Exercícios Deep Learning
Aula 1

June 24, 2019

1 Retas

1- Esboce num gráfico as seguintes retas:

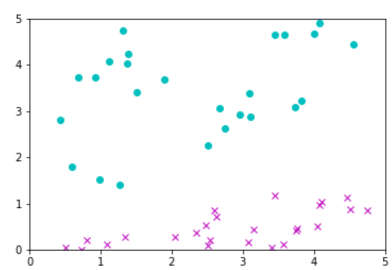
a) $2x_2 + x_1 = 0$

b) $x_2 - 2x_1 + 1 = 0$

c) $x_2 - 1 = 0$

d) $x_1 - 1 = 0$

2- Especifique uma reta que divide as duas categorias de itens no gráfico abaixo onde x_1 é o eixo abscissas e e x_2 é o eixo das ordenadas.



3- Considere a reta $x_2 = 3 + 2x_1$. Obtenha a expressão analítica do conjunto de todas as retas paralelas e o conjunto de todas as retas perpendiculares à reta acima.

2 Álgebra Linear

4- Sejam $w = w_1, ...w_n$ e $x = x_1, ..., x_n$ vetores coluna de dimensão $n \times 1$. Expresse $w'x$ em termos de um somatório.

5- Seja $x = (x_1, ..., x_n)$ um vetor-coluna $n \times 1$ e A uma matriz $n \times n$. A' indica a matriz transposta de A. Verifique que as seguintes identidades matriciais estão corretas, checando se o lado direito é igual ao lado esquerdo.

a) $x'Ax = \sum_{i,j} x_i x_j A_{ij}$

b) $x'x = \sum_i x_i^2$

c) xx' é uma matriz simétrica $n \times n$ com elemento (i, j) dado por $x_i x_j$

3 Derivadas

6- Encontre a derivada $F'(x)$ de

$$F(x) = \sqrt{x^2 + 1}$$

7- Encontre a derivada $F'(x)$ de

$$F(x) = e^{\sin x}$$

8- Função sigmóide:

$$S(x) = \frac{1}{1 + e^{-x}}$$

a) Esboce o gráfico de $S(x)$.

b) Mostre que $S(-x) = 1 - S(x)$

c) Calcule a derivada em termos da própria sigmóide, isto é, mostre que $S'(x) = S(x)(1 - S(x))$. Esboce o gráfico da derivada.

d) Qual o valor máximo de $S'(x)$? Para qual valor de x ela atinge esse máximo?

e) Considere $S(z) = \frac{1}{1+e^{-z}}$, sendo que $z = b + W_1 x$ encontre $\frac{\partial S}{\partial b}$ e $\frac{\partial S}{\partial W_1}$.

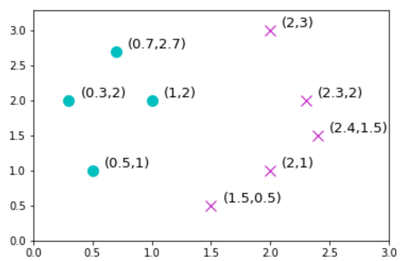
f) Considere $S(h(x)) = \frac{1}{1+e^{-h(x)}}$, calcule $\frac{\partial S}{\partial x}$ em função de $h(x)$.

9- Suponha que você tenha dados da forma $\mathbf{X} = (X_1, X_2, ..., X_n)$, onde $X_i \in \mathbb{R}$ e que seu classificador seja da forma $\hat{Y}_i = \beta X_i$. Considerando o erro quadrático, ou seja, $L(\hat{Y}, Y) = \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$, qual o valor de β que minimiza o erro?

4 Perceptron

10- Considere um perceptron com duas features x_1, x_2, onde $w_0 = 3, w_1 = 2, w_2 = 1$, qual é a fórmula da reta que divide as duas classes? Qual é o vetor normal à reta?

11- Considere um perceptron com $w_0 = 0, w_1 = -1, w_2 = 1$, com os pontos mostrados no gráfico abaixo (círculos pertencem à classe 1 e 'x' à classe 0), esboce o gráfico da reta de separação das classes. Execute uma iteração do algoritmo do perceptron com $m = 0.1$, esboce a nova reta de separação.



12- Seja o conjunto de entrada dado por um total de 4 amostras, onde cada amostra é representada pela tupla $(\boldsymbol{x_i}, \boldsymbol{t})$, composta pelo vetor $\boldsymbol{x_i} = (x_0, x_1, x_2)$ e um rotulo $\boldsymbol{t}$ associado a amostra.

	x_0	x_1	x_2	t
Entrada 1	1	0	0	0
Entrada 2	1	0	1	0
Entrada 3	1	1	0	0
Entrada 4	1	1	1	1

a) Execute a quinta iteração do algoritmo do perceptron com pesos iniciais $w_0, w_1 e w_2$ iguais a 0, taxa de aprendizado η igual a 0.5, e utilizando a função de ativação degrau bipolar. Abaixo segue o exemplo das quatro primeiras iterações do algoritmo:

1º Iteração:

Entrada 1:

$$\begin{aligned} s_{out} &= f(w_0 x_0 + w_1 x_1 + w_2 x_2) \\ &= f(0 * 1 + 0 * 0 + 0 * 0) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{1}$$

Entrada 2:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0*1 + 0*0 + 0*1) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{2}$$

Entrada 3:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0*1 + 0*1 + 0*0) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{3}$$

Entrada 4:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0*1 + 0*1 + 0*1) = f(0) = 0;\ logo\ s_{out} \neq t \end{aligned} \tag{4}$$

Atualiza Pesos:

$$\begin{aligned} w_0 &= w_0 + \eta(t - s_{out}) * x_0 = 0 + 0.5(1-0)*1 = 0.5 \\ w_1 &= w_1 + \eta(t - s_{out}) * x_1 = 0 + 0.5(1-0)*1 = 0.5 \\ w_2 &= w_2 + \eta(t - s_{out}) * x_2 = 0 + 0.5(1-0)*1 = 0.5 \end{aligned}$$

2º Iteração:
Entrada 1:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0.5*1 + 0.5*0 + 0.5*0) = f(0.5) = 1;\ logo\ s_{out} \neq t \end{aligned} \tag{5}$$

Atualiza Pesos:

$$\begin{aligned} w_0 &= w_0 + \eta(t - s_{out}) * x_0 = 0.5 + 0.5(0-1)*1 = 0 \\ w_1 &= w_1 + \eta(t - s_{out}) * x_1 = 0.5 + 0.5(0-1)*0 = 0.5 \\ w_2 &= w_2 + \eta(t - s_{out}) * x_2 = 0.5 + 0.5(0-1)*0 = 0.5 \end{aligned}$$

Entrada 2:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0*1 + 0.5*0 + 0.5*1) = f(0.5) = 1;\ logo\ s_{out} \neq t \end{aligned} \tag{6}$$

Atualiza Pesos:

$$\begin{aligned} w_0 &= w_0 + \eta(t - s_{out}) * x_0 = 0 + 0.5(0-1)*1 = -0.5 \\ w_1 &= w_1 + \eta(t - s_{out}) * x_1 = 0.5 + 0.5(0-1)*0 = 0.5 \\ w_2 &= w_2 + \eta(t - s_{out}) * x_2 = 0.5 + 0.5(0-1)*1 = 0 \end{aligned}$$

Entrada 3:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-0.5*1 + 0.5*1 + 0*0) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{7}$$

Entrada 4:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-0.5*1 + 0.5*1 + 0*1) = f(0) = 0;\ logo\ s_{out} \neq t \end{aligned} \tag{8}$$

Atualiza Pesos:

$$w_0 = w_0 + \eta(t - s_{out}) * x_0 = -0.5 + 0.5(1-0)*1 = 0$$
$$w_1 = w_1 + \eta(t - s_{out}) * x_1 = 0.5 + 0.5(1-0)*1 = 1$$
$$w_2 = w_2 + \eta(t - s_{out}) * x_2 = 0 + 0.5(1-0)*1 = 0.5$$

3º Iteração:
Entrada 1:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0*1 + 1*0 + 0.5*0) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{9}$$

Entrada 2:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(0*1 + 1*0 + 0.5*1) = f(0.5) = 1;\ logo\ s_{out} \neq t \end{aligned} \tag{10}$$

Atualiza Pesos:

$$w_0 = w_0 + \eta(t - s_{out}) * x_0 = 0 + 0.5(0-1)*1 = -0.5$$
$$w_1 = w_1 + \eta(t - s_{out}) * x_1 = 1 + 0.5(0-1)*0 = 1$$
$$w_2 = w_2 + \eta(t - s_{out}) * x_2 = 0.5 + 0.5(0-1)*1 = 0$$

Entrada 3:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-0.5*1 + 1*1 + 0*0) = f(0.5) = 1;\ logo\ s_{out} \neq= t \end{aligned} \tag{11}$$

Atualiza Pesos:

$$w_0 = w_0 + \eta(t - s_{out}) * x_0 = -0.5 + 0.5(0-1)*1 = -1$$
$$w_1 = w_1 + \eta(t - s_{out}) * x_1 = 1 + 0.5(0-1)*0 = 1$$
$$w_2 = w_2 + \eta(t - s_{out}) * x_2 = 0.5 + 0.5(0-1)*1 = 0$$

Entrada 4:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-1*1 + 1*1 + 0*1) = f(0) = 0;\ logo\ s_{out} \neq t \end{aligned} \tag{12}$$

Atualiza Pesos:

$$w_0 = w_0 + \eta(t - s_{out}) * x_0 = -1 + 0.5(1-0)*1 = -0.5$$
$$w_1 = w_1 + \eta(t - s_{out}) * x_1 = 1 + 0.5(1-0)*1 = 1.5$$
$$w_2 = w_2 + \eta(t - s_{out}) * x_2 = 0 + 0.5(1-0)*1 = 0.5$$

4º Iteração:
Entrada 1:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-0.5 * 1 + 1.5 * 0 + 0.5 * 0) = f(-0.5) = 0;\ logo\ s_{out} = t \end{aligned} \tag{13}$$

Entrada 2:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-0.5 * 1 + 1.5 * 0 + 0.5 * 1) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{14}$$

Entrada 3:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-0.5 * 1 + 1.5 * 1 + 0.5 * 0) = f(1) = 1;\ logo\ s_{out} \neq= t \end{aligned} \tag{15}$$

Atualiza Pesos:

$$\begin{aligned} w_0 &= w_0 + \eta(t - s_{out}) * x_0 = -0.5 + 0.5(0-1) * 1 = -1 \\ w_1 &= w_1 + \eta(t - s_{out}) * x_1 = 1.5 + 0.5(0-1) * 0 = 1 \\ w_2 &= w_2 + \eta(t - s_{out}) * x_2 = 0.5 + 0.5(0-1) * 1 = 0.5 \end{aligned}$$

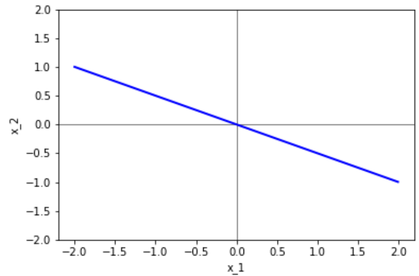
Entrada 4:

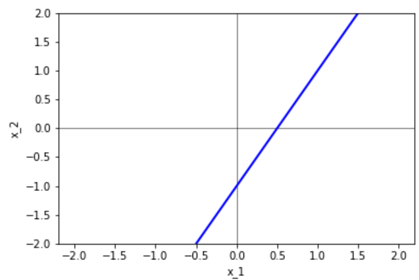
$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-1 * 1 + 1 * 1 + 0.5 * 1) = f(0.5) = 1;\ logo\ s_{out} = t \end{aligned} \tag{16}$$

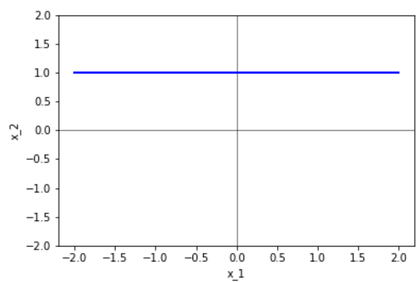
b) Esboce o gráfico da reta gerada após a quinta iteração do algoritmo. A equação da reta após a quinta iteração é dado pela equação: $x_1w_1 + x_2w_2 = -w_0$

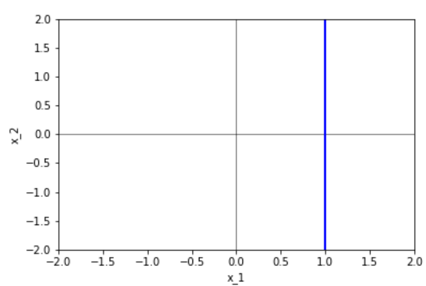
Solução

1-









2- $x_2 - \frac{1}{2}x_1 = 0$

3- Paralelas : $x_2 = 2x_1 + c$ com $c \in \mathbb{R}$
Perpendiculares: $x_2 = -\frac{1}{2}x_1 + c$ com $c \in \mathbb{R}$

4- $\sum_{i=1}^{n} w_i x_i$

5- a)

$$\begin{aligned}\mathbf{x'Ax} &= \begin{bmatrix}\sum_j x_1 A_{1,j} & \cdots & \sum_j x_n A_{n,j}\end{bmatrix} \mathbf{x} \\ &= \sum_{i,j} x_i A_{i,j} x_j\end{aligned}$$

b)

$$\begin{aligned}\mathbf{x'x} &= \begin{bmatrix}x_1 & \cdots & x_n\end{bmatrix} \begin{bmatrix}x_1 \\ \vdots \\ x_n\end{bmatrix} \\ &= \sum_i x_i^2\end{aligned}$$

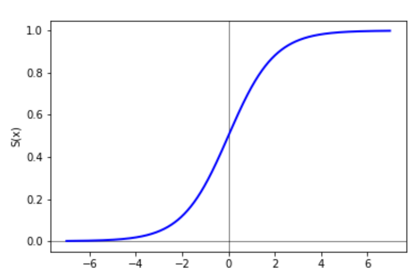
c)

$$\begin{aligned}\mathbf{xx'} &= \begin{bmatrix}x_1 \\ \vdots \\ x_n\end{bmatrix} \begin{bmatrix}x_1 & \cdots & x_n\end{bmatrix} \\ &= \begin{bmatrix}x_1x_1 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2x_2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_nx_n\end{bmatrix}\end{aligned}$$

6- $x(x^2+1)^{-\frac{1}{2}}$

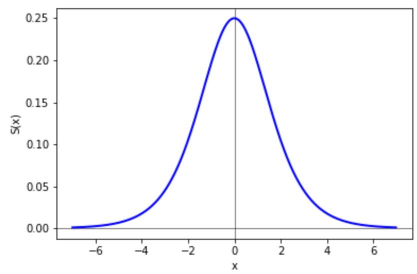
7- $\cos x e^{\sin x}$

8- a)



b) $S(-x) = \frac{1}{1+e^{x}} = \frac{1}{1+\frac{1}{e^{-x}}} = \frac{1}{\frac{e^{-x}+1}{e^{-x}}} = \frac{e^{-x}}{e^{-x}+1} = 1 - \frac{1}{1+e^{-x}} = 1 - S(x)$

c) $S'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} = S(x)(1 - S(x))$



d) Valor máximo é 0.25 quando $x = 0$.

e) $\frac{\partial S}{\partial b} = \frac{e^{-(b+W_1x)}}{(1+e^{-(b+W_1x)})^2}$ e $\frac{\partial S}{\partial W_1} = \frac{xe^{-(b+W_1x)}}{(1+e^{-(b+W_1x)})^2}$

f) $\frac{\partial S}{\partial x} = \frac{h'(x)e^{-h(x)}}{(1+e^{-h(x)})^2}$

9- Para encontrar o valor mínimo da perda, devemos derivar a função em relação a β e encontrar onde ela é 0. $L(\hat{Y}, Y) = \sum_{i=1}^{n}(\beta X_i - Y_i)^2$

$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} 2X_i(\beta X_i - Y_i) = 2(\sum_{i=1}^{n} \beta X_i^2 - \sum_{i=1}^{n} X_i Y_i) = 0$

$\beta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$

10- Reta: $x_2 = -3 - 2x_1$ Vetor normal: (2 1)

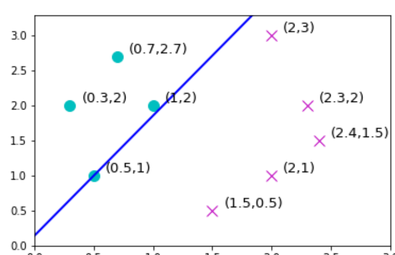
11- Reta: $x_2 = x_1$

Algoritmo:

O único ponto onde $y \neq \hat{y}$ é $(2, 3)$

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} + 0.1(-1)\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -0.1 \\ -1.2 \\ 0.7 \end{bmatrix}$$

Nova reta : $x_2 = \frac{1}{7} + \frac{12}{7}x_1$



12-

a)

5º Iteração:

Entrada 1:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-1*1+1*0+0.5*0) = f(-1) = 0;\ logo\ s_{out} = t \end{aligned} \tag{17}$$

Entrada 2:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-1*1+1*0+0.5*1) = f(-0.5) = 0;\ logo\ s_{out} = t \end{aligned} \tag{18}$$

Entrada 3:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-1*1+1*1+0.5*0) = f(0) = 0;\ logo\ s_{out} = t \end{aligned} \tag{19}$$

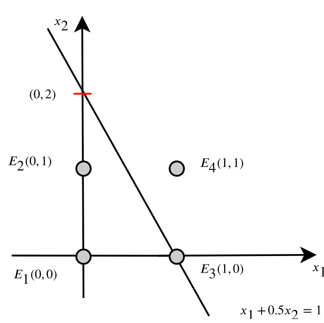
Entrada 4:

$$\begin{aligned} s_{out} &= f(w_0x_0 + w_1x_1 + w_2x_2) \\ &= f(-1*1+1*1+0.5*1) = f(0.5) = 1;\ logo\ s_{out} = t \end{aligned} \tag{20}$$

Logo temos o resultado: $w_0 = -1, w_1 = 1, w_2 = 0.5$

b)

Esboçando a reta $x_1 * 1 + x_2 * 0.5 = 1$ temos:



Exercícios Deep Learning
Aula 2

June 26, 2019

1 Derivadas

1- Seja $x, y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$ e $\mathbf{y} \in \mathbb{R}^m$ de forma que:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} \quad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}$$

e

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}}{\partial x_1} \\ \frac{\partial \mathbf{y}}{\partial x_2} \\ \vdots \\ \frac{\partial \mathbf{y}}{\partial x_3} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ & & \vdots & \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Assuma que $\mathbf{y} = f(\mathbf{u})$ e $\mathbf{u} = g(\mathbf{x})$, escreva a derivada (usando a regra da cadeia) para $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

2- Seja $\mathbf{X} \in \mathbb{R}^{m \times n},\ \mathbf{w} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^m$ e $z = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, calcule $\frac{\partial z}{\partial \mathbf{w}}$.

3- Sejam $u, \mathbf{x} \in R^n$ (vetores colunas).

a) Calcule a derivada de $u'\mathbf{x}$ em respeito a $\mathbf{x}$, ou seja: $\frac{\partial (u'\mathbf{x})}{\partial \mathbf{x}}$

b) Calcule a derivada de $\mathbf{x}'\mathbf{x}$ em respeito a $\mathbf{x}$, ou seja: $\frac{\partial (\mathbf{x}'\mathbf{x})}{\partial \mathbf{x}}$

2 Regressão Linear

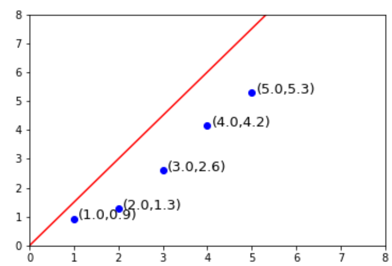
4- Em uma regressão linear, o valor estimado $\hat{Y}_i$ é dado por

$$\hat{Y}_i = b + w_1 x_1 + ... + w_n x_n$$

onde n é o número de features. O erro quadrático é dado por $L(\hat{Y}, Y) = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$. Neste exercício, $n = 1$, portanto $\hat{Y}_i = b + w_1 x_1$.

a) Considerando os pontos mostrados no gráfico abaixo e a reta com $w_1 = 1.5$ e $b = 0$, calcule o erro quadrático dessa reta.

b) Utilize a derivada do erro quadrático para atualizar os valores de w_1 e b, encontrando uma nova reta (Dica: use uma taxa de aprendizado menor que 0.1). Qual é o erro quadrático desta nova reta?



3 Regressão Logística

5- No modelo de regressão logística com um regressor apenas: $P(Y_i = 1) = p(x_i) = \frac{1}{1+\exp(-b-w_1 x_i)}$. Deduza que a log-verossimilhança de $\theta = (b, w_1)$ no caso do modelo logístico com um único regressor é dada por:

$$l(\theta) = b \sum_{i=1}^{n} y_i + w_1 \sum_{i=1}^{n} x_i y_i - \sum_{i} \log(1 + e^{b+w_1 x_i})$$

6- No modelo logístico com um regressor apenas, mostre que o vetor gradiente de $l(\theta)$ é dado por:

$$Dl(\theta) = \begin{bmatrix} \frac{\partial logl}{\partial b} \\ \frac{\partial logl}{\partial w_1} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i - p_i \\ \sum_{i=1}^{n} x_i y_i - p_i x_i \end{bmatrix}$$

onde $p_i = p(x_i)$.

7- Ainda no modelo logístico com um regressor apenas, mostre que a matriz hessiana de $l(\theta)$ é dada por:

$$Dl(\theta) = -\begin{bmatrix} \frac{\partial^2 logl}{\partial b^2} & \frac{\partial^2 logl}{\partial b \partial w_1} \\ \frac{\partial^2 logl}{\partial b \partial w_1} & \frac{\partial^2 logl}{\partial w_1^2} \end{bmatrix} = -\begin{bmatrix} \sum_{i=1}^n p_i(1-p_i) & \sum_{i=1}^n p_i(1-p_i)x_i \\ \sum_{i=1}^n p_i(1-p_i)x_i & \sum_{i=1}^n p_i(1-p_i)x_i^2 \end{bmatrix}$$

4 Notebook

8- Faça download do notebook S01A02 no drive e utilize o collab para completar os exercícios.

5 Newton e SGA

9- Aplique duas iterações do método de newton para encontrar o ponto máximo da função $f(x) = -(x-3)^4$ para $x_0 = 1$.

10- Utilizando a a mesma função do exercicio anterior, aplique o método do gradiente ascendente usando learning rate $\alpha = 0.01$ e $x_0 = 1$. Compare os dois resultados.

Solução

1- $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$

2-

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\mathbf{Xw}\ \textbf{-}\ \mathbf{Y} = \begin{bmatrix} \sum_i x_{i1} w_i - y_1 \\ \sum_i x_{i2} w_i - y_2 \\ \vdots \\ \sum_i x_{im} w_i - y_m \end{bmatrix}$$

$z = \|\mathbf{Xw} - \mathbf{y}\|^2$

$$\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial \sum_{i,j} (x_{ij} w_i - y_j)^2}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial z}{\partial w_1} \\ \frac{\partial z}{\partial w_2} \\ \vdots \\ \frac{\partial z}{\partial w_n} \end{bmatrix} \quad \begin{bmatrix} 2\sum_{ij} (x_{ij} w_i - y_j) \sum_j x_{1j} \\ 2\sum_{ij} (x_{ij} w_i - y_j) \sum_j x_{2j} \\ \vdots \\ 2\sum_{ij} (x_{ij} w_i - y_j) \sum_j x_{mj} \end{bmatrix}$$

3-

a) A derivada de $u'\mathbf{x} = \sum_{i=1}^n u_i x_i$ em respeito a $\mathbf{x}$:

$$\frac{\partial \sum_{i=1}^n u_i x_i}{\partial x_i} = u_i \Rightarrow \frac{\partial u'\mathbf{x}}{\partial \mathbf{x}} = (u_1, \ldots, u_n) = u'$$

b) A derivada de $\mathbf{x}'\mathbf{x} = \sum_{i=1}^n x_i^2$ em respeito a $\mathbf{x}$:

$$\frac{\partial \sum_{i=1}^n x_i^2}{\partial x_i} = 2x_i \Rightarrow \frac{\partial \mathbf{x}'\mathbf{x}}{\partial \mathbf{x}} = (2\mathbf{x}_1, \ldots, 2\mathbf{x}_n) = 2\mathbf{x}'$$

4-

a)

$(1.5 - 0.9)^2 + (3 - 1.3)^2 + (4.5 - 2.6)^2 + (6 - 4.2)^2 + (7.5 - 5.3)^2 = 14.94$

b)

$$\frac{\partial L}{\partial b} = 2\sum_{i=1}^{n}(b + w_1 x_i - Y_i) = 2nb + 2\sum_{i=1}^{n}(w_1 x_i - Y_i)$$

$$\frac{\partial L}{\partial w_1} = 2\sum_{i=1}^{n}\big((b + w_1 x_i - Y_i)x_i\big) = 2b\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n}(w_1 x_i^2 - Y_i x_i)$$

Colocando a taxa de aprendizado $\alpha = 0.01$

$$\begin{bmatrix} w^* \\ b^* \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix} - \alpha \begin{bmatrix} 2\sum_{i=1}^{n}(1.5x_i^2 - Y_i x_i) \\ 2\sum_{i=1}^{n}(1.5x_i - Y_i) \end{bmatrix} = \begin{bmatrix} 0.942 \\ -0.164 \end{bmatrix}$$

$$L = (0.78-0.9)^2 + (1.72-1.3)^2 + (2.66-2.6)^2 + (3.6-4.2)^2 + (4.55-5.3)^2 = 1.117$$

5-

$$
\begin{aligned}
l(\theta) &= log\left(\prod_{i=1}^{n}\mathbb{P}(Y_i = y_i)\right)\\
&= log\left(\prod_{i=1}^{n}\mathbb{P}(Y_i = y_i)\right)\\
&= log\left(\prod_{i=1}^{n}p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right)\\
&= \sum_{i=1}^{n} log(p(x_i)^{y_i}(1-p(x_i))^{1-y_i})\\
&= \sum_{i=1}^{n} log(p(x_i)^{y_i}) + \sum_{i=1}^{n} log((1-p(x_i))^{1-y_i})\\
&= \sum_{i=1}^{n} y_i log(p(x_i)) + \sum_{i=1}^{n}(1-y_i)log(1-p(x_i))\\
&= \sum_{i=1}^{n} y_i log\left(\frac{1}{1+exp(-b-w_1x_i)}\right) + \sum_{i=1}^{n}(1-y_i)log\left(1-\frac{1}{1+exp(-b-w_1x_i)}\right)\\
&= \sum_{i=1}^{n} y_i log\left(\frac{exp(b+w_1x_i)}{1+exp(b+w_1x_i)}\right) + \sum_{i=1}^{n}(1-y_i)log\left(\frac{1}{1+exp(b+w_1x_i)}\right)\\
&= \sum_{i=1}^{n} y_i log(exp(b+w_1x_i)) - \sum_{i=1}^{n} y_i log(1+exp(b+w_1x_i)) + \sum_{i=1}^{n}(1-y_i)log\left(\frac{1}{1+exp(b+w_1x_i)}\right)\\
&= \sum_{i=1}^{n} y_i log(exp(b+w_1x_i)) - \sum_{i=1}^{n} y_i log(1+exp(b+w_1x_i)) - \sum_{i=1}^{n}(1-y_i)log(1+exp(b+w_1x_i))\\
&= \sum_{i=1}^{n} y_i(b+w_1x_i) - \sum_{i=1}^{n} y_i log(1+exp(b+w_1x_i)) - \sum_{i=1}^{n}(1-y_i)log(1+exp(b+w_1x_i))\\
&= b\sum_{i=1}^{n} y_i + w_1\sum_{i=1}^{n} y_ix_i - \sum_{i=1}^{n} y_i log(1+exp(b+w_1x_i)) - \sum_{i=1}^{n}(1-y_i)log(1+exp(b+w_1x_i))\\
&= b\sum_{i=1}^{n} y_i + w_1\sum_{i=1}^{n} y_ix_i - \sum_{i=1}^{n} log(1+exp(b+w_1x_i))
\end{aligned}
$$

6-

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial b} &= \frac{\partial}{\partial b}\left(b\sum_{i=1}^{n} y_i + w_1\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} log(1+exp(b+w_1x_i))\right)\\
&= \frac{\partial}{\partial b}\left(b\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} log(1+exp(b+w_1x_i))\right)\\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\partial}{\partial b}(log(1+exp(b+w_1x_i)))\\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\frac{\partial}{\partial b}(1+exp(b+w_1x_i))}{1+exp(b+w_1x_i)}\\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{exp(b+w_1x_i)}{1+exp(b+w_1x_i)}\\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} p_i\\
&= \sum_{i=1}^{n} (y_i - p_i)\\
\frac{\partial l(\theta)}{\partial w_1} &= \frac{\partial}{\partial w_1}\left(b\sum_{i=1}^{n} y_i + w_1\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} log(1+exp(b+w_1x_i))\right)\\
&= \frac{\partial}{\partial w_1}\left(w_1\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} log(1+exp(b+w_1x_i))\right)\\
&= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\partial}{\partial w_1}(log(1+exp(b+w_1x_i)))\\
&= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\frac{\partial}{\partial w_1}(1+exp(b+w_1x_i))}{1+exp(b+w_1x_i)}\\
&= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i exp(b+w_1x_i)}{1+exp(b+w_1x_i)}\\
&= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i p_i\\
&= \sum_{i=1}^{n} (y_i x_i - x_i p_i)
\end{aligned}
$$

7-

$$\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial b^2} &= \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} p_i\right) \\
&= \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial b}\left(\frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} \left(-\frac{-exp(-b-w_1x_i)}{(1+exp(-b-w_1x_i))^2}\right) \\
&= -\sum_{i=1}^{n} \left(\frac{exp(-b-w_1x_i)}{(1+exp(-b-w_1x_i))^2}\right) \\
&= -\sum_{i=1}^{n} \left(\frac{exp(-b-w_1x_i)}{1+exp(-b-w_1x_i)}\frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} \left(\frac{1}{1+exp(b+w_1x_i)}\frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} (1-p_i)p_i
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial b \partial w_1} = \frac{\partial^2 l(\theta)}{\partial w_1 \partial b} &= \frac{\partial}{\partial w_1}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} p_i\right) \\
&= \frac{\partial}{\partial w_1}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial w_1}\left(\frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n}\left(-\frac{-x_i exp(-b-w_1x_i)}{(1+exp(-b-w_1x_i))^2}\right) \\
&= -\sum_{i=1}^{n}\left(\frac{x_i exp(-b-w_1x_i)}{(1+exp(-b-w_1x_i))^2}\right) \\
&= -\sum_{i=1}^{n}\left(\frac{x_i exp(-b-w_1x_i)}{1+exp(-b-w_1x_i)} \frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n}\left(x_i \frac{1}{1+exp(b+w_1x_i)} \frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} x_i(1-p_i)p_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial w_1^2} &= \frac{\partial}{\partial w_1}\left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i p_i\right) \\
&= \frac{\partial}{\partial w_1}\left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i p_i\right) \\
&= \frac{\partial}{\partial w_1}\left(-\sum_{i=1}^{n} x_i p_i\right) \\
&= \frac{\partial}{\partial w_1}\left(-\sum_{i=1}^{n} x_i \frac{1}{1+exp(-b-w_1x_i)}\right) \\
&= -\sum_{i=1}^{n} x_i \frac{x_i exp(-b-w_1x_i)}{(1+exp(-b-w_1x_i))^2} \\
&= -\sum_{i=1}^{n} x_i x_i \frac{exp(-b-w_1x_i)}{1+exp(-b-w_1x_i)} \frac{1}{1+exp(-b-w_1x_i)} \\
&= -\sum_{i=1}^{n} x_i^2(1-p_i)p_i
\end{aligned}
$$

8- Jupyter notebook

9-
Seja $x_0 = 1$ e f(x) e suas derivadas:

$$\begin{aligned} f(x) &= -(x-3)^4 \\ f^{'}(x) &= -4(x-3)^3 \\ f^{''}(x) &= -12(x-3)^2 \end{aligned} \tag{1}$$

Utilizando método de Newton: $x_{n+1} = x_n - \frac{f^{'}(x_n)}{f^{''}(x_n)}$

- Para $n = 0$ e $x_0 = 1$
  $x_1 = x_0 - \frac{f^{'}(x_0)}{f^{''}(x_0)} = 1.66$
- Para $n = 1$ e $x_1 = 1.66$
  $x_2 = x_1 - \frac{f^{'}(x_1)}{f^{''}(x_1)} = 2.11$
- Para $n = 2$ e $x_1 = 2.11$
  $x_3 = x_2 - \frac{f^{'}(x_2)}{f^{''}(x_2)} = 2.407$

10-

Utilizando método do Gradient Ascendente: $x_{n+1} = x_n + \alpha * f^{'}(x_n)$, onde $\alpha = 0.01$

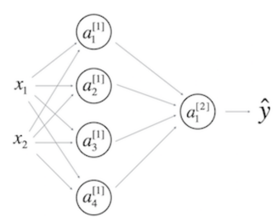
- Para $n = 0$ e $x_0 = 1$
  $x_1 = x_0 + \alpha * f^{'}(x_0) = 1.32$
- Para $n = 1$ e $x_1 = 1.32$
  $x_2 = x_1 + \alpha * f^{'}(x_1) = 1.509$
- Para $n = 2$ e $x_2 = 1.509$
  $x_3 = x_2 + \alpha * f^{'}(x_2) = 1.64$

Exercícios Deep Learning
Aula 3

June 26, 2019

1 Redes Neurais

1- Considere a rede com uma camada escondida mostrada abaixo:



Assumindo que cada neurônio de uma camada [l] aplica uma função $f(Z^{[l]}) = A^{[l]}$ onde Z é calculado como $W^{[l]}A^{[l-1]} + b^{[l]}$ e que a entrada pode ser escrita como a camada zero: $A[0] = \mathbf{x}$.

a) Quais são as dimensões de $W^{[1]}$, $b^{[1]}$, $W^{[2]}$ e $b^{[2]}$?

b) Quais são as dimensões de $Z^{[1]}$ e $A^{[1]}$?

2- Além da sigmóide, a tangente hiperbólica e ReLU são funções de ativação comumente utilizadas em redes neurais.

a) Encontre a derivada da tangente hiperbólica $\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ com respeito a x. Em qual valor de x a derivada atinge seu valor máximo?

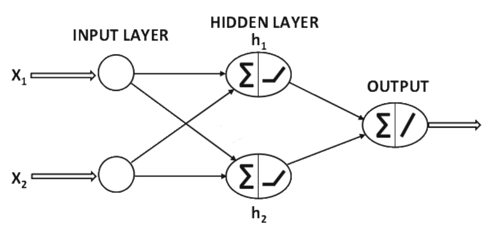
b) Encontre a derivada da função $ReLU = \max(0, x)$ para $x < 0$ e $x > 0$.

c) Sejam a,b constantes positivas. Expresse a derivada da função $f(x) = a * tanh(bx)$ (sigmoide anti-simétrica) em termos da própria função.

3- Considere uma rede neural com dois inputs X_1 e X_2 e com duas camadas escondidas, cada uma com dois nós. Assuma que os pesos estão distribuídos de forma que os nós em cima na camada aplicam a sigmóide na soma dos seus inputs e que os nós em baixo aplicam a função tanh em seus inputs. O nó de saída aplica a ReLU na soma dos dois inputs. Desenhe esta rede. Escreva a saída dessa rede neural como uma função de x_1 e x_2 em forma fechada.

4- Suponha que você tenha uma sequência de dados com duas features binárias, isto é, $X = \{(x_1^{(0)}, x_2^{(0)}), ..., (x_1^{(n)}, x_2^{(n)})\}$, onde $x_1^{(i)}, x_2^{(i)} \in \{0, 1\}$. Imagine uma rede neural com 3 nós, como na figura abaixo, todos utilizam a função de ativação $ReLU(x)$, o último nó (output) deve retornar 0 se o resultado é Falso e um valor maior que zero se o resultado é Verdadeiro. Mostre como você pode usar esta rede para calcular $XOR(x_1, x_2)$. Ou seja, mostre os pesos dos parâmetros dos nós $h1, h2$ e output para calcular o XOR.

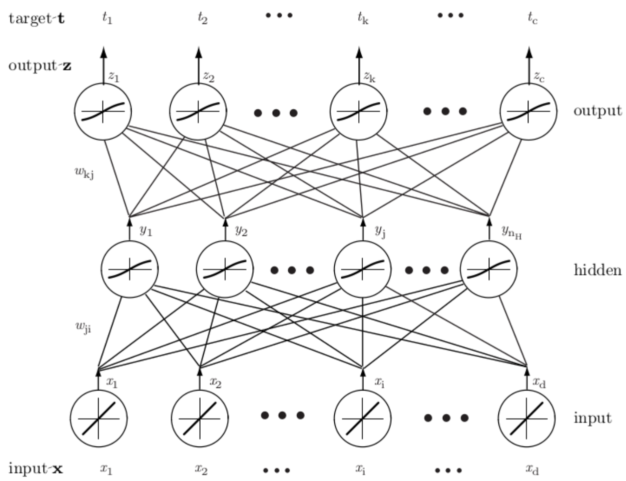
Lembre-se que, o $XOR(x_1, x_2)$, também chamado de *ou exclusivo*, é Verdadeiro quando $x_1 = 1$ OU $x_2 = 1$, porém é Falso quando $x_1 = 1$ E $x_2 = 1$ e quando $x_1 = 0$ E $x_2 = 0$.



5- Suponha que todos os nós de uma rede neural utilizam uma função de ativação linear, isto é, $g(x) = x$. Mostre que, quando utilizada essa função, qualquer rede neural com 2 camadas é equivalente a uma rede de 1 camada. Verifique intuitivamente que esse resultado é válido independentemente do número de camadas (não precisa fazer contas).

Esse tipo de rede consegue classificar corretamente inputs de acordo com o XOR (problema 4)?

6- Considere uma rede padrão de 3 camadas cuja entrada $\mathbf{x}$ possui dimensão $d \times 1$, a primeira camada da rede possui d unidades de entrada e possui somente uma ativação linear do tipo $f(\mathbf{x}) = \mathbf{x}$, a camada escondida possui n_H unidades escondidas e a camada final possui c unidades de saída e o bias.



a) Qual o número total de pesos que existem na rede?

b) Mostre que se o sinal de cada peso da rede for trocado, a função da rede permanece inalterada, caso a função de ativação usada for uma função par. Lembre-se que, uma função é par se e somente se $f(-x) = f(x)$.

Solução

1-

a)
b[1] tem dimensão (4, 1)
W[1] tem dimensão (4, 2)
W[2] tem dimensão (1, 4)
b[2] tem dimensão (1, 1)

b) Z[1] e A[1] tem dimensão (4,m).

2-

a)

$$\begin{aligned}
\tanh x' &= \frac{(e^x - e^{-x})'(e^x + e^{-x}) - (e^x - e^{-x})(e^x + e^{-x})'}{(e^x + e^{-x})^2} \\
&= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \\
&= \frac{e^{2x} + 2e^0 + e^{-2x} - (e^{2x} - 2e^0 + e^{-2x})}{(e^x + e^{-x})^2} \\
&= \frac{4}{(e^x + e^{-x})^2} = \left(\frac{2}{e^x + e^{-x}}\right)^2
\end{aligned}$$

b)
Para $x < 0$:

$$ReLU(x)' = (0)' = 0$$

Para $x > 0$:

$$ReLU(x)' = x' = 1$$

c) Podemos expressar tanh(x) como:

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$$

Seja $g(x) := e^{-2bx}$, temos $g'(x) = -2bg(x)$ e podemos reescrever f(x) da seguinte forma:

$$f(x) = a\frac{1 - e^{-2bx}}{1 + e^{-2bx}} := a\frac{1 - g(x)}{1 + g(x)} = a(1 - g(x))(1 + g(x))^{-1}$$

Derivando f(x) usando a regra da cadeia e do produto, obtemos:

$$\begin{aligned} f'(x) &= \frac{-ag'(x)(1+g(x)) - ag'(x)(1-g(x))}{(1+g(x))^2} \\ &= \frac{4abg(x)}{(1+g(x))^2} = \underbrace{a\frac{1-g(x)}{1+g(x)}}_{f'(x)} \frac{4bg(x)}{(1-g(x))(1+g(x))} \end{aligned}$$

Substituindo de volta $g(x) := e^{-2bx}$, podemos ver que sua derivada pode ser escrita como:

$$f'(x) = f(x)\frac{4be^{-2bx}}{1-e^{-4bx}}$$

3- $max\Big(0, \sigma\big(\sigma(x_1+x_2)+\tanh{(x_1+x_2)}\big)+\tanh{\big(\sigma(x_1+x_2)+\tanh{(x_1+x_2)}\big)}\Big)$

4- Seja $a_{h1} = max(0, b^{h1} + W_1^{h1}x_1 + W_2^{h1}x_2)$ a saída do nó h_1, então os pesos devem ser colocados de forma que somente a saída para entrada $x_1 = 1, x_2 = 0$ seja positiva, por exemplo:

$$b^{h1} = 0, W_1^{h1} = 1, W_2^{h1} = -1$$

Seja $a_{h2} = max(0, b^{h2} + W_1^{h2}x_1 + W_2^{h2}x_2)$ a saída do nó h_2, então os pesos devem ser colocados de forma que somente a saída para entrada $x_1 = 0, x_2 = 1$ seja positiva, por exemplo:

$$b^{h2} = 0, W_1^{h2} = -1, W_2^{h2} = 1$$

(Ou vice-versa)

Seja $a_{out} = max(0, b^{out} + W_1^{out}a_{h1} + W_2^{out}a_{h2})$ a saída da rede (do nó output). Os pesos devem ser distribuídos de forma que a saída seja positiva somente quando a_{h1} ou a_{h2} seja positivo, ou seja, quando $XOR(x_1, x_2)$ é verdadeiro. Por exemplo:

$$b^{h2} = 0, W_1^{h2} = 1, W_2^{h2} = 1$$

5- O nó de output na rede de duas camadas receberá a seguinte função:

$$b^{[2]} + W^{[2]}(b^{[1]} + W^{[1]}X)$$

$$= (b^{[2]} + W^{[2]}b^{[1]}) + (W^{[2]}W^{[1]})X$$

Seja $b^{[1]*} = (b^{[2]} + W^{[2]}b^{[1]})$ e $W^{[1]*} = (W^{[2]}W^{[1]})$, então se uma rede de uma camada estimar $b^{[1]*}$ e $W^{[1]*}$ como seus parâmetros, então ela gerará o mesmo resultado da rede de duas camadas acima, logo elas são equivalentes.

A rede que utiliza função de ativação linear não consegue classificar corretamente *inputs* de acordo com o XOR. Isso porque o XOR não é linearmente separável e a saída da rede com função de ativação linear é uma reta.

6-

a) O número total de pesos então é dado por: $n_H + c + d * n_H + n_H * c = n_H(1 + d + c) + c$

- O bias é conectado com n_H + c pesos.
- Cada uma das d entradas são conectadas com cada uma das n_H unidades escondidas, por um total de $d * n_H$ pesos
- Cada unidade escondida n_H é conectado com cada unidade de saída c, por $n_H * c$ pesos.

b) Considere a equação para a saída de uma rede neural:

$$z_k = f\left[\sum_{j=1}^{n_H} w_{kj} f\left(\sum_{i=1}^{d} w_{ji}x_i + b_j\right) + b_k\right]$$

Caso o sinal for trocado para cada peso indo para uma unidade oculta, e os pesos saindo das unidades ocultas também forem trocados, então o resultado da rede não é alterado se a função de ativação possuir a seguinte propriedade: $f(-x) = f(x)$. Em outras palavras, se $w_{ji} \mapsto -w_{ji}$ e $b_j \mapsto -b_j$ na equação acima, então:

$$f\left(\sum_{i=1}^{d} -w_{ji}x_i - b_j\right) = f\left(-(\sum_{i=1}^{d} w_{ji}x_i + b_j)\right) = f\left(\sum_{i=1}^{d} w_{ji}x_i + b_j\right)$$

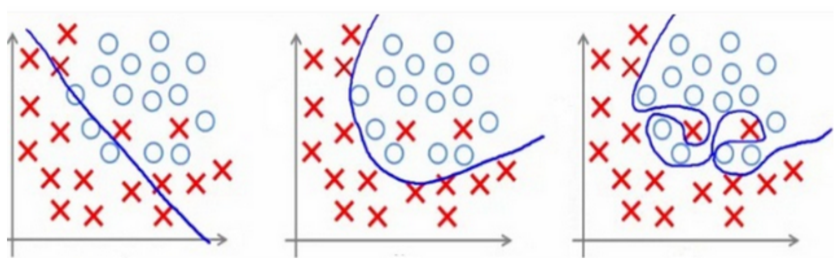
Analogamente, o mesmo ocorre na função de ativação para os nós da última camada. Note que esse resultado só é válido se aplicado para funções de ativação pares, onde $f(-x) = f(x)$

Exercícios Deep Learning
Aula 4

January 23, 2020

1 Ajuste de modelos

1- Observe as figuras abaixo e procure identificar as características de cada um dos modelos que foram gerados para classificar os dados em duas categorias. Para aqueles modelos cujo ajuste é ruim, cite um problema causado por ele e proponha sugestões para solucioná-lo.



2 Regularização

2- Considere a seguinte função objetiva de mínimos quadrados regularizada L2 para uma regressão linear: $f(w) = \|\mathbf{w}'\mathbf{x} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2$. Como a escolha de λ afeta a reta estimada?

3- Considere uma rede neural com duas features de entrada e uma camada escondida de dois nós. Sejam $W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}$ os pesos da rede:

$$W^{[1]} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} b^{[1]} = \begin{bmatrix} -2 \\ 4 \end{bmatrix} W^{[2]} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} b^{[2]} = \begin{bmatrix} -4 \end{bmatrix}$$

E os gradientes em relação a função de perda:

$$\frac{\partial J}{\partial W^{[1]}} = \begin{bmatrix} -10 & 5 \\ 1 & 2.5 \end{bmatrix} \frac{\partial J}{\partial b^{[1]}} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} \frac{\partial J}{\partial W^{[2]}} = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \frac{\partial J}{\partial b^{[2]}} = \begin{bmatrix} -2 \end{bmatrix}$$

a) Seja $\alpha = 0.1$, faça uma iteração do backpropagation nos pesos da rede. Além disso, considerando a regularização L2 com $\lambda = 0.5$, atualize os pesos considerando a regularização.

b) Qual diferença você nota entre os pesos com e sem regularização?

c) Ao final de várias iterações, após a rede convergir, o que você espera que seja a diferença entre os pesos das duas redes? Explique qual será o efeito provável da regularização no erro no conjunto de testes e porque isso ocorre.

3 Normalização

4- Explique qual é o efeito causado pela normalização das entradas, $\mathbf{x}$, no treinamento da rede.

4 Dropout

5- Suponha que você esteja treinando uma rede com dropout e a probabilidade de um nó ser mantido muda de 0.6 para 0.5.

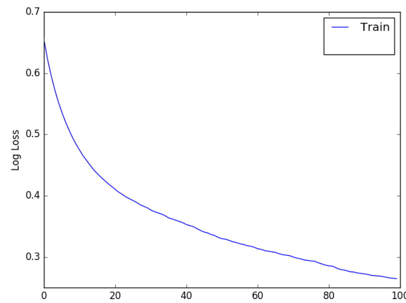
a) O que acontece com o efeito de regularização ao diminuir a probabilidade de um nó ser mantido na rede?

b) Qual o impacto dessa mudança no erro calculado com o conjunto de treino?

c) Durante o treinamento com dropout, a rede é modificada diversas vezes. Alguma modificação deve ser feita na rede em tempo de teste?

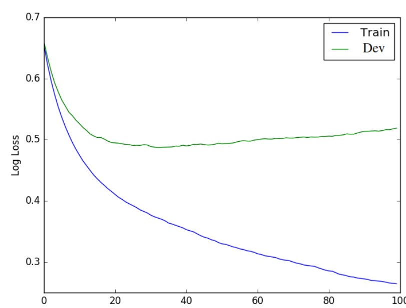
5 Treino, teste e validação

6- Imagine que você está treinando um modelo e ao plotar a curva de erro ao longo das iterações com os dados de treino você obtenha o seguinte gráfico:



a) Estime o número da iteração onde o modelo obteve o melhor resultado, ou seja, sua predição foi melhor.

Considere agora que, além de plotar a curva de erro ao longo das iterações com os dados de treino, a curva de erro para o conjunto de dados de validação ao longo das mesma iterações também é considerada, conforme mostrado no gráfico abaixo:



b) Estime o número da iteração onde o modelo obteve o melhor resultado, ou seja, onde de forma geral, sua predição foi melhor.

c) Explique a importância do conjunto de validação nesse caso e indique uma estratégia para obter o erro mínimo nesse conjunto.

6 Gradiente descendente, Mini-batch e SGD

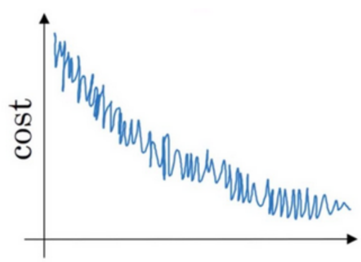
7- Sobre o gradiente descendente em mini-batch, discorra sobre as vantagens e desvantagens de diferentes tamanhos para o batch. Por que o melhor tamanho de mini-batch geralmente não é 1 e nem m, mas algo intermediário?

8- Para cada uma das afirmativas abaixo, diga porque ela é falsa.

a) Treinar uma época (uma passagem pelo conjunto de treinamento) usando gradiente descendente em mini-batch é mais rápido do que treinar uma época usando descida gradiente em batch.

b) Uma iteração de gradiente descendente em mini-batch (computação em um único mini-batch) é mais lenta que uma iteração de gradiente descendente em batch.

9- Suponha que o custo J do seu algoritmo de aprendizado, plotado como uma função do número de iterações, seja representado no gráfico abaixo. A partir da análise do gráfico, explique qual deve ter sido o algoritmo utilizado: gradiente descendente em batch ou gradiente descendente em mini-batch.



7 Exponentially Weighted Average

10- Suponha que a temperatura (em graus Celsius) em Casablanca nos primeiros três dias de janeiro seja a mesma:

1º de janeiro: $\theta_1 = 10$
2 de janeiro: $\theta_2 = 10$

Digamos que você use uma média exponencialmente ponderada com $\beta = 0.5$ para acompanhar a temperatura: $v_0 = 0$, $v_t = \beta v_{t-1} + (1-\beta)\theta_t$. Se v_2 é o valor calculado após o dia 2 sem correção de viés, e $v_t^{corrigido} = \frac{v_t}{1-\beta^t}$ é o valor que você calcula com correção de viés. Quais são esses valores?

Soluções

1-

Se o seu algoritmo tem bias alto:

- Tente uma RN maior (mais hidden layers, mais neurônios).
- Tente um modelo diferente que seja mais adequado para seus dados.
- Tente aumentar o número de iterações.
- Tente algoritmos de otimização diferentes.

Se o seu algoritmo tem variância alta:

- Obtenha mais dados.
- Tente regularização (por ex. L1, L2, Dropout).
- Parar de treinar quando o erro no conjunto de teste aumentar (early stopping)
- Tente um modelo diferente que seja mais adequado para seus dados.

2- A regularização é um método que ajuda a prevenir o overfitting, controlando a complexidade do modelo. O λ controla um trade-off entre encaixar bem o conjunto de treinamento e manter os pesos pequenos. Um λ grande pode levar a underfitting (um modelo mais linear e simples) enquanto que um λ pequeno pode levar a overfitting (um modelo mais complicado - maior intervalo de valores para os parâmetros).

3-

a)

Sem regularização:

$$W^{[1]} = \begin{bmatrix} 2 & 1.5 \\ 1.9 & 2.75 \end{bmatrix} b^{[1]} = \begin{bmatrix} -2.3 \\ 4.3 \end{bmatrix} W^{[2]} = \begin{bmatrix} 1.1 \\ 4.8 \end{bmatrix} b^{[2]} = \begin{bmatrix} -3.8 \end{bmatrix}$$

Com regularização:

$$W^{[1]} = \begin{bmatrix} 1.9 & 1.3 \\ 1.7 & 2.45 \end{bmatrix} b^{[1]} = \begin{bmatrix} -2.1 \\ 3.9 \end{bmatrix} W^{[2]} = \begin{bmatrix} 1 \\ 4.3 \end{bmatrix} b^{[2]} = \begin{bmatrix} -3.4 \end{bmatrix}$$

b) Os pesos na rede com regularização tem valor absoluto menor.

c) A rede final sem regularização terá alguns valores de pesos altos que funcionam para o conjunto de treinamento, mas podem ocasionar grandes erros nas previsões para o conjunto de teste. A rede com regularização terá pesos menores e, portanto, a complexidade do modelo é menor, o que reduz overfitting.

4- Normalizar o input reduz as regiões da função de perda que são relativamente planas ao deixar a função mais simétrica (em média). Isso torna o algoritmo gradiente descendente mais rápido, pois existem menos regiões onde o gradiente é pequeno e, portanto, os pesos demoram a mudar significativamente.

5-

a) O efeito de regularização aumenta já que menos nós tendem a permanecer na rede.

b) O erro do conjunto de treino aumentará.

c) Nenhuma modificação deve ser feita, a previsão durante o teste é feita com todos nós da rede.

6-

a) Na iteração 100.

b) Aproximadamente na iteração 40.

c) Analisar os efeitos do treinamento da rede com o conjunto de validação pode ser útil para detectar o overfitting do modelo nos dados de treino. O decaimento no erro dos dados de treino não significa decaimento no erro dos dados de validação e teste. Sendo assim, acompanhando o erro para o conjunto de validação, é possível parar de treinar o modelo assim que o erro no conjunto de validação subir (early-stopping).

7- Se o tamanho do mini-batch for 1, você perderá os benefícios da vetorização entre os exemplos no mini-batch. Se o tamanho do mini-batch é m, você acaba com o gradiente descente em batch, que tem que processar todo o conjunto de treinamento antes de fazer progresso.

8-

a) Falso, porque treinar uma época consiste em passar por todo o conjunto de treinamento (forward e backpropagation). Usando gradiente descendente em mini-batch você executa esse processo várias vezes, já que seu conjunto de treino é dividido em batchs e, no caso do gradiente em batch, o processo é executado uma só vez para todo o conjunto de dados, além disso, o gradiente descendente em batch se aproveita melhor da eficiência através da vetorização.

b) Uma iteração de gradiente descendente em mini-batch (computação em um único mini-batch) é mais rápida que uma iteração de gradiente descendente em batch porque a quantidade de dados de treino é menor.

9- Se você estiver usando gradiente descendente em mini-batch, isso parece aceitável. Mas se você estiver usando gradiente descendente em batch, algo está errado. Haverá algumas oscilações quando você estiver usando gradiente descendente em mini-batch, pois pode haver algum exemplo de dados ruidosos em batchs. No entanto, a descida gradiente em batch sempre garante um menor J antes de atingir o ótimo.

10- $v_2 = 7.5$ e $v_2^{corrigido} = 10$