UFAM

Disciplina: ICC204 – Aprendizagem de Máquina e Mineração de Dados 07/07/2019
Prof. Rafael Giusti (rgiusti@icomp.ufam.edu.br)

# Lista de Exercícios 6

## Resolução

1. Qual é a relação entre a a dimensionalidade dos dados, a capacidade de um modelo e o fenômeno do *overfitting*?

   Existem múltiplas relações entre esses termos. Para contextualizar, vamos começar revendo o que eles significam.

   a) A dimensionalidade de um conjunto de dados é o número de dimensões que necessitamos para representar os exemplos;

   b) A capacidade de um modelo está relacionada com o tamanho do espaço de hipóteses que ele consegue representar;

   c) *Overfitting* é o fenômeno no qual um modelo superajustado aos exemplos de treinamento é selecionado, não havendo generalização do conceito.

   Tendo isso em mente, podemos pensar nas seguintes relações:

   - Entre dimensionalidade e capacidade: a capacidade do modelo não muda com a dimensionalidade dos dados, mas quanto maior a dimensionalidade, mais complexos podem ser os conceitos que queremos representar. Nesse caso, a tendência é que precisaremos aumentar a capacidade do modelo para continuar generalizando bem;

   - Entre capacidade e *overfitting*: quanto maior a capacidade do modelo, maior o espaço de hipóteses que ele consegue explorar, portanto maior o risco de encontrarmos uma hipótese que está superajustada aos exemplos de treinamento. Podemos reduzir o risco de *overfitting* obtendo mais exemplos a fim de caracterizar bem o espaço de atributos ou utilizando técnicas como regularização;

   - Entre dimensionalidade, capacidade e *overfitting*: nós reagimos ao aumento da dimensionalidade aumentando a capacidade do modelo. Para evitar o *overfitting*, iremos coletar mais dados ou tornar a regularização mais agressiva. Entretanto, quanto maior a dimensionalidade, mais difícil é obter mais dados, pois a quantidade de exemplos necessária para caracterizar bem o espaço de características aumenta exponencialmente com a dimensinalidade. A regularização funciona apenas até um certo ponto; eventualmente, iremos precisar de mais dados.

2. Normalize o conjunto de dados abaixo utilizando o método da estandardização (*z*-escores) e classifique o exemplo (4, 16, 18) utilizando o método dos k vizinhos mais próximos com k=1 e distância euclidiana.

| X1 | X2 | X3 | Classe |
|---|---|---|---|
| 1 | 10 | 3 | + |
| 1 | 12 | 16 | + |
| 1 | 16 | 33 | + |
| 2 | 14 | 3 | + |
| 2 | 17 | 18 | + |
| 2 | 18 | 34 | - |
| 3 | 25 | 26 | - |
| 3 | 29 | 18 | - |

Para normalizar, precisamos encontrar os valores de média e desvio padrão de todos os atributos. As médias de cada atributo são µ1 = 1,88, µ2 = 17,63 e µ3 = 18,88. O desvio padrão de uma amostra x = (x1, x2, ..., xn) com média µ é calculado por meio da seguinte fórmula:

$$s = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n - 1}}$$

Comecemos com o atributo X1 e calculemos os quadrados das diferenças com respeito:

d1 = d2 = d3 = $(1 - 1{,}88)^2$ = 0,7744
d4 = d5 = d6 = $(2 - 1{,}88)^2$ = 0,0144
d7 = d8 = $(3 - 1{,}88)^2$ = 1,2544
d1 + d2 + d3 + d4 + d5 + d6 + d7 + d8 = 4,8752
Então o desvio padrão do atributo X1 será sqrt(4,8752/7) = sqrt(0,6965) = 0,8346

Repetindo o procedimento para os outros três atributos, encontraremos que os desvios padrão são s1 = 0,84, s2 = 6,44 e s3 = 11,91. Caso algum desvio padrão fosse zero, selecionaríamos si = 1. Agora é só calcular os escores (xi - µi) / si

| X1std | X2std | X3std | Classe |
|---|---|---|---|
| -1.06 | -1.18 | -1.33 | + |
| -1.06 | -0.87 | -0.24 | + |
| -1.06 | -0.25 | 1.19 | + |
| 0.14 | -0.56 | -1.33 | + |
| 0.14 | -0.1 | -0.07 | - |
| 0.14 | 0.06 | 1.27 | - |
| 1.35 | 1.14 | 0.6 | - |
| 1.35 | 1.77 | -0.07 | - |

O exemplo de teste (4, 16, 18) deve ser normalizado com os valores de média e desvio padrão que calculamos para os exemplos do conjunto de treinamento. Portanto nosso exemplo de teste será (2.55, -0.25, -0.07).

X1_teste = (4 - 1,88) / 0,83 = 2,55
X2_teste = (16 - 17,63) / 6,44 = -0,25
X3_teste = (18 - 18,88) / 11,91 = -0,07

A distância ao quadrado do exemplo de teste para cada instância de treinamento será

$d^2(t, \#1) = (2.55 - -1.06)^2 + (-0.25 - -1.18)^2 + (-0.07 - -1.33)^2 = 15.48$
$d^2(t, \#2) = (2.55 - -1.06)^2 + (-0.25 - -0.87)^2 + (-0.07 - -0.24)^2 = 13.45$
$d^2(t, \#3) = (2.55 - -1.06)^2 + (-0.25 - -0.25)^2 + (-0.07 - 1.19)^2 = 14.62$
$d^2(t, \#4) = (2.55 - 0.14)^2 + (-0.25 - -0.56)^2 + (-0.07 - -1.33)^2 = 7.49$
$d^2(t, \#5) = (2.55 - 0.14)^2 + (-0.25 - -0.1)^2 + (-0.07 - -0.07)^2 = 5.83$
$d^2(t, \#6) = (2.55 - 0.14)^2 + (-0.25 - 0.06)^2 + (-0.07 - 1.27)^2 = 7.7$
$d^2(t, \#7) = (2.55 - 1.35)^2 + (-0.25 - 1.14)^2 + (-0.07 - 0.6)^2 = 3.82$
$d^2(t, \#8) = (2.55 - 1.35)^2 + (-0.25 - 1.77)^2 + (-0.07 - -0.07)^2 = 5.52$

Para obter os valores exatos das distâncias poderíamos extrair a raiz quadrada de cada termo, porém isso não é necessário, já que apenas queremos saber qual vizinho tem a menor distância. O vizinho mais próximo será o exemplo de treinamento #7, com distância 3,82, cuja classe é negativa. Portanto o exemplo de teste será classificado como "-".

3. Monte uma pequena base de treino e treine um modelo perceptron, utilizando função sigmoide, para classificar a função ((X1 **e** X2) **ou** (X3)). Faça uma única época, isto é, uma única passagem pelo conjunto de dados.

Conjunto de treino:

| X1 | X2 | X3 | Classe |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |

Vamos treinar o perceptron pela *regra do perceptron*, utilizando $\eta$=0,5. Os pesos iniciais serão $w_0$=0.1, $w_1$=-0.1, $w_2$=0.2 e $w_3$=0.4. A regra do perceptron consiste de atualizar os pesos de acordo com

erro = (*t* - *o*), sendo que *o* é a saída do perceptron após a função degrau
$\delta_i = \eta \cdot \text{erro} \cdot x_i$

Para o primeiro exemplo x=(0, 0, 0), t=0

Função soma: $w \cdot x = 1 \cdot 0,1 - 0 \cdot 0,1 + 0 \cdot 0,2 + 0 \cdot 0,4 = 0,1$

Saída: $o_1 = 0$

Erro: $erro_1 = 0 - 0$

Calculando os termos de erro:

$\delta_0 = \eta \cdot erro \cdot 1 = 0,5 \cdot 0 \cdot 1 = 0$

$\delta_1 = \eta \cdot erro \cdot X1 = 0,5 \cdot 0 \cdot 0 = 0$

$\delta_2 = \eta \cdot erro \cdot X2 = 0,5 \cdot 0 \cdot 0 = 0$

$\delta_3 = \eta \cdot erro \cdot X2 = 0,5 \cdot 0 \cdot 0 = 0$

Nenhum peso será atualizado (os pesos só são atualizados quando há erro)

Para o segundo exemplo x=(0, 0, 1), t=1

Função soma: $1 \cdot 0,1 - 0 \cdot 0,1 + 0 \cdot 0,2 + 1 \cdot 0,4 = 0,5$

Saída: $o_2 = 0$ (função soma $\leq 0,5$)

Erro: $erro_2 = 1 - 0 = 1$

Calculando os termos de erros

$\delta_0 = \eta \cdot erro \cdot 1 = 0,5 \cdot 1 \cdot 1 = 0,5$

$\delta_1 = \eta \cdot erro \cdot X1 = 0,5 \cdot 1 \cdot 0 = 0$

$\delta_2 = \eta \cdot erro \cdot X2 = 0,5 \cdot 1 \cdot 0 = 0$

$\delta_3 = \eta \cdot erro \cdot X2 = 0,5 \cdot 1 \cdot 1 = 0,5$

Atualizando os pesos

$w_0 \leftarrow w_0 + \delta_0$ $w_0 \leftarrow 0,1 + 0,5 = 0,6$

$w_3 \leftarrow w_3 + \delta_3$ $w_3 \leftarrow 0,4 + 0,5 = 0,9$

Novos pesos

$w_0=0,6$ $w_1=-0,1$ $w_2=0,2$ $w_3=0,9$

Para o terceiro exemplo x=(0, 1, 0), t=0

Função soma: $1 \cdot 0,6 - 0 \cdot 0,1 + 1 \cdot 0,2 + 0 \cdot 0,9 = 0,9$

Saída: $o_3 = 1$ (função soma > 0,5)

Erro: $erro_3 = 0 - 1 = -1$

Calculando os termos de erros

$\delta_0 = \eta \cdot erro \cdot 1 = 0,5 \cdot (-1) \cdot 1 = -0,5$

$\delta_1 = \eta \cdot erro \cdot X1 = 0,5 \cdot (-1) \cdot 0 = 0$

$\delta_2 = \eta \cdot erro \cdot X2 = 0,5 \cdot (-1) \cdot 1 = -0,5$

$\delta_3 = \eta \cdot erro \cdot X2 = 0,5 \cdot (-1) \cdot 0 = 0$

Atualizando os pesos

$w_0 \leftarrow w_0 + \delta_0$ $w_0 \leftarrow 0,6 - 0,5 = 0,1$

$w_2 \leftarrow w_2 + \delta_2$ $w_2 \leftarrow 0,2 - 0,5 = -0,3$

Novos pesos

$w_0=0,1$ $w_1=-0,1$ $w_2=-0,3$ $w_3=0,9$

Para o quarto exemplo x=(0, 1, 1), t=1
Função soma: $1\cdot0{,}1 - 0\cdot0{,}1 - 1\cdot0{,}3 + 1\cdot0{,}9 = 0{,}7$
Saída: $o_4 = 1$ (função soma > 0,7)
Erro: $erro_4 = 1 - 1 = 0$
Todos os termos de erro serão $\delta_i = 0$
Pesos permanecem
$w_0=0{,}1$ $w_1=-0{,}1$ $w_2=-0{,}3$ $w_3=0{,}9$

Para o quinto exemplo x=(1, 0, 0), t=0
Função soma: $1\cdot0{,}1 - 1\cdot0{,}1 - 0\cdot0{,}3 + 0\cdot0{,}9 = 0$
Saída: $o_5 = 0$
Erro: $erro_5 = 0 - 0 = 0$
Todos os termos de erro serão $\delta_i = 0$
Pesos permanecem
$w_0=0{,}1$ $w_1=-0{,}1$ $w_2=-0{,}3$ $w_3=0{,}9$

Para o sexto exemplo x=(1, 0, 1), t=1
Função soma: $1\cdot0{,}1 - 1\cdot0{,}1 - 0\cdot0{,}3 + 1\cdot0{,}9 = 0{,}9$
Saída: $o_6 = 1$
Erro: $erro_6 = 1 - 1 = 0$
Todos os termos de erro serão $\delta_i = 0$
Pesos permanecem
$w_0=0{,}1$ $w_1=-0{,}1$ $w_2=-0{,}3$ $w_3=0{,}9$

Para o sétimo exemplo x=(1, 1, 0), t=1
Função soma: $1\cdot0{,}1 - 1\cdot0{,}1 - 1\cdot0{,}3 + 0\cdot0{,}9 = -0{,}3$
Saída: $o_7 = 0$
Erro: $erro_7 = 1 - 0 = 1$

Calculando os termos de erros
$\delta_0 = \eta\cdot erro\cdot1 = 0{,}5\cdot1\cdot1 = 0{,}5$
$\delta_1 = \eta\cdot erro\cdot X1 = 0{,}5\cdot1\cdot1 = 0{,}5$
$\delta_2 = \eta\cdot erro\cdot X2 = 0{,}5\cdot1\cdot1 = 0{,}5$
$\delta_3 = \eta\cdot erro\cdot X2 = 0{,}5\cdot1\cdot0 = 0$

Atualizando os pesos
$w_0 \leftarrow w_0 + \delta_0$ $w_0 \leftarrow 0{,}1 + 0{,}5 = 0{,}6$
$w_1 \leftarrow w_1 + \delta_1$ $w_1 \leftarrow -0{,}1 + 0{,}5 = 0{,}4$
$w_2 \leftarrow w_2 + \delta_2$ $w_2 \leftarrow -0{,}3 + 0{,}5 = 0{,}2$

Novos pesos
$w_0=0{,}6$ $w_1=0{,}4$ $w_2=0{,}2$ $w_3=0{,}9$

Para o oitavo exemplo x=(1, 1, 1), t=1
Função soma: $1\cdot0{,}6 + 1\cdot0{,}4 + 1\cdot0{,}2 + 1\cdot0{,}9 = 2{,}1$
Saída: $o_8 = 1$
Erro: $erro_8 = 1 - 1 = 0$
Todos os termos de erro serão $\delta_i = 0$
Pesos permanecem $w_0=0{,}6$ $w_1=0{,}4$ $w_2=0{,}2$ $w_3=0{,}9$

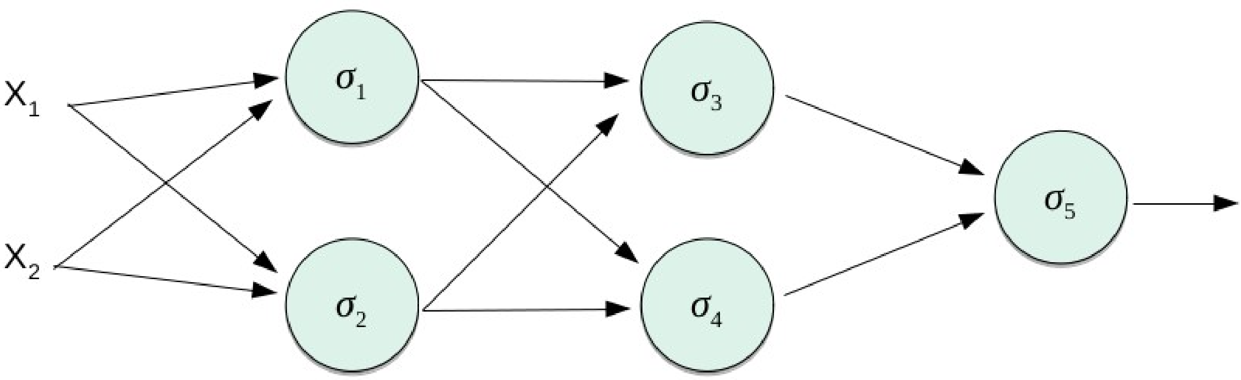
4. Desenhe a rede neural representada pelos pesos abaixo. Faça uma única iteração feed-forward e propagação retrógada para um exemplo cujos valores são $x_1$=1 e $x_2$=0,5 e cuja saída esperada é t=0.

   Primeira camada:
   $w_{01}$ = -1,55 $w_{11}$ = 6,64 $w_{21}$ = -15
   $w_{02}$ = -6,37 $w_{12}$ = -8,98 $w_{22}$ = -19

   Segunda camada:
   $w_{03}$ = 0,76 $w_{13}$ = 2,4 $w_{23}$ = 9
   $w_{04}$ = 10,1 $w_{14}$ = -28 $w_{24}$ = -20

   Terceira camada:
   $w_{05}$ = -0,5 $w_{15}$ = 0,24 $w_{25}$ = 32

A rede pode ser representada pelo seguinte grafo




**Primeira parte: forward**

$net_1 = w_{01} + X_1 \cdot w_{11} + X_2 \cdot w_{21} = 16{,}034$
$\sigma_1 = 1 / (1 + \exp(-net_1)) = 1$

$net_2 = w_{02} + X_1 \cdot w_{12} + X_2 \cdot w_{22} = -38{,}008$
$\sigma_2 = 1 / (1 + \exp(-net_2)) = 0$

$net_3 = w_{03} + \sigma_1 \cdot w_{13} + \sigma_2 \cdot w_{23} = 3{,}16$
$\sigma_3 = 1 / (1 + \exp(-net_3)) = 0{,}9593$

$net_4 = w_{04} + \sigma_1 \cdot w_{14} + \sigma_2 \cdot w_{24} = -17{,}9$
$\sigma_4 = 1 / (1 + \exp(-net_4)) = 0$

$net_5 = w_{05} + \sigma_3 \cdot w_{15} + \sigma_4 \cdot w_{25} = -0{,}2698$
$\sigma_5 = 1 / (1 + \exp(-net_5)) = 0{,}4329$

**Segunda parte: propagação retrógada**

Para cada neurônio, vamos calcular o erro e o termo de erro, que é a parte comum do gradiente para todas as derivadas parciais. Recorde que, para neurônios das camadas escondidas, o erro é ponderado pelos termos de erro dos neurônios a frente.

Camada de saída

$erro_5 = t - \sigma_5 = 0 - 0{,}43296 = -0{,}43296$
$\delta_5 = \sigma_5(1 - \sigma_5)erro_5 = -0{,}1063$

Segunda camada:

$erro_3 = \delta_5 \cdot w_{15} = -0{,}1063 \cdot 0{,}24 = -0{,}0255$
$\delta_3 = \sigma_3(1 - \sigma_3)erro_3 = 0{,}9593(1 - 0{,}9593)(-0{,}0255) = -0{,}000996$

$erro_4 = \delta_5 \cdot w_{25} = -0{,}1063 \cdot 32 = -3{,}4016$
$\delta_4 = \sigma_4(1 - \sigma_4)erro_4 = 0(1 - 0)(-3{,}4016) = 0$

Primeira camada:

$erro_1 = \delta_3 \cdot w_{13} + \delta_4 \cdot w_{14} = -0{,}000996 \cdot 2{,}4 + 0 \cdot (-28) = 0{,}0024$
$\delta_1 = \sigma_1(1 - \sigma_1)erro_1 = 1(1 - 1)0{,}0024 = 0$

$erro_2 = \delta_3 \cdot w_{23} + \delta_4 \cdot w_{24} = -0{,}000996 \cdot 9 + 0 \cdot (-20) = -0{,}009$
$\delta_2 = \sigma_2(1 - \sigma_2)erro_2 = 0(1 - 0)(-0{,}009) = 0$

Atualização dos pesos (como o exercício não especificou uma taxa de aprendizado, vamos definir $\eta$=0,1)

$w_{05} \leftarrow w_{05} + \eta\delta_5$ $\quad w_{05} \leftarrow -0{,}5 + 0{,}1 \cdot (-0{,}1063) = -0{,}51063$
$w_{15} \leftarrow w_{15} + \eta\delta_5\sigma_3$ $\quad w_{15} \leftarrow 0{,}24 + 0{,}1 \cdot (-0{,}1063) \cdot (0{,}9593) = 0{,}2812$
$w_{25} \leftarrow w_{25} + \eta\delta_5\sigma_4$ $\quad w_{25} \leftarrow 32 + 0{,}1 \cdot (-0{,}1063) \cdot 0 = 32$

$w_{04} \leftarrow w_{04} + \eta\delta_4$ $\quad w_{04} \leftarrow 10{,}1 + 0{,}1 \cdot 0 = 10{,}1$
$w_{14} \leftarrow w_{14} + \eta\delta_4\sigma_1$ $\quad w_{14} \leftarrow -28 + \eta \cdot 0 \cdot \sigma_1 = -28$
$w_{24} \leftarrow w_{24} + \eta\delta_4\sigma_2$ $\quad w_{24} \leftarrow -20 + \eta \cdot 0 \cdot \sigma_2 = -20$

$w_{03} \leftarrow w_{03} + \eta\delta_3$ $\quad w_{03} \leftarrow 0{,}76 + 0{,}1 \cdot (-0{,}000996) = -0{,}7599$
$w_{13} \leftarrow w_{13} + \eta\delta_3\sigma_1$ $\quad w_{13} \leftarrow 2{,}4 + 0{,}1 \cdot (-0{,}000996) \cdot 1 = 2{,}3999$
$w_{23} \leftarrow w_{23} + \eta\delta_3\sigma_2$ $\quad w_{23} \leftarrow 9 + 0{,}1 \cdot (-0{,}000996) \cdot 1 = 8{,}9999$

$w_{02} \leftarrow w_{02} + \eta\delta_2$ $\quad w_{02} \leftarrow -6{,}37 + \eta \cdot 0 = -6{,}37$
$w_{12} \leftarrow w_{12} + \eta\delta_2 X_1$ $\quad w_{12} \leftarrow -8{,}98 + \eta \cdot 0 \cdot X_1 = -8{,}98$
$w_{22} \leftarrow w_{22} + \eta\delta_2 X_2$ $\quad w_{22} \leftarrow -19 + \eta \cdot 0 \cdot X_2 = -19$

$w_{01} \leftarrow w_{01} + \eta\delta_1$ $\quad w_{01} \leftarrow -1{,}55 + \eta \cdot 0 = -1{,}55$
$w_{11} \leftarrow w_{11} + \eta\delta_1 X_1$ $\quad w_{11} \leftarrow 6{,}64 + \eta \cdot 0 \cdot X_1 = 6{,}64$
$w_{21} \leftarrow w_{21} + \eta\delta_1 X_2$ $\quad w_{21} \leftarrow -15 + \eta \cdot 0 \cdot X_2 = -15$