任务一：基于机器学习的文本分类

刘向阳

2020 年 8 月 22 日

1 题目

“Rotten Tomatoes” 数据集包含了多条关于电影的评论，这些评论表达 5 种人类情感，分别为消极的（0）、有点消极的（1）、中立的（2）、有点积极的（3）、积极的（4）五种情感。现给定这样一个包含句子和其对应情感标签的数据集，请设计一个机器学习方法来从该数据集中获取一定的知识并对没有标签的测试集进行情感标签预测。

2 模型

针对该数据集，我设计了一个基于”Softmax Regression” 的文本分类方法，该方法的数学表示形式如下：

$$P(y^{(i)}=j|x^{(i)};\theta)=Softmax(\theta_j^T x^{(i)}+b)=\frac{e^{(\theta_j^T x^{(i)}+b)}}{\sum_{c=1}^{C} e^{(\theta_c^T x^{(i)}+b)}},$$

该式子表示模型将当前样例的标签预测为 j 的概率。其中，$x \in R^{m\times d}$ 为 m 个输入样例的特征向量组成的矩阵，d 为特征向量的维度，θ 和 b 为模型参数，C 为分类的类别数。

为了能够学习模型中的参数，我们使用了交叉熵损失函数来计算预测值和真实值之间的差距，其数学形式如下：

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{C}1\left\{y^{(i)}=j\right\}\log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{c=1}^{C} e^{\theta_l^T x^{(i)}}}\right],$$

其中，$1\{\cdot\}$ 为示性函数，即 $1\{TRUE\}=1, 1\{FALSE\}=0$. 我们在训练的过程中通过对损失函数求导得到其梯度，然后利用梯度下降法来学习模型中的参数。

3 实验及结果

在实验中，我们首先加载数据并进行预处理，主要包括去除非字母的符号，词形还原，去除停用词。在预处理过后，数据集总条目为 154626，我们利用”Bag of words”、”N-gram”、”TF-IDF” 等方法将文本语句转换为对应的特征向量。然后划分训练集和验证集，划分比例为 4:1，其中训练集条目数为 123700，验证集条目数为 30926。下面我们根据实验来分析提取特征的方法、学习率、Batch size 对结果的影响。

3.1 特征提取

参数设置

学习率：2，Batch size：64，训练周期：30。

实验结果

Methods	Dimension	Time/s	Accuary
BOW	13606	6334.374	0.6444
Bi-Gram	64015	90949.675	0.6345
TF-DF	13606	5375.971	0.6361

结果分析

从对比实验结果来看，词袋模型（BOW）的最终准确率最高，并且用时也处于相对优势，而 Bi-Gram 方法使特征维度增大了将近 5 倍，训练空间和时间都消耗极大，但是准确率却是最低的，可能的原因是该方法提取得到的特征向量变得极为稀疏。

对于 TF-IDF 而言，根据其原理可知一个词语在一篇文档中出现次数越多, 同时在所有文档中出现次数越少, 越能够代表该文档。因此，这种方法可能更适用于篇幅较长的文档分类。而在本数据集中，所有数据样例的长度都很短，所以一个词语可能很重要但是在样例中出现的次数也不是很多。

3.2 学习率

参数设置

特征提取: BOW, Batch size: 64, 训练周期: 30

实验结果

Learning rate	Best epoch	Loss	Accuary
0.01	30	1.1637	0.5367
0.1	29	1.0143	0.5985
1	28	0.9339	0.6425
2	28	0.9426	0.6444
4	22	0.9623	0.6404

结果分析

从对比结果中可以看出，当学习率为 2 的时候，最终的准确率最高。当准确率偏小的时候，模型收敛较慢，比如学习率为 0.01 的时候，在跑完 30 个周期后，验证集的损失还比较大，而学习率比较大的时候，梯度下降法每次的迭代步长相对较大，模型可能在较小的训练周期就达到比较低的损失和准确率，但是也可能会使训练震荡，导致模型不能找到最优解。因此，在训练周期比较大的时候，使用较小的学习率会更有助于模型找到最优解。

3.3 Batch size

参数设置

特征提取方法: BOW，学习率：2，训练周期：30

实验结果

Batch size	Best epoch	Time/s	Accuary
1	23	47819.690	0.5897
64	28	6334.374	0.6444
128	28	6261.490	0.6424
123700	30	5686.931	0.5082

结果分析

从对比结果可以看出，随着 Batch size 的增大，训练所需时间逐渐变少。当 Batch size 为最小值 1 时，训练 30 个周期迭代的次数较多，并且由于每次迭代仅用一个样本，所以其训练过程会因为噪声的存在而变得极其震荡，因此在相同的周期下，能够达到的效果也会变差。当 Batch size 为最大值 123700 时，由于每次迭代只进行了一次矩阵运算，所以时间有所减短，但是收敛效果变差，可能收敛到了局部最优值。

4 结论

综合以上实验结果对比，在训练一个模型时可能存在很多超参数影响着模型学习的效率和质量，本文只研究了特征提取方法、学习率和 Batch size 这三个，当然，还有其他的因素，比如优化算法，使用了 Momentum 加速的优化算法可能和普通的梯度下降算法得到的结果也会有所不同。因此我们在实现并训练一个模型时，需要有耐心并仔细地调试好每一个超参数，使得到的学习效果尽可能最优。