**自然语言处理实践作业一报告**

1120210526 马煜

1. **实验数据，平台和工具**

实验数据来自老师提供的酒店评价。使用Python代码。分词部分直接使用jieba分词工具。停词表选用了互联网查找到的最简单的中文停词表。在逻辑回归方法中用到的

1. **概率统计方法实现汉语句子的情感极性判断--基于Naïve Bayes的情感分析系统**

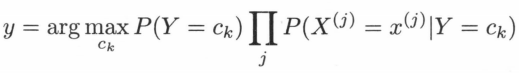
主要参考链接：https://blog.csdn.net/weixin\_43093481/article/details/115031477

1. **模型原理解释**
2. **贝叶斯定理**

P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}

1. **朴素贝叶斯Naïve Bayes**

基于贝叶斯定理和条件独立假设的分类方法；(条件独立假设定义：用于分类的特征在类确定的条件下都是条件独立的。)

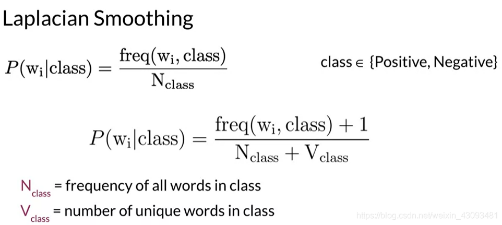


多特征的贝叶斯公式P(A | F1, F2, ..., Fn) = (P(A) \* P(F1 | A) \* P(F2 | A) \* ... \* P(Fn | A)) / P(F1) \* P(F2) \* ... \* P(Fn)

通过极大似然估计对先验概率P(neg)和条件概率p(word|Ci=neg)进行估计

1. **拉普拉斯平滑**

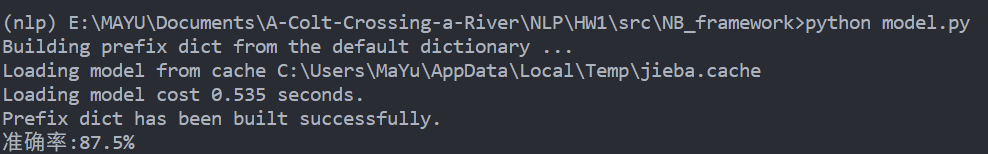
由于朴素贝叶斯法中通过累乘来计算最终概率，因此只要其中一个值为0，最终结果就会变为0，失去意义。因此加入平滑因子。



1. **对数概率（**log probabilities**）**

将连乘改为连续加法，避免概率过小连乘导致数值下溢。

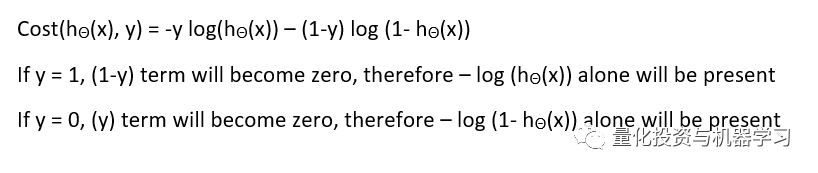
1. **实验过程**
2. 先对数据进行预处理，分出训练集和测试集（preprocess.py）
3. 读取训练和监测数据，按照积极和消极分类。读取的时候消除结尾换行，使用jieba分词，并用停词表去掉停词。计算得到正负语料的词频表备用。
4. 测试：分类按照积极和消极统计正确率，按照朴素贝叶斯方法计算每个词的后验概率，句子中的每个词的概率取log然后相加。得到该句为积极的可能性和消极的可能性，比较得出结论，与结果比较
5. **实验结果展示**

****

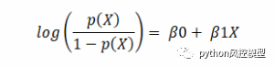
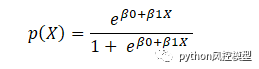
可以看到，整体正确率高达87.5%。由此说明，使用朴素贝叶斯方法做情感分析的效果还是不错的。朴素贝叶斯作为一种分类器效果还是相当不错的。

我的代码最主要的问题是没有处理语句顺序（不同的语序可能表达完全不同的含义）。因此这么高的正确率应该只是训练集合测试集过于相像的结果。

1. **神经网络方法实现汉语句子的情感极性判断--基于逻辑回归的情感分析系统**
2. **模型原理解释**
3. **损失函数**



1. **Sigmoid函数**

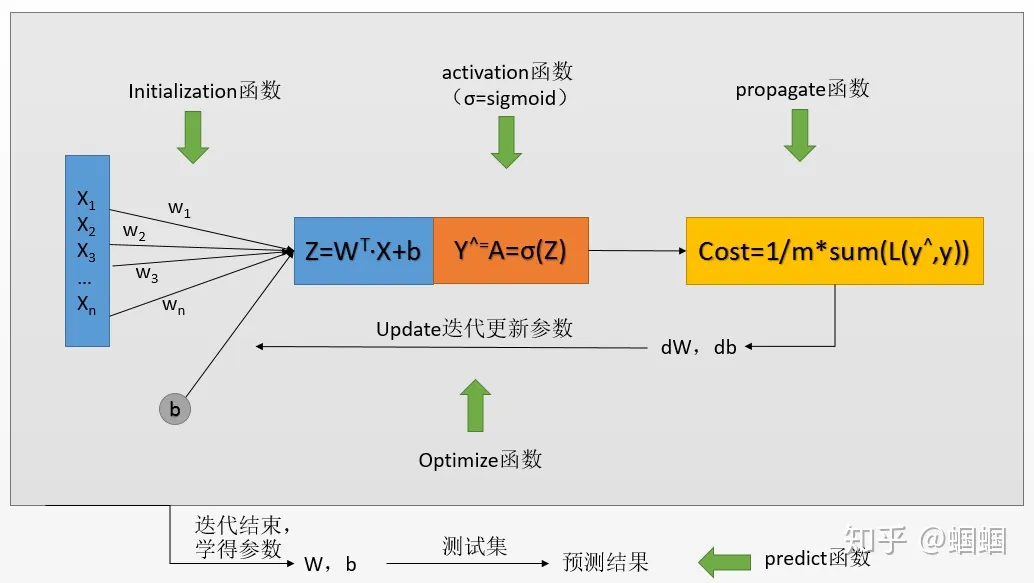
逻辑回归公式（反函数）->sigmoid函数

1. **梯度下降算法**

d\_weights = np.dot(X.T, (out - y)) / X.shape[0]

d\_bias = np.mean(out - y)

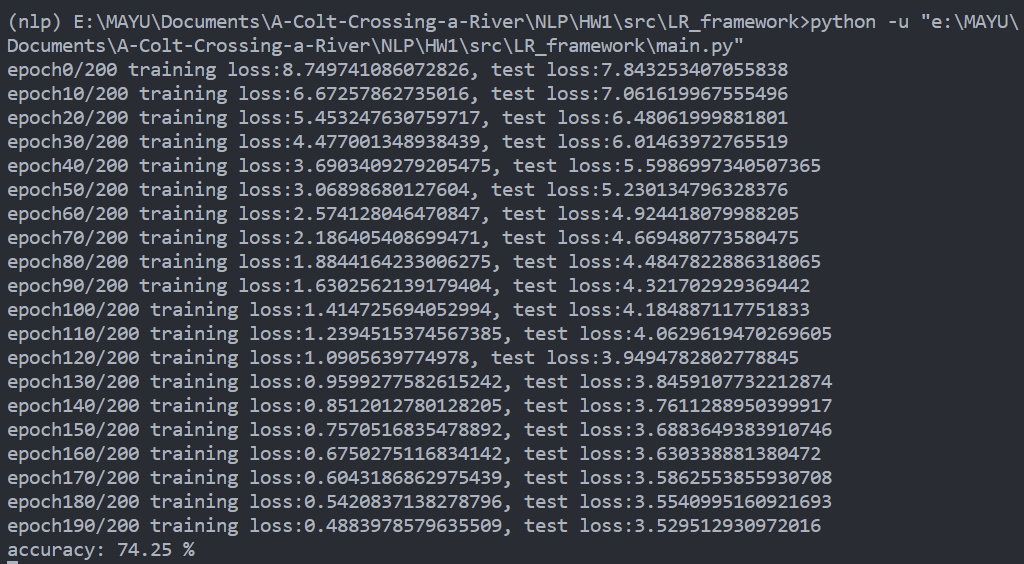
1. **实验过程**



首先预处理训练集和数据集。借助网图说明网络结构，模型分为激活函数，损失函数，正向传播和梯度下降反响传播部分。每个epoch通过正向传播计算loss，同时使用梯度下降更新权重矩阵，作为训练过程。测试时用训练好的权重矩阵进行正向传播预测，得到测试loss和准确率。

1. **实验结果展示**

**学习率0.01 batch= 32 epoch=200 每10个epoch打印一次损失**

****

可以看到准确率也可以达到74%，由于训练样本较小，所以效果不如贝叶斯，但也能做到情感识别功能。

**备注：因词向量过大，因此未放在压缩包中**