2023-2024学年第2学期
《自然语言处理导论》
实验2：垃圾邮件分类

专业：计算机科学与技术
学号：210340170
姓名：薛皓天
教师：卢敏

一、实验目的

1、使用四种线性分类算法和神经网络进行文本的分类。
2、理解五种算法的基本思路。

二、实验内容

1、朴素贝叶斯

如公式（1）所示，通过计算当前标签下的出现w这个词的概率和出现这个标签的概率，通过取对数求和得到的每一个数据预测的结果，在这之中找到概率最大的对应的标签，作为这一条数据的分类结果。如图Figure1 所示给出模型训练的准确度和损失。

$$\begin{aligned} \mathrm{Y}^{*} = \arg\max_{\mathrm{y}} P(\mathrm{y} \mid X) &= \arg\max_{\mathrm{y}} \frac{P(x \mid \mathrm{y}) \cdot P(\mathrm{y})}{P(x)} \\ &= \arg\max_{\mathrm{y}} \sum_{i=1}^{n} \log P\left(w_i \mid \mathrm{y}\right) + \log P(\mathrm{y}) \end{aligned} \tag{1}$$

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4827
1	0.99	0.99	0.99	747
accuracy			1.00	5574
macro avg	0.99	0.99	0.99	5574
weighted avg	1.00	1.00	1.00	5574

Naive Bayes accuracy is: 99.75%

Figure 1: 朴素贝叶斯模型准确率

2、支持向量机

支持向量机（SVM）模型是经典的二分类模型，目的是寻找一个超平面来对样本进行分割，如公式（2）所示分割的原则是间隔最大化，即找到使距离最大的w和b值，距离超平面最近的点被称为支持向量，如图Figure2所示。如图Figure3，为支持向量机模型训练结果。

$$\begin{aligned} &\max_{w,b} \frac{2}{\|w\|} \\ &s.t. y_i \left(w^T \cdot x_i + b\right) \geq 1, i = 1, 2, \ldots, m. \end{aligned} \tag{2}$$

$$Y = \operatorname{sgn}[\mathbf{w}^\top \mathbf{x} + b > 0] = \begin{cases} +1 & \mathbf{w}^\top \mathbf{x} + b > 0 \\ -1 & \mathbf{w}^\top \mathbf{x} + b \leq 0 \end{cases}$$

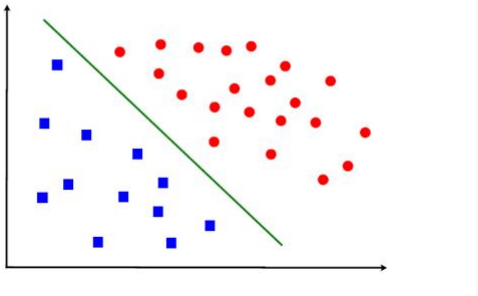

Figure 2: 支持向量机模型

	precision	recall	f1-score	support
0	1.00	0.97	0.99	747
1	1.00	1.00	1.00	4827
accuracy			1.00	5574
macro avg	1.00	0.99	0.99	5574
weighted avg	1.00	1.00	1.00	5574

SVM classifier accuracy is: 99.66%

Figure 3: 支持向量机模型准确率

3、感知机模型

感知机模型是二分类的线性分类模型。输入为实例的特征向量，输出为实例的类别（取+1和-1）。旨在求出将输入空间中的实例划分为两类的分离超平面。为求得超平面，感知机导入了基于误分类的损失函数，利用梯度下降法对损失函数进行最优化求解。在迭代更新过程中，权重的更新，当输入数据x是正例，但是得到了x和权重的点积小于0，则将x加上权重进行权重的更新，当输入数据x是负例，但是得到了x和权重的点积大于0，则将权重减去x进行权重的更新。最后使模型的损失最小化得到最优的结果。如图Figure4，所示是该模型训练的结果。

	precision	recall	f1-score	support
-1.0	1.00	1.00	1.00	747
0.0	0.00	0.00	0.00	0
1.0	1.00	1.00	1.00	4827
accuracy			1.00	5574
macro avg	0.67	0.67	0.67	5574
weighted avg	1.00	1.00	1.00	5574

perceptron classifier accuracy is: 99.89%

Figure 4: 感知机模型准确率

4、逻辑回归模型

逻辑回归是一种用于处理二分类问题的统计学习方法，虽然它名字中包含“回归”，但实际上是一种分类算法。如公式（3）所示，如果输入的数据是大于0的值，通过函数计算出来的值是大于0.5的可以归为一类，反之输入数据是小于0的值，通过函数计算出来的值是小于0.5的可以归为一类，以达到二分类的问题。

$$g(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

0	0.95	0.36	0.52	747
1	0.91	1.00	0.95	4827
accuracy			0.91	5574
macro avg	0.93	0.68	0.74	5574
weighted avg	0.92	0.91	0.89	5574

Logistical classifier accuracy is: 91.17%

Figure 5: 逻辑回归模型准确率

5、神经网络

如代码块listing1所示，神经网络定义了两个全连接层，层间使用Relu函数作为激活函数，dropout概率表示以概率0.2进行丢弃，最后softmax层进行分类结果的归一化，概率最大的即是最后判别的结果。如图Figure6即为训练的结果。

Listing 1 viterbi算法

```
#定义神经网络模型
model = nn.Sequential(nn.Linear(V,int(np.sqrt(V))),
                      nn.ReLU(),
                      nn.Dropout(0.2),
                      nn.Linear(int(np.sqrt(V)),2),
                      nn.Dropout(0.2),
                      nn.Softmax(dim=-1))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       747
           1       1.00      1.00      1.00      4827

    accuracy                           1.00      5574
   macro avg       1.00      1.00      1.00      5574
weighted avg       1.00      1.00      1.00      5574

NN classifier accuracy is: 99.91%
```

Figure 6: 神经网络模型准确率

三、实验结果和结论

本次实验主要探讨了五种不同的分类模型在特定数据集上的性能表现，包括朴素贝叶斯、支持向量机（SVM）、感知机模型、逻辑回归和神经网络。每种模型都有其独特的原理和适用场景。

1、朴素贝叶斯模型：该模型基于贝叶斯定理和特征条件独立假设，通过计算给定类别下特征的条件概率来进行分类。它在文本分类等领域表现良好，尤其适用于特征之间相互独立的情况。然而，当特征之间相关性较强时，朴素贝叶斯的性能可能会受到影响。

2、支持向量机（SVM）：SVM是一种广泛使用的分类算法，它通过寻找一个能够将不同类别的样本分隔开的最大间隔超平面来实现分类。SVM在处理高维数据和非线性可分问题时表现优秀，同时对于小样本数据也有较好的泛化能力。

3、感知机模型：感知机是一种二分类的线性分类模型，它通过误分类点到超平面的距离来定义损失函数，并使用梯度下降法进行优化。感知机模型简单易懂，但在实际应用中往往存在收敛速度慢和分类效果不稳定的问题。

4、逻辑回归模型：逻辑回归虽然名字中包含“回归”，但实际上是一种用于处理二分类问题的分类算法。它通过sigmoid函数将线性回归的输出映射到[0,1]区间内，从而得到属于某个类别的概率。逻辑回归具有计算效率高、易于实现和解释性强的特点，但在处理非线性可分问题时可能表现不佳。

5、神经网络：神经网络是一种模拟人脑神经元工作方式的机器学习模型，具有强大的非线性拟合能力。通过构建多层神经网络并使用反向传播算法进行训练，可以实现对复杂数据的分类和预测。神经网络在处理高维数据和非线性问题时表现出色，但需要大量的训练数据和计算资源。

四、实验过程分析与建议

1、对于朴素贝叶斯模型，其优点在于文本分类等特征条件独立或近似独立的场景下表现优异，计算效率高。其缺点在于对于特征间存在强相关性的数据集，朴素贝叶斯的性能会受到影响。

2、对于支持向量机模型，其优点在于对于高维数据和非线性可分问题表现出色，泛化能力强，对异常值和噪声数据鲁棒性较好。其缺点在于对于大规模数据集，训练时间较长；参数选择对性能影响较大，需要调整优化。

3、对于感知机模型，其优点在于模型简单，易于理解和实现；对于线性可分数据集能快速收敛。其缺点在于对于非线性可分数据集表现不佳；收敛速度可能较慢，且分类效果不稳定。

4、对于逻辑回归，其优点在于计算效率高，易于实现和解释；对于二分类问题具有良好的性能。其缺点在于对于非线性可分问题表现不佳；对特征缩放敏感，需要进行预处理。

5、对于神经网络其优点在于具有强大的非线性拟合能力，能够处理复杂的数据分布；适用于大规模数据集和高维特征。其缺点在于需要大量训练数据和计算资源；模型结构复杂，调参困难；容易过拟合。