矩阵求导——MLP神经网络

覃雄派

矩阵求导——MLP神经网络

提纲

- 矩阵求导——MLP神经网络

π@RUC

矩阵求导

X
10元素
$A^{[0]}$

$W^{[1]}$、$b^{[1]}$

隐藏层
64神经元
$Z^{[1]} \rightarrow A^{[1]}$

$W^{[2]}$、$b^{[2]}$

输出层
1个神经元
$Z^{[2]} \rightarrow A^{[2]}$

损失函数J
二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)= (64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**

有7个样本，每个样本为10×1的列向量，7个样本就是10*7

每列加64*1

每列加1*1

从输入层计算隐藏层

矩阵求导

X 10元素 $A^{[0]}$	$W^{[1]}$、$b^{[1]}$	隐藏层 64神经元 $Z^{[1]} \rightarrow A^{[1]}$	$W^{[2]}$、$b^{[2]}$	输出层 1个神经元 $Z^{[2]} \rightarrow A^{[2]}$	损失函数J 二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)= (64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**

有7个样本，每个样本为10×1的列向量，7个样本就是10*7

每列加64*1

σ针对每个元素操作

每列加1*1

隐藏层的非线性传导

矩阵求导

X 10元素 $A^{[0]}$	$W^{[1]}$、$b^{[1]}$	隐藏层 64神经元 $Z^{[1]}\rightarrow A^{[1]}$	$W^{[2]}$、$b^{[2]}$	输出层 1个神经元 $Z^{[2]}\rightarrow A^{[2]}$	损失函数J 二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)**= **(64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**

有7个样本，每个样本为10×1的列向量，7个样本就是10*7

每列加64*1

每列加1*1

从隐藏层计算输出层

矩阵求导

X
10元素
$A^{[0]}$

$W^{[1]}$、$b^{[1]}$

隐藏层
64神经元
$Z^{[1]} \rightarrow A^{[1]}$

$W^{[2]}$、$b^{[2]}$

输出层
1个神经元
$Z^{[2]} \rightarrow \boxed{A^{[2]}}$

损失函数J
二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)= (64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - $\boxed{\text{(4) } \hat{y} = A^{[2]} = \sigma(Z^{[2]}) = (1* 7)}$

有7个样本，每个样本为10×1的列向量，7个样本就是10*7

每列加64*1

每列加1*1

输出层的非线性传导

矩阵求导

X
10元素
$A^{[0]}$

$W^{[1]}$、$b^{[1]}$

隐藏层
64神经元
$Z^{[1]}\rightarrow A^{[1]}$

$W^{[2]}$、$b^{[2]}$

输出层
1个神经元
$Z^{[2]}\rightarrow A^{[2]}$

损失函数J
二元交叉熵

- 神经网络
  - X即$A^{[0]}$
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - 最后用$A^{[2]}$构造损失函数，，注意$A^{[2]}$即预测值$\hat{y}$
    - 二值分类器(0/1)的交叉熵损失函数的形式为
    - $J = -\frac{1}{n}((Ylog\left(A^{[2]}\right) + (1 - Y)\log\left(1 - A^{[2]}\right))$
    - Y=(1*7)
    - $\hat{y} = A^{[2]}$=(1*7)

有7个样本，每个样本为10×1的列向量，7个样本就是10*7

$dA^{[2]} = -\frac{Y}{A^{[2]}} + \frac{1-Y}{1-A^{[2]}}$=(1*7)

根据损失函数计算损失值

π@RUC

矩阵求导

X
10元素
$A^{[0]}$

$W^{[1]}$、$b^{[1]}$

隐藏层
64神经元
$Z^{[1]} \rightarrow A^{[1]}$

$\boxed{W^{[2]}}$、$b^{[2]}$

输出层
1个神经元
$Z^{[2]} \rightarrow A^{[2]}$

损失函数J
二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)= (64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**
  - 最后用$A^{[2]}$构造损失函数，，注意$A^{[2]}$即预测值$\hat{y}$
    - 二值分类器(0/1)的交叉熵损失函数的形式为
    - $J = -\frac{1}{n}((Ylog(A^{[2]}) + (1-Y)\log(1-A^{[2]}))$
  - 计算损失函数对$W^{[1]}$、$b^{[1]}$、$W^{[2]}$、$b^{[2]}$的导数

$$dA^{[2]} = -\frac{Y}{A^{[2]}} + \frac{1-Y}{1-A^{[2]}} = (1*7)$$

矩阵的各个位置点乘

$$dZ^{[2]} = dA^{[2]} g'(Z^{[2]}) \ \mathbf{(1*7)\ (1*7)}$$

$$dW^{[2]} = \frac{dJ}{dW^{[2]}} = \frac{\boldsymbol{dJ}}{\mathbf{dA^{[2]}}}\frac{\mathbf{dA^{[2]}}}{\mathbf{dZ^{[2]}}}\frac{dZ^{[2]}}{dW^{[2]}} = \frac{\boldsymbol{dJ}}{\mathbf{dZ^{[2]}}}\frac{dZ^{[2]}}{dW^{[2]}}$$

$$= dZ^{[2]}(A^{[1]})^T$$

(1*7)(7*64)

矩阵乘法

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T$$

$$\frac{dZ^{[2]}}{dW^{[2]}} = \frac{d(W^{[2]}A^{[1]}+b^{[2]})}{dW^{[2]}} = \frac{d(EW^{[2]}A^{[1]}+b^{[2]})}{dW^{[2]}} = E^T(A^{[1]})^T$$

E为单位矩阵

矩阵求导

X 10元素 $A^{[0]}$ | $W^{[1]}$、$b^{[1]}$ | 隐藏层 64神经元 $Z^{[1]}\to A^{[1]}$ | $W^{[2]}$、$b^{[2]}$ | 输出层 1个神经元 $Z^{[2]}\to A^{[2]}$ | 损失函数J 二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \to Z^{[1]}\to A^{[1]}\to Z^{[2]}\to A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)**= **(64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**
  - (5)最后用$A^{[2]}$构造损失函数，，注意$A^{[2]}$即预测值$\hat{y}$
    - 二值分类器(0/1)的交叉熵损失函数的形式为
    - $J = -\frac{1}{n}((Y\log(A^{[2]}) + (1-Y)\log(1-A^{[2]}))$
  - 计算损失函数对$W^{[1]}$、$b^{[1]}$、$W^{[2]}$、$b^{[2]}$的导数

$$\frac{dJ}{db^{[2]}} = \frac{dJ}{dA^{[2]}}\frac{dA^{[2]}}{dZ^{[2]}}\frac{dZ^{[2]}}{db^{[2]}}$$
$$= [A^{[2]} - Y][1]= [A^{[2]} - Y]$$
$$= dZ^{[2]}\ [1]$$

(1*7)(7*1) 相当于每行的各列累加

为下一步准备

$$\frac{dJ}{dA^{[1]}} = \frac{dJ}{dA^{[2]}}\frac{dA^{[2]}}{dZ^{[2]}}\frac{dZ^{[2]}}{dA^{[1]}} = \frac{dJ}{dZ^{[2]}}\frac{dZ^{[2]}}{dA^{[1]}}$$
$$= dZ^{[2]}(W^{[2]})^T$$

$(W^{[2]})^T dZ^{[2]}$才能完成矩阵相乘

(64*1)(1*7)

$$\frac{dZ^{[2]}}{dA^{[1]}} = \frac{d(W^{[2]}A^{[1]}+b^{[2]})}{dA^{[1]}}=(W^{[2]})^T$$

矩阵求导

X
10元素
$A^{[0]}$

$\boxed{W^{[1]}}$、 $b^{[1]}$

隐藏层
64神经元
$Z^{[1]} \rightarrow A^{[1]}$

$W^{[2]}$、$b^{[2]}$

输出层
1个神经元
$Z^{[2]} \rightarrow A^{[2]}$

损失函数J
二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)**= **(64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**
  - (5)最后用$A^{[2]}$构造损失函数，，注意$A^{[2]}$即预测值$\hat{y}$
    - 二值分类器(0/1)的交叉熵损失函数的形式为
    - $J = -\frac{1}{n}((Y log(A^{[2]}) + (1-Y)\log(1-A^{[2]}))$
  - 计算损失函数对$W^{[1]}$、$b^{[1]}$、$W^{[2]}$、$b^{[2]}$的导数

$dZ^{[1]} = dA^{[1]} g'(Z^{[1]})$
(64*7) (64*7)

矩阵的各个位置点乘

$$dW^{[1]} = \frac{dJ}{dW^{[1]}} = \frac{\boldsymbol{dJ}}{\mathbf{dA^{[2]}}} \frac{\boldsymbol{dA^{[2]}}}{\mathbf{dZ^{[2]}}} \frac{dZ^{[2]}}{dA^{[1]}} \frac{dA^{[1]}}{dZ^{[1]}} \frac{dZ^{[1]}}{dW^{[1]}} = \frac{\boldsymbol{dJ}}{\mathbf{dZ^{[1]}}} \frac{dZ^{[1]}}{dW^{[1]}}$$

$= dZ^{[1]}(A^{[0]})^T$
(64*7)(7*10)

矩阵乘法

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$
$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$
$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T$$

$$\frac{dZ^{[1]}}{dW^{[1]}} = \frac{d(W^{[1]}A^{[0]}+b^{[1]})}{dW^{[1]}} = \frac{d(EW^{[1]}A^{[0]}+b^{[1]})}{dW^{[1]}} = E^T(A^{[0]})^T$$

E为单位矩阵

矩阵求导

X
10元素
$A^{[0]}$

$W^{[1]}$、$\boxed{b^{[1]}}$

隐藏层
64神经元
$Z^{[1]} \rightarrow A^{[1]}$

$W^{[2]}$、$b^{[2]}$

输出层
1个神经元
$Z^{[2]} \rightarrow A^{[2]}$

损失函数J
二元交叉熵

- 神经网络
  - 处理过程 $A^{[0]} \rightarrow Z^{[1]} \rightarrow A^{[1]} \rightarrow Z^{[2]} \rightarrow A^{[2]}$
  - X即$A^{[0]}$
  - (1) $Z^{[1]} = W^{[1]}X + b^{[1]}$=**(64*10)**(10* 7)+**(64*1)**= **(64* 7)**
  - (2) $A^{[1]} = \sigma(Z^{[1]})$=(64* 7)
  - (3) $Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}$=**(1*64)**(64* 7) +**(1*1)**=(1* 7)
  - (4) $\hat{y} = A^{[2]} = \sigma(Z^{[2]})$ =**(1* 7)**
  - (5)最后用$A^{[2]}$构造损失函数，，注意$A^{[2]}$即预测值$\hat{y}$
    - 二值分类器(0/1)的交叉熵损失函数的形式为
    - $J = -\frac{1}{n}((Ylog(A^{[2]}) + (1-Y)\log(1-A^{[2]}))$
  - 计算损失函数对$W^{[1]}$、$b^{[1]}$、$W^{[2]}$、$b^{[2]}$的导数

$$db^{[1]} = \frac{dJ}{db^{[1]}} = \frac{dJ}{dA^{[2]}}\frac{dA^{[2]}}{dZ^{[2]}}\frac{dZ^{[2]}}{dA^{[1]}}\frac{dA^{[1]}}{dZ^{[1]}}\frac{dZ^{[1]}}{db^{[1]}} =$$

$$\frac{dJ}{dZ^{[1]}}\frac{dZ^{[1]}}{db^{[1]}} =$$

$dZ^{[1]}[1]$ =

$dZ^{[1]}[1]$

(64*7)(7*1)　　相当于每行的各列累加

矩阵求导