**1.传导过程，第一种记法**

备注Ndim\*V\*V\*1🡪Ndim \*1

备注V\*Ndim\*Ndim\*1🡪V\*1

备注🡪V\*1

**2.一个例子**

假设词汇表|V| =10000, 所有词向量的维度为300

现在有999个样本，每个词汇都是one hot encoding(10000,1)的向量

**3.传导和反向传播**

Layer1的前向传导过程，红色表示1个样本

X = (1, 10000)

=(10000, 1)

==(300, 10000)

=

=(300,10000)\*(10000, 1)

= (300, 1)

=(300, 1)，备注直接等于，没有做非线性变换

Layer2的前向传导过程

== (10000, 300)

=(10000, 300)\*(300, 1)=(10000, 1)

=(10000, 1) ，备注直接等于，没有做非线性变换

最后

=(10000, 1)

L为多分类的交叉熵损失函数，

Y=(10000,1), =(10000，1)

Layer2的反向传播过程

= ，各个元素相乘

=(10000,1)×( 1, 300) = (10000,300)，矩阵相乘

==(300,10000)(10000,1) = (300, 1)

Layer1的反向传播过程

，各个元素相乘

= (300, 1) ×( 1, 10000) = (300\*10000) ，矩阵相乘

**4.带数据的例子**

现有如下的句子

|  |
| --- |
| The man who passes the sentence should swing the sword. (Ned Stark) |

在这句话里，有10个词项，8个唯一的词项，于是字典表大小|V|为8，每个词项编码为one hot encoding的时候，是一个8维向量。

如果我们设定c=1，那么我们针对passes这个词项，就可以把它之前的词项who和它之后的词项the提取出来，和passes一起构成2个训练样本。

|  |
| --- |
| Passes who  Passes the |

我们准备利用Skip-Gram进行训练，把8维的one hot encoding通过训练，降维为3维。

W(8\*3)初始值如下。

|  |  |  |  |
| --- | --- | --- | --- |
| Man | -0.078 | 0.018 | 0.033 |
| Passes | 0.068 | 0.170 | -0.109 |
| Sentence | -0.158 | -0.081 | -0.151 |
| Should | 0.150 | 0.064 | 0.145 |
| swing | -0.097 | -0.055 | 0.188 |
| sword | 0.036 | 0.071 | 0.059 |
| The | 0.168 | -0.060 | -0.058 |
| who | 0.098 | 0.015 | 0.096 |

即W[1]的初始值为（W的转置）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| -0.078 | 0.068 | -0.158 | 0.150 | -0.097 | 0.036 | 0.168 | 0.098 |
| 0.018 | 0.170 | -0.081 | 0.064 | -0.055 | 0.071 | -0.060 | 0.015 |
| 0.033 | -0.109 | -0.151 | 0.145 | 0.188 | 0.059 | -0.058 | 0.096 |

词项passes对应的one hot encoding如下，它作为输入X，即A[0]。

|  |  |
| --- | --- |
| Man | 0 |
| Passes | 1 |
| Sentence | 0 |
| Should | 0 |
| swing | … |
| sword | … |
| The | 0 |
| who | 0 |

h=WTX得到的中间结果如下，即A[1]= Z[1]= W[1] A[0]。

|  |
| --- |
| 0.068 |
| 0.170 |
| -0.109 |

W'(3\*8)初始值如下。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0.192 | 0.070 | -0.066 | 0.014 | -0.012 | 0.013 | 0.016 | -0.028 |
| 0.176 | 0.061 | 0.117 | 0.006 | 0.067 | 0.111 | 0.175 | -0.016 |
| 0.012 | -0.046 | 0.083 | -0.044 | 0.147 | -0.097 | -0.198 | 0.148 |

即W[2]的初始值为（W'的转置）。

|  |  |  |
| --- | --- | --- |
| 0.192 | 0.176 | 0.012 |
| 0.070 | 0.061 | -0.046 |
| -0.066 | 0.117 | 0.083 |
| 0.014 | 0.006 | -0.044 |
| -0.012 | 0.067 | 0.147 |
| 0.013 | 0.111 | -0.097 |
| 0.016 | 0.175 | -0.198 |
| -0.028 | -0.016 | 0.148 |

W'Th的结果如下，即==W[2]A[1]，再经过Softmax处理，就是。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Man | 0.042 |  | Man | 0.128 |
| Passes | 0.020 |  | Passes | 0.125 |
| Sentence | 0.006 |  | Sentence | 0.124 |
| Should | 0.007 | 🡪Softmax🡪 | Should | 0.124 |
| swing | -0.005 |  | swing | 0.122 |
| sword | 0.030 |  | sword | 0.127 |
| The | 0.052 |  | The | 0.130 |
| who | -0.021 |  | who | 0.120 |

Softmax处理的目的是给出每个分类的概率，各个类别的概率之和为1，如图所示。

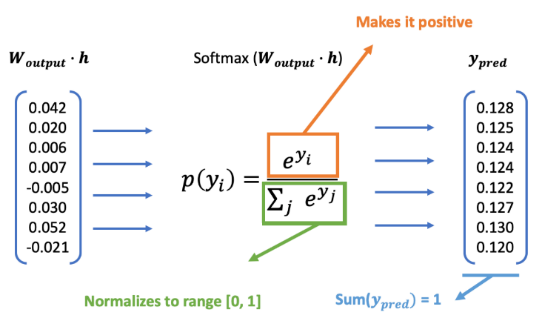


图1. Softmax原理

由于Skip-Gram针对passes的输出是the和who，它们的one hot encoding具体如下。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Man | 0 |  | Man | 0 |
| Passes | 0 |  | Passes | 0 |
| Sentence | 0 |  | Sentence | 0 |
| Should | 0 |  | Should | 0 |
| swing | 0 |  | swing | 0 |
| sword | 0 |  | sword | 0 |
| The | 1 |  | The | 0 |
| who | 0 |  | who | 1 |

现在开始进行反向传播。根据，有样本(Passes who)和样本(Passes the)的误差分别如下，即有两个误差。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Man | 0.128 |  | Man | 0.128 |
| Passes | 0.125 |  | Passes | 0.125 |
| Sentence | 0.124 |  | Sentence | 0.124 |
| Should | 0.124 |  | Should | 0.124 |
| swing | 0.122 |  | swing | 0.122 |
| sword | 0.127 |  | sword | 0.127 |
| The | 0.130 - 1 |  | The | 0.130 |
| who | 0.120 |  | who | 0.120 - 1 |

这两个误差相加，于是两个样本的总的误差如下，根据这个误差进行反向传播，注意=。

|  |  |
| --- | --- |
| Man | 0.256 |
| Passes | 0.251 |
| Sentence | 0.247 |
| Should | 0.248 |
| swing | 0.245 |
| sword | 0.253 |
| The | -0.741 |
| who | -0.759 |

反向传播修正W’即W[2]

=(8,1) ×( 1, 3) = (8,3)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.256 |  |  | |  |  | |  | 0.017 | 0.044 | -0.028 |
| 0.251 |  |  | |  |  | |  | 0.017 | 0.043 | -0.027 |
| 0.247 |  |  | |  |  | |  | 0.017 | 0.042 | -0.027 |
| 0.248 | × | 0.068 | 0.17 | | | -0.109 | = | 0.017 | 0.042 | -0.027 |
| 0.245 |  |  | |  |  | |  | 0.017 | 0.042 | -0.027 |
| 0.253 |  |  | |  |  | |  | 0.017 | 0.043 | -0.028 |
| -0.741 |  |  | |  |  | |  | -0.050 | -0.126 | 0.081 |
| -0.759 |  |  | |  |  | |  | -0.052 | -0.129 | 0.083 |

反向传播修正W即W[1]

dA[1] = (W[2])TdZ[2]有 ==(3,8) ×(8,1) = (3, 1)

= (3, 1)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  | 0.256 |  |  |
|  |  |  |  |  |  |  |  |  | 0.251 |  |  |
| 0.192 | 0.070 | -0.066 | 0.014 | -0.012 | 0.013 | 0.016 | -0.028 |  | 0.247 |  | 0.064 |
| 0.176 | 0.061 | 0.117 | 0.006 | 0.067 | 0.111 | 0.175 | -0.016 | × | 0.248 | = | 0.018 |
| 0.012 | -0.046 | 0.083 | -0.044 | 0.147 | -0.097 | -0.198 | 0.148 |  | 0.245 |  | 0.047 |
|  |  |  |  |  |  |  |  |  | 0.253 |  |  |
|  |  |  |  |  |  |  |  |  | -0.741 |  |  |
|  |  |  |  |  |  |  |  |  | -0.759 |  |  |

= (3, 1) ×( 1, 8) = (3\*8)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.064 |  |  |  |  |  |  |  |  |  |  | 0 | 0.064 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.018 | × | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | = | 0 | 0.018 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.047 |  |  |  |  |  |  |  |  |  |  | 0 | 0.047 | 0 | 0 | 0 | 0 | 0 | 0 |

，注意W[1]是W转置过的，即它是W的转置。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -0.078 | 0.065 | 0.066 | -0.158 | 0.150 | -0.097 | 0.036 | 0.168 | 0.098 |
| 0.018 | 0.169 | 0.168 | -0.081 | 0.064 | -0.055 | 0.071 | -0.060 | 0.015 |
| 0.033 | -0.111 | -0.112 | -0.151 | 0.145 | 0.188 | 0.059 | -0.058 | 0.096 |

=

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -0.078 | 0.068 | -0.158 | 0.150 | -0.097 | 0.036 | 0.168 | 0.098 |  | 0 | 0.064 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.018 | 0.170 | -0.081 | 0.064 | -0.055 | 0.071 | -0.060 | 0.015 | -0.05\* | 0 | 0.018 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.033 | -0.109 | -0.151 | 0.145 | 0.188 | 0.059 | -0.058 | 0.096 |  | 0 | 0.047 | 0 | 0 | 0 | 0 | 0 | 0 |

得到新的W[1]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **-0.078** | **0.065** | **0.066** | **-0.158** | **0.150** | **-0.097** | **0.036** | **0.168** | **0.098** |
| **0.018** | **0.169** | **0.168** | **-0.081** | **0.064** | **-0.055** | **0.071** | **-0.060** | **0.015** |
| **0.033** | **-0.111** | **-0.112** | **-0.151** | **0.145** | **0.188** | **0.059** | **-0.058** | **0.096** |

，注意W[2]是W'转置过的，即它是W'的转置。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.191 | 0.174 | 0.013 |  | 0.192 | 0.176 | 0.012 |  | 0.017 | 0.044 | -0.028 |
| 0.069 | 0.059 | -0.045 |  | 0.070 | 0.061 | -0.046 |  | 0.017 | 0.043 | -0.027 |
| -0.067 | 0.115 | 0.084 |  | -0.066 | 0.117 | 0.083 |  | 0.017 | 0.042 | -0.027 |
| 0.013 | 0.004 | -0.043 | = | 0.014 | 0.006 | -0.044 | -0.05× | 0.017 | 0.042 | -0.027 |
| -0.013 | 0.065 | 0.148 |  | -0.012 | 0.067 | 0.147 |  | 0.017 | 0.042 | -0.027 |
| 0.012 | 0.109 | -0.096 |  | 0.013 | 0.111 | -0.097 |  | 0.017 | 0.043 | -0.028 |
| 0.019 | 0.181 | -0.202 |  | 0.016 | 0.175 | -0.198 |  | -0.050 | -0.126 | 0.081 |
| -0.025 | 0.010 | 0.144 |  | -0.028 | -0.016 | 0.148 |  | -0.052 | -0.129 | 0.083 |

得到新的W[2]

|  |  |  |
| --- | --- | --- |
| **0.191** | **0.174** | **0.013** |
| **0.069** | **0.059** | **-0.045** |
| **-0.067** | **0.115** | **0.084** |
| **0.013** | **0.004** | **-0.043** |
| **-0.013** | **0.065** | **0.148** |
| **0.012** | **0.109** | **-0.096** |
| **0.019** | **0.181** | **-0.202** |
| **-0.025** | **0.010** | **0.144** |

重新正向传导一下，看看误差有没有缩小；请打开“2021-11-20SkipGram-正向传播、反向传导、在正向传播.xlsm”文件进行实验。