

Universidade do Minho

Escola de Engenharia

TRABALHO PRÁTICO EXERCÍCIO 3

Mestrado Integrado em Engenharia Informática

Sistemas de Representação de Conhecimento e Raciocínio

(2º semestre, 3º ano)

A78322 André Filipe Ferreira de Mira Vieira
A77048 Eduardo Gil Ribeiro Rocha
A78764 Ricardo André Araújo Neves

Data:

20 de maio de 2018

Braga,

20 de maio de 2018

Índice

Resumo	3
Introdução	4
Preliminares	
Normalização dos Dados	
Redes Neuronais Artificiais	6
Implementação em R	7
Conclusões e Sugestões	15
Referências	16

Resumo

Este relatório serve como complemento ao terceiro exercício do trabalho prático da Unidade Curricular de Sistemas de Representação de Conhecimento e Raciocínio.

Aqui iremos expressar a nossa linha de pensamento e como realizamos o exercício proposto, seguida de uma breve conclusão crítica sobre o trabalho realizado pelo grupo.

Introdução

Este **terceiro exercício** do trabalho prático tem em vista a utilização de sistemas não simbólicos na representação de conhecimento e no desenvolvimento de mecanismos de raciocínio, nomeadamente, Redes Neuronais Artificiais (RNAs) para a resolução de problemas.

Preliminares

O número do nosso grupo de trabalho é o o1, logo o dataset utilizado foi o Bank Marketing. Neste trabalho, o grupo procedeu ao tratamento e análise do conhecimento descrito pelos dados, desenvolvendo uma solução baseada em RNAs, com recurso à linguagem de programação R.



Figura 1 – Logotipo da linguagem R

Normalização dos Dados

Depois de descarregar o ficheiro .csv que contém todas as informações necessárias, contatamos que os valores não se encontravam normalizados.

```
["age";"job"; "marital"; "education"; "default"; "housing"; "loan"; "contact"; "month"; "day_of_week"; "duration"; "campaign"; "pdays"; "previous"; "poutcome"; "emp.var.rate"; "cons.price.idx"; "cons.conf.idx"; "euribor3m"; "n. employed"; "y"

2 56; "housemaid"; "married"; "basic.49"; "no"; "no"; "telephone"; "may"; "mon"; 149;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

3 77; "services"; "married"; "high.school"; "unknown"; "no"; "telephone"; "may"; "mon"; 126;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

48; "admin."; "married"; "basic.69"; "no"; "no"; "telephone"; "may"; "mon"; 226;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

48; "admin."; "married"; "basic.59"; "no"; "no"; "roelphone"; "may"; "mon"; 307;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

45; "services"; "married"; "basic.9y"; "unknown"; "no"; "no"; "telephone"; "may"; "mon"; 199;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

45; "services"; "married"; "basic.9y"; "unknown"; "no"; "no"; "telephone"; "may"; "mon"; 199;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

59; "admin."; "married"; "professional.course"; "no"; "no"; "telephone"; "may"; "mon"; 191;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

41; "blue-collar"; "married"; "unknown"; "no"; "no"; "no"; "telephone"; "may"; "mon"; 191;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"

41; "blue-collar"; "married"; "unknown"; "no"; "no"; "no"; "telephone"; "may"; "mon"; 191;1;999;0; "nonexistent"; 1.1;93.994; -36.4;4.857;5191; "no"
```

Figura 2 – Dados não normalizados

Posto isto, o grupo recorreu à ferramenta Excel para normalizar estes mesmos dados, obtendo o sequinte resultado.

```
age;job;marital;education;default;housing;loan;contact;month;day_of_week;duration;campaign;pdays;previous;poutcome;emp.var.rate;cons.price.idx;cons.conf.idx;euribor3m;nr.employed;y
0.481481481;housemaid;married;basic.4y;no;no;no;telephone;may;mon;0.053070354;0;1;0;nonexistent;0.9375;0.698752923;0.60251046;0.957379279;0.85973535;no
0.49382716;services;married;high.school;no;yes;no;telephone;may;mon;0.0392564;0;1;0;nonexistent;0.9375;0.698752923;0.60251046;0.957379279;0.85973535;no
0.26959617;admin:;married;basic.6y;no;no;no;no;telephone;may;mon;0.030703538;0;1;0;nonexistent;0.9375;0.698752923;0.60251046;0.957379279;0.85973535;no
0.481481481;services;married;high.school;no;no;no;yes;telephone;may;mon;0.062423749;1);0;nonexistent;0.9375;0.698752923;0.60251046;0.957379279;0.85973535;no
0.345679012;services;married;basic.9y;unknowm;no;no;telephone;may;mon;0.040260268;0;1;0;nonexistent;0.9375;0.698752923;0.60251046;0.957379279;0.85973535;no
0.518518519;admin:;married;professional.course;no;no;no;telephone;may;mon;0.02824263522;0;1;0;nonexistent;0.9375;0.698752923;0.60251046;0.957379279;0.85973535;no
```

Figura 3 – Dados normalizados

Para esta normalização, o grupo utilizou a seguinte fórmula.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Redes Neuronais Artificiais

Com os dados normalizados, está agora na altura de implementar as RNAs.

Uma Rede Neuronal Artificial é um modelo computacional inspirado pelo sistema nervoso central, responsável por realizar o aprendizado de uma máquina e reconhecimento de padrões.

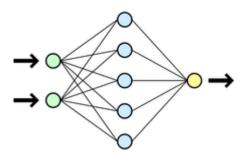


Figura 4 – Diagrama simplificado de uma rede neuronal

As RNAs têm várias aplicações possíveis, como o reconhecimento de caracteres, robótica, biometria, processamento de voz, entre outros exemplos.

Neste caso que foi trabalhado pelo grupo, os dados são relativos a uma campanha de marketing de um banco português. O objetivo principal é prever se um determinado cliente, com uma série de informações associadas, irá subscrever a um depósito com o banco respetivo.

Implementação em R

As librarias importadas foram:

```
library(neuralnet)
library(hydroGOF)
library(arules)
library(leaps)
```

Figura 5 – Librarias importadas

De seguida, é necessário indicar onde se encontra o ficheiro .csv com todas as informações precisas.

```
# ler dataset do ficheiro csv dados <- read.csv("C:\\Users\\Ricardo\\Desktop\\Universidade\\SRCR\\TP3\\bank-additional\\bank-additional-full.csv", header=TRUE, sep=";")
```

Figura 6 – Indicação do ficheiro .csv

Podemos observar que o separador dos dados foi selecionado como um ponto e vírgula.

Algumas variáveis não são discretas, ou seja, não se tratam de números, mas sim de "strings". Assim, foi necessário transformar estas "strings" em outros números, para ser possível o cálculo da RNA.

```
dados$job <- as.numeric(dados$job)
dados$marital <- as.numeric(dados$marital)
dados$education <- as.numeric(dados$education)
dados$default <- as.numeric(dados$default)
dados$housing <- as.numeric(dados$housing)
dados$loan <- as.numeric(dados$loan)
dados$contact <- as.numeric(dados$contact)
dados$month <- as.numeric(dados$month)
dados$day_of_week <- as.numeric(dados$poutcome)
dados$y <- as.numeric(dados$poutcome)</pre>
```

Figura 7 – Discretização da informação

Antes de efetuar estas instruções, os dados não eram discretos, como é o caso da variável job, education, loan, entre muitas outras.

```
age : num 0.481 0.494 0.247 0.284 0.481 ...
job : Factor w/ 12 levels "admin.", "blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ... marital : Factor w/ 4 levels "divorced", "married",..: 2 2 2 2 2 2 2 3 3 ...
education : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...
default : Factor w/ 3 levels "no", "unknown", ...: 1 2 1 1 1 2 1 2 1 1 ...
housing : Factor w/ 3 levels "no", "unknown",..: 1 1 3 1 1 1 1 1 3 3 ...
loan : Factor w/ 3 levels "no", "unknown",...: 1 1 1 1 3 1 1 1 1 1 ...
contact : Factor w/ 2 levels "cellular", "telephone": 2 2 2 2 2 2 2 2 2 2 ...
month : Factor w/ 10 levels "apr", "aug", "dec", ...: 7 7 7 7 7 7 7 7 7 7 7 ...

day_of_week : Factor w/ 5 levels "fri", "mon", "thu", ...: 2 2 2 2 2 2 2 2 2 2 ...
duration : num 0.0531 0.0303 0.046 0.0307 0.0624 ...
campaign : num 0 0 0 0 0 0 0 0 0 0 ...
pdays : num 1 1 1 1 1 1 1 1 1 1 ...
previous : num 0 0 0 0 0 0 0 0 0 0 ...
poutcome : Factor w/ 3 levels "failure", "nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
emp.var.rate : num 0.938 0.938 0.938 0.938 ...
cons.price.idx: num 0.699 0.699 0.699 0.699 ...
cons.conf.idx : num 0.603 0.603 0.603 0.603 ...
euribor3m : num 0.957 0.957 0.957 0.957 ...
nr.employed : num 0.86 0.86 0.86 0.86 ...
y : Factor w/ 2 levels "no", "yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Figura 8 – Dados não discretos

Com todos os comandos executados, podemos ver que todos estas variáveis já se encontram discretas, com todos os seus valores numéricos.

```
age : num 0.481 0.494 0.247 0.284 0.481 ...
job : num 4 8 8 1 8 8 1 2 10 8 ...
marital : num 2 2 2 2 2 2 2 3 3 ...
education : num 1 4 4 2 4 3 6 8 6 4 ...
default : num 1 2 1 1 1 2 1 2 1 1 ...
housing: num 1 1 3 1 1 1 1 1 3 3 ...
loan : num 1 1 1 1 3 1 1 1 1 1 ...
contact : num 2 2 2 2 2 2 2 2 2 2 ...
month : num 7 7 7 7 7 7 7 7 7 7
day_of_week : num 2 2 2 2 2 2 2 2 2 2 . . .
duration : num 0.0531 0.0303 0.046 0.0307 0.0624 ...
campaign : num 0 0 0 0 0 0 0 0 0 0 ...
pdays : num 1 1 1 1 1 1 1 1 1 1 ...
previous : num 0 0 0 0 0 0 0 0 0 0 ...
poutcome : num 2 2 2 2 2 2 2 2 2 2 ...
emp.var.rate : num 0.938 0.938 0.938 0.938 ...
cons.price.idx: num 0.699 0.699 0.699 0.699 ...
cons.conf.idx : num 0.603 0.603 0.603 0.603 ...
euribor3m : num 0.957 0.957 0.957 0.957 ...
nr.employed : num 0.86 0.86 0.86 0.86 0.86 ...
y: num 1 1 1 1 1 1 1 1 1 1 ...
```

Figura 9 – Dados discretos (numéricos)

Agora já é possível ao programa, mostrar a cabeça do dataset, dividindo os dados em casos para treino e casos para teste.

```
head(dados)
# dividir os dados iniciais em casos para treino
treino <- dados[1:700, ]
# dividir os dados iniciais em casos para teste
teste <- dados[701:41188, ]
```

Figura 10 – Apresentação e divisão dos dados

```
> head(dados)
                                                                                                                                            duration campaign pdays previous poutcome emp.var.rate 053070354 0 1 0 2 0.9375 030298669 0 1 0 2 0.9375 045953640 0 1 0 2 0.9375 030703538 0 1 0 2 0.9375 030703538 0 1 0 2 0.9375 062423749 0 1 0 2 0.9375 040260268 0 1 0 2 0.9375
age job marital education default housing loan contact month day_of_week duration 1 0.481481481 4 2 1 1 1 1 2 7 2 0.053070354
0.030296869
```

Figura 11 — Apresentação da cabeça dos dados

A primeira fórmula, ou seja, a definição das camadas de entrada e saída da RNA, que o grupo usou foi a sequinte.

```
# definicao das camadas de entrada e saida da RNA formulatp3 <- y \sim age + education + duration
```

Figura 12 — Primeira fórmula utilizada

Esta fórmula representa que a resposta final do cliente depende da sua idade, das suas habilitações e da duração do seu último contrato. O grupo chegou a esta fórmula depois de analisar todas as variáveis e escolhemos 3 delas que achamos que mais se adequavam à resposta final do cliente. Por exemplo, estas variáveis fazem mais sentido, ao nosso ver, do que o tipo de contacto da pessoa em questão ou do mês que ela foi contactada.

Depois de formulada, é tempo de treinar a rede neuronal com as variáveis age, education e duration como input e y como output, estabelecendo o número de nodos intermédios a 3,

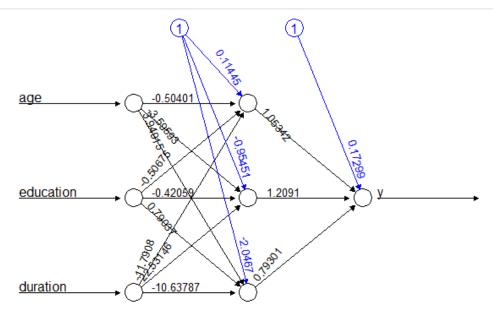
```
 rnabanco <- \ neuralnet( \ formulatp3, \ treino, \ hidd \ ene (3), lifesign = "full", \ linear.output = \ TRUE, \ threshold = 0.1)
```

Figura 13 – Treino da rede com o comando neuralnet

Com a rede neuronal treinada, podemos observar o desenho da mesma com o comando:

```
plot(rnabanco, rep = "best")
```

Figura 14 – Comando para desenho da RNA



Error: 3.67452 Steps: 652

Figura 15 – Desenho da rede neuronal

Podemos observar acima que o erro desta rede reside em 3.67452. O próximo objetivo é minimizar ao máximo este erro, alterando a fórmula acima descrita, acrescentando e retirando algumas variáveis.

De seguida, o grupo definiu as variáveis de input para teste, as mesmas que já foram referidas acima, testando, logo a seguir, a rede com estes novos casos.

```
teste.01 <- subset(teste, select = c("age", "education", "duration"))
# testar a rede com os novos casos
rnabanco.resultados <- compute(rnabanco,| teste.01)
# imprimir resultados
resultados <- data.frame(atual = teste$y, previsao = rnabanco.resultados$net.result)
# imprimir resultados arredondados
resultados$previsao <- round(resultados$previsao, digits=0)</pre>
```

Figura 16 — Teste da rede

Com os dados atuais e os dados previstos para a rede, é possível agora calcular o rootmeans-square error (RMSE), que é frequentemente usado para medir as diferenças entre valores previstos e os valores atualmente observados.

rmse(c(teste\$y),c(resultados\$previsao))

Figura 17 — Cálculo do RMSE

Este valor, dados os valores acima calculados, é igual a 0.3109335141. Para além do erro determinado mais acima, também o RMSE deverá descer, de modo a ter uma RNA mais viável.

Agora está na altura de selecionar as variáveis mais significativas, as variáveis que têm um maior impacto na resposta final do cliente: criar (ou não) um depósito no banco.

```
funcao <- y ~ age + education + duration + job + housing + marital + loan + month + day_of_week + previous
selecao <- regsubsets(funcao,dados,nvmax=3)
summary(selecao)
selecao <- regsubsets(funcao,dados,method="backward")
summary(selecao)</pre>
```

Figura 18 – Seleção das variáveis mais significativas

Desta vez, a função principal vai ser caracterizada por 10 variáveis, acima referidas.

Assim, no terminal do R Studio, podemos observar uma tabela que nos indica quais as variáveis mais impactantes da RNA.

```
> selecao <- regsubsets(funcao, dados, nvmax=4)
> summary(selecao)
Subset selection object
call: regsubsets.formula(funcao, dados, nvmax = 4)
19 Variables (and intercept)
Forced in Forced out
age FALSE FALSE
job FALSE FALSE
marital FALSE FALSE
education FALSE FALSE
default FALSE FALSE
loan FALSE FALSE
contact FALSE
day_of_week FALSE
fALSE
fALSE
fALSE
contact FALSE
fALSE
fALSE
contact FALSE
  education
default
housing
loan
contact
month
day_of_week
duration
 cons.price.idx euribor3m nr.employed
```

Figura 19 – Tabelas de seleção das variáveis mais significantes

Depois de uma análise cuidada, constatamos que as 4 variáveis mais significativas são as month, duration, pdays e nr.employed, que representam, respetivamente, o último mês de contacto com o cliente, a duração do último contrato, o número de dias que passaram desde a última comunicação com o cliente e o número de empregados do banco em questão. Nenhuma das variáveis agora escolhidas estão presentas na primeira fórmula designada pelo grupo.

Assim, os testes aplicados anteriormente vão ser aplicados de novo a esta segunda fórmula.

formulatp3 <- y \sim month + duration + pdays + nr.employed Figura 20 - Segunda fórmula da RNA

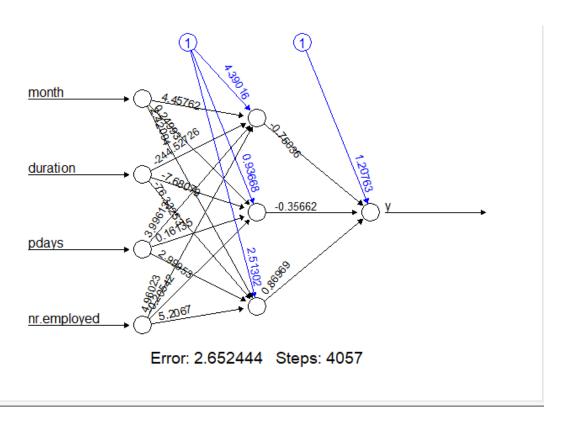


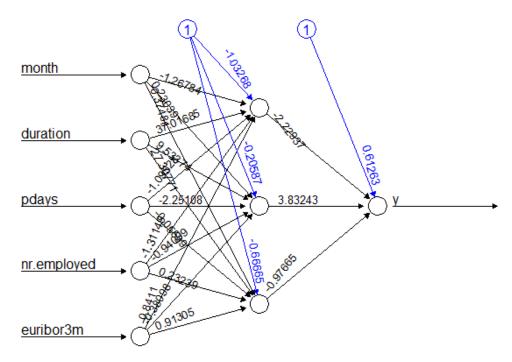
Figura 21 – Novo desenho da segunda rede neuronal

Podemos ver que, com esta nova fórmula e nova RNA, o erro desceu para 2.652, sensivelmente. O RNSE é agora igual a 0.31897171.

Analisaremos agora uma função objetivo com 5 variáveis, as 4variáveis que já conhecemos, mais a variável euribor3m, que representa a Euribor a 3 meses atual.

Deste modo, aplica-se de novo os testes.

formulatp3 <- y ~ month + duration + pdays + nr.employed + euribor3m Figura 22 - Terceira fórmula, com 5 variáveis



Error: 2.915296 Steps: 4640

Figura 23 – Novo desenho da segunda RNA

Neste terceiro teste, o valor do erro é igual a 2.915296 e o RNSE é igual a 0.7781586831.

Concluindo estes três testes, constatamos que a rede neuronal artificial mais fiável foi a segunda, com a fórmula com 4 variáveis, que podemos observar mais acima. Esta opinião baseia-se na rede com menor erro e com menor RNSE.

Conclusões e Sugestões

Em conclusão, este trabalho permitiu aprofundar os nossos conhecimentos em relação à linguagem R e Redes Neuronais Artificiais.

Podemos afirmar que surgiram algumas dificuldades ao longo da realização deste exercício, uma vez que nenhum dos elementos do grupo teve disponibilidade para ir à

primeira aula de RNAs e de R. No entanto, dedicamo-nos a perceber ao máximo como tudo funciona, com várias pesquisas na Internet e nos slides da disciplina.

De resto, concordamos que terminamos esta fase e todo o trabalho prático com sucesso.

Referências

[Leite, 1978] LEITE, J. A. Américo,

"Metodologia de Elaboração de Teses",

McGraw-Hill do Brasil, São Paulo, 1978.

[Analide, 2001] ANALIDE, Cesar, NOVAIS, Paulo, NEVES, José,

"Sugestões para a Elaboração de Relatórios",

Relatório Técnico, Departamento de Informática, Universidade do Minho, Portugal, 2001.