ) "*"                  "*"     "*" "*"       "*"
10  ( 1 ) "*"                  "*"     "*" "*"       "*"
11  ( 1 ) "*"                  "*"     "*" "*"       "*"
```

Fazendo uma análise dos resultados obtidos, rapidamente concluímos que os 4 atributos mais relevantes são:

- volatile.acidity;
- residual_sugar;
- density;
- pH;
- alcohol.

3.2.2.2. Fórmulas

Foram desenvolvidas várias formas de modo a se ir ao encontro com o pretendido.

- **Fórmula 1:** corresponde à fórmula que possui todas variáveis de entrada como atributos.

```
# Definição das camadas de entrada e saída da RNA
formula00 <- quality ~ fixed.acidity + volatile.acidity + citric.acid +
                       residual.sugar + chlorides + free.sulfur.dioxide +
                       total.sulfur.dioxide + density + pH + sulphates +
                       alcohol
```

- **Restantes fórmulas:** exatamente o mesmo princípio do *dataset* anterior.

3.2.2.3. Escolha da melhor fórmula

Nº Teste	Fórmula	Hidden	Algoritmo	Threshold	Rep	Steps	Error	RMSE
1	formula.white01	c(4)	rprop+	0,1	1	2948	679.05687	0.8805872024
2	formula.white02	c(4)	rprop+	0,1	1	21152	612.57265	0.8017079421
3	formula.white03	c(4)	rprop+	0,1	1	16806	608.20989	0.7980923155
4	formula.white04	c(4)	rprop+	0,1	1	31471	581.28114	0.8376151762
5	formula.white05	c(4)	rprop+	0,1	1	24959	563.6103	0.8069956128
6	formula.white06	c(4)	rprop+	0,1	1	73979	552.08182	0.8292459588
7	formula.white07	c(4)	rprop+	0,1	1	69440	508.16202	0.8101516499
8	formula.white08	c(4)	rprop+	0,1	1	77487	506.17347	0.8006462174
9	formula.white09	c(4)	rprop+	0,1	1	29761	495.34472	0.8310903238
10	formula.white10	c(4)	rprop+	0,1	1	17272	501.3502	0.8355816331
11	formula.white11	c(4)	rprop+	0,1	1	94652	486.76196	0.8017079421

Comprovando pela tabela anterior desenvolvida, a **menor raiz do erro médio quadrático (RMSE)** é a ***formula.white03*** que contém os atributos : *volatile.acidity*, *residual.sugar* e *alcohol*. Com esta fórmula definida e analisada, ficam assim mais do que patenteados aqueles que são os atributos mais significativos deste novo conjunto de dados.

3.2.2.4. Discussão dos Resultados

Nº Teste	Fórmula	Hidden	Algoritmo	Rep	Steps	Error	RMSE
1	formula.white11	c(4)	rprop+	1	9251	607.81003	0.7890885685
2	formula.white11	c(4)	sag	1	stepmax	-	-
3	formula.white11	c(4)	slr	2	26364	602.78126	0.7997958229
4	formula.white11	c(4)	rprop+	1	20016	603.60347	0.7797627474
5	formula.white11	c(4,2)	rprop+	2	37039	599.34287	0.8337472036
6	formula.white11	c(4,2)	slr	1	stepmax	-	-
7	formula.white11	c(4,2)	sag	1	27639	593.6498	0.7843310715
8	formula.white11	c(4,2)	rprop+	1	73094	582.38933	0.8465047324
9	formula.white11	c(5)	rprop+	1	25327	598.77152	0.7970257744
10	formula.white11	c(5)	sag	2	stepmax	-	-
11	formula.white11	c(5)	slr	1	7494	609.92831	0.8072063991
12	formula.white11	c(3,2)	rprop+	1	35033	599.80995	0.8158017492
13	formula.white11	c(4,2)	sag	3	stepmax	-	-
14	formula.white11	c(5,2)	slr	1	stepmax	-	-
15	formula.white11	c(5,3)	rprop+	1	90505	570.50394	0.7987315565

Analisando a tabela anterior, verificamos que a melhor rede cujo RMSE corresponde ao mínimo é a que diz respeito ao teste nº 4.

Esta é uma rede *Feedforward* de MultiCamada, constituída por apenas uma camada intermediária e com 4 nodos escondidos. O algoritmo de aprendizagem utilizado para a mesma foi o rprop+ (*Resilient back Propagation* com *backtracking*).

4. Conclusões

Fazendo parte do objetivo final do leque de objetivos definidos para a Unidade Curricular Sistemas de Representação de Conhecimento e Raciocínio, este foi o exercício que nos forçou a pôr de lado toda a ideia daquilo que é o Conhecimento Simbólico, baseado na parte lógica da programação, começando a trabalhar num novo tipo de conhecimento. Assim, iniciou-se toda uma busca mais focada na aprendizagem de conhecimento em si e não totalmente na lógica, entrando naquilo se denomina de Conhecimento Não Simbólico.

Com este novo tipo de conhecimento surgiram as tais Redes Neuronais Artificiais que se baseiam num modelo bem mais simplificado daquilo que é o sistema nervoso central do ser humano. Sendo algo mais sério e complexo daquilo que foi abordado até então, exigiu uma necessidade de entendimento maior e mais precisa para se poder estudar todo o funcionamento dos conjuntos de dados fornecidos.

Fazendo uso de todo o conhecimento arrecadado relativamente a este novo tema, focando sempre naquilo que os conjuntos de dados e respetivos atributos significavam para a modelação da qualidade dos dois tipos de vinho e percebendo-se como os vários testes e treinos das redes são supostos funcionar, foi mais do que possível chegar a uma conclusão face ao que se pretendia.