

2024-2025学年机器学习实践大作业

姓名：郭采昕

学号：2022040211

班级：智能2201

学院：信息科学与技术学院

指导教师：项德良

目录

[基于多类机器学习的数据集分类与预测实践 4](#_Toc185612678)

[一、CIFAR-10数据集背景及介绍 4](#_Toc185612679)

[二、所选方法概述 5](#_Toc185612680)

[1. 逻辑回归（Logistic Regression） 5](#_Toc185612681)

[2. 支持向量机（**S**upport **V**ector **M**achines） 6](#_Toc185612682)

[3. 随机森林算法（Random Forest） 7](#_Toc185612683)

[4. K-means聚类(Unsupervised Learning) 8](#_Toc185612684)

[5. 主成分分析方法（**P**rincipal **C**omponent **A**nalysis） 8](#_Toc185612685)

[三、结果展示与总结 9](#_Toc185612686)

[1. 基于逻辑回归的CIFAR-10数据集分类 11](#_Toc185612687)

[2. 基于SVM的CIFAR-10数据集分类 12](#_Toc185612688)

[3. 基于随机森林的CIFAR-10数据集分类 12](#_Toc185612689)

[4. 基于改进的随机森林算法的CIFAR-10数据集分类 12](#_Toc185612690)

[5. 基于K-means的CIFAR-10数据集分类 13](#_Toc185612691)

[总结 14](#_Toc185612692)

[基于深度学习的数据集分类与预测实践 14](#_Toc185612693)

[一、BBC文本数据集背景及介绍 14](#_Toc185612694)

[二、RNN(Recurrent Neural Network)介绍 15](#_Toc185612695)

[三、结果展示与总结 16](#_Toc185612696)

[实验过程 16](#_Toc185612697)

[总结 19](#_Toc185612698)

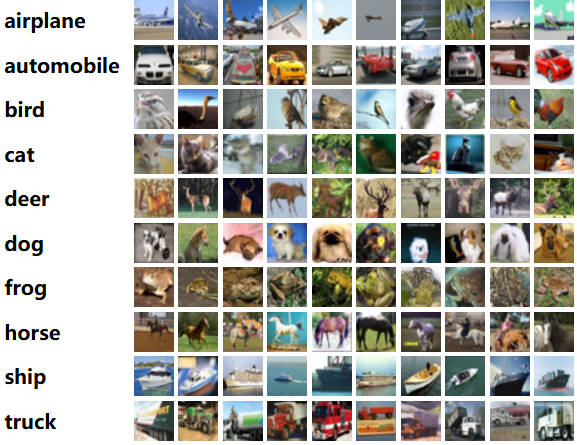
# 基于多类机器学习的数据集分类与预测实践

在该实验中，分别基于逻辑回归、支持向量机、随机森林三种有监督机器学习方法和一种无监督机器学习方法K均值聚类执行CIFAR-10的分类任务，输出各个方法得到的分类准确率、召回率、精确率，可视化其多分类混淆矩阵及ROC曲线进行对比，选择表现最好的算法，引入了降维、特征选择等方法并优化了模型参数，以提高算法分类效果。

## 一、CIFAR-10数据集背景及介绍

CIFAR-10数据集，由加拿大高级研究所（CIFAR）于2009年发布，是计算机视觉领域的重要基准数据集之一。该数据集由Alex Krizhevsky、Vinod Nair和Geoffrey Hinton共同创建，旨在推动图像分类技术的发展。

CIFAR-10包含了60,000张32x32像素的彩色图像，分为10个类别，每类6,000张。图像的尺寸为32x32像素，具有红、绿、蓝三个颜色通道，因此每张图像的数据量为3072字节。CIFAR-10的10个类别分别是飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车，如下图所示，它从每个类别当中随机截取了十张图片：



该数据集通常被划分为50000张训练图像和10000张测试图像，用于机器学习训练和评估。CIFAR-10数据集的文件结构相对简单，主要包括以下几个文件：batches.meta记录分类信息的元文件，包含类别名称与标签的对应关系；data\_batch\_1…data\_batch\_5为训练集的五个批次，每个批次包含10,000张图像的数据；test\_batch为测试集，包含10,000张图像的数据；readme.html为数据集介绍文件，提供数据集的基本信息和使用说明。

CIFAR-10数据集以其丰富的图像类别和适中的数据规模，成为计算机视觉领域入门的理想选择。它被广泛用于训练和评估图像分类任务中的深度学习模型，如卷积神经网络（CNN）、支持向量机（SVM）和其他各种机器学习算法。该数据集在类别方面的多样性和彩色图像的存在，使其成为机器学习和计算机视觉领域研究和开发的全面数据集。

*数据集下载地址：[CIFAR-10 and CIFAR-100 datasets](http://www.cs.toronto.edu/~kriz/cifar.html)*

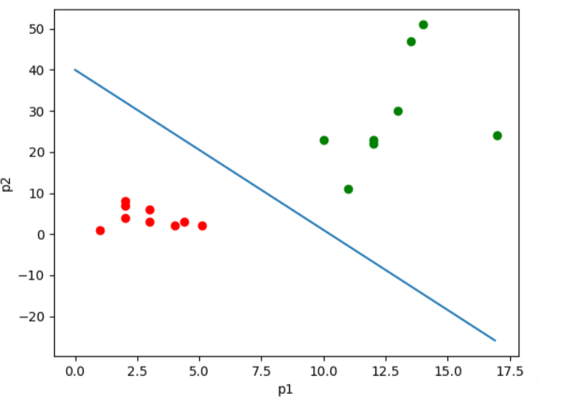
## 二、所选方法概述

### 1. 逻辑回归（Logistic Regression）

逻辑回归也称作logistic回归分析，是一种广义的线性回归分析模型，属于机器学习中的监督学习。其推导过程与计算方式类似于回归的过程，但实**际上主要是用来解决二分类问题（也可以解决多分类问题）**。通过给定的n组数据（训练集）来训练模型，并在训练结束后对给定的一组或多组数据（测试集）进行分类。其中每一组数据都是由p 个指标构成。

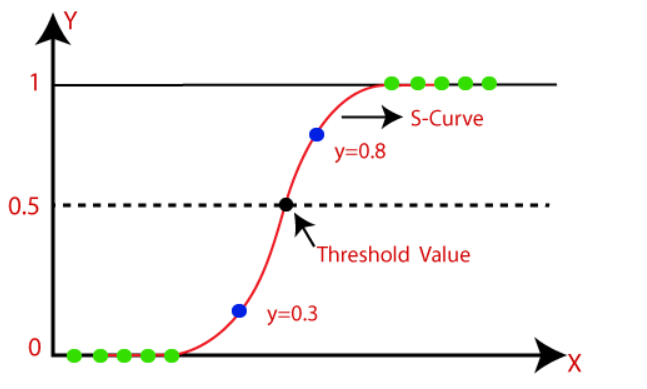
逻辑回归是用来进行分类的。例如，我们给出一个人的 [身高，体重] 这两个指标，然后判断这个人是属于”胖“还是”瘦“这一类。对于这个问题，我们可以先测量n个人的身高、体重以及对应的指标”胖“,"瘦”，把胖和瘦分别用0和1来表示，把这n组数据输入模型进行训练。训练之后再把待分类的一个人的身高、体重输入模型中，看这个人是属于“胖”还是“瘦”。

如果数据是有两个指标，可以用平面的点来表示数据，其中一个指标为x轴，另一个为y轴；如果数据有三个指标，可以用空间中的点表示数据；如果是p维的话(p>3)，就是p维空间中的点。

从本质上来说，逻辑回归训练后的模型是平面的一条直线（p=2),或是平面（p=3)，超平面（p>3)。并且这条线或平面把空间中的散点分成两半，属于同一类的数据大多数分布在曲线或平面的同一侧。

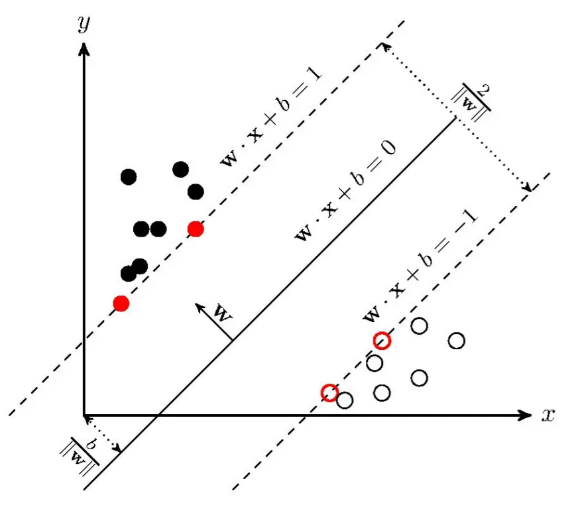
 如上图所示，其中点的个数是样本个数，两种颜色代表两种指标。这个直线可以看成经这些样本训练后得出的划分样本的直线。那么对于之后的样本的p1与p2的值，就可以根据这条直线来判断它属于哪一类了。

逻辑回归的实现基于线性回归模型，并在此基础上**引入了sigmoid函数作为映射工具**。Sigmoid函数的作用是将线性回归的输出（其值域原本为整个实数范围）巧妙地压缩至（0, 1）区间，从而该输出可以被解释为属于某一类别的概率。

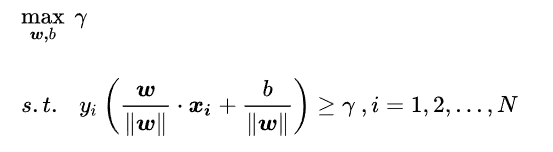
通俗地讲，逻辑回归模型首先通过线性函数计算出一个分数，然后使用sigmoid函数将这个分数转换为概率，最终根据设定的阈值（通常是0.5）来决定样本属于哪个类别。它使用**交叉熵损失函数**进行模型的训练，通常采用梯度下降法优化模型参数

### 2. 支持向量机（**S**upport **V**ector **M**achines）

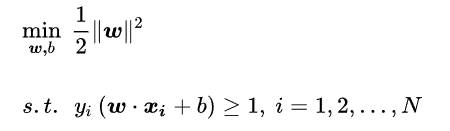
支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的**间隔最大的线性分类器**，间隔最大使它有别于感知机；SVM还包括**核技巧**，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的的学习算法就是求解凸二次规划的最优化算法。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是**几何间隔**（实际上就是所谓支持向量到超平面的距离）最大的分离超平面却是唯一的。

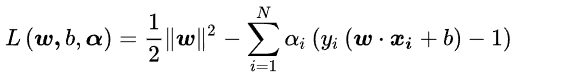
所以，SVM模型的求解最大分割超平面问题可以表示为以下约束最优化问题（即为几何间隔）：



等价于以下最优化问题：



使用拉格朗日乘子法得到其对偶问题：

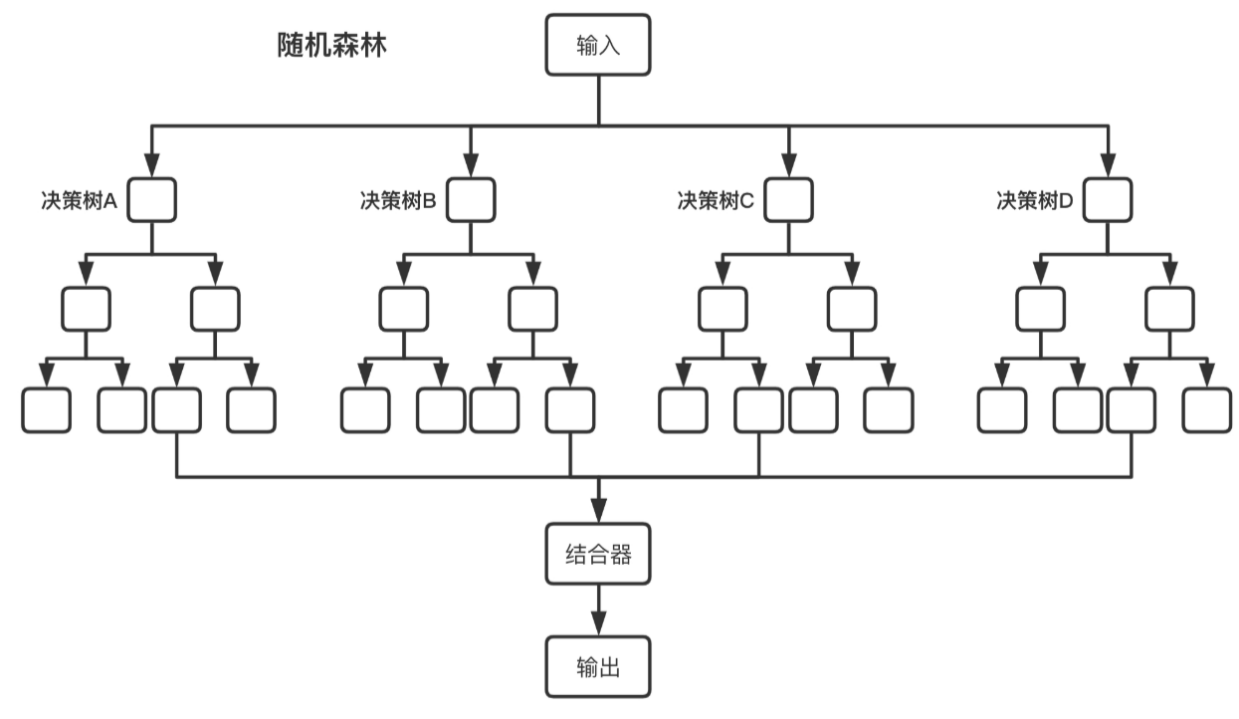


求解该问题可得到分类决策函数：



### 3. 随机森林算法（Random Forest）

随机森林由众多独立的决策树组成（数量从几十至几百不等），类似于一片茂密的森林。它通过汇总所有决策树的预测结果来形成最终预测，最终结果是通过对所有树的预测进行投票或加权平均计算而获得。

实际上，随机森林是一种**基于 Bagging（Bootstrap Aggregating）策略的集成学习**模型，它能够有效地处理非线性问题，并且擅长处理大量样本和特征。此外，随机森林还注重降低方差。Bagging 方法在训练过程中，各基学习器之间无依赖，可实现并行训练。通过集成多个模型，它可以有效地处理过拟合问题，提高模型的预测准确性和泛化能力。此模型可应用于分类、回归和异常检测任务。

上图展示了随机森林算法的具体流程，其中结合器在分类问题中，选择多数分类结果作为最后的结果，在回归问题中，对多个回归结果取平均值作为最后的结果。

  使用Bagging算法能**降低过拟合**的情况，从而带来了更好的性能。单个决策树对训练集的噪声非常敏感，但通过Bagging算法降低了训练出的多颗决策树之间关联性，有效缓解了上述问题。

随机森林的内在机理主要体现在其**随机性和集成性**。随机性来自于样本随机和特征随机，这使得每个决策树都有所不同，增加了模型的多样性。集成性则体现在通过投票或平均的方式，将多个决策树的预测结果结合起来，这可以有效地降低模型的方差，提高模型的稳定性和准确性。要让随机森林做出准确的预测，我们需要提供有用的信息，并且确保森林里的每棵小树都能独立地给出自己的看法。这样，当它们一起做决定时，我们就能得到一个既准确又可靠的答案。

此外，随机森林还具有一定程度的可解释性，可以通过特征重要性来解释模型的预测结果。这些特性使得随机森林在许多实际问题中都有很好的应用效果。

### 4. K-means聚类(Unsupervised Learning)

不同于以上三种有监督学习方法(supervised learning)，K-means聚类是一种无需已知样本标签的**无监督学习方法**(unsupervised learning)，目标是通过对无标记训练样本的学习来揭示数据的内在性质及规律，为进一步的数据分析提供基础。

K-means算法，又称为k均值算法。K-means算法中的k表示的是聚类为k个簇，means代表取每一个聚类中数据值的均值作为该簇的中心，或者称为质心，即用每一个的类的质心对该簇进行描述。

* 即，K-Means算法接受参数K；然后将样本数据集划分为K个聚类。获得的聚类需要满足：同一个聚类中的样本数据集相似度较高；而不同聚类中的样本数据集相似度较小。
* 算法思想为：以空间中K个点为中心进行聚类（即先从样本集中随机选取 k个样本作为簇中心），对最靠近他们的对象归类（所有样本与这 k个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中）。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

算法流程：

1. 先从没有标签的元素集合A中随机取K个元素，作为K个子集各自的质心。
2. 分别计算剩下的元素到K个子集质心的距离，根据距离将元素分别划分到最近的子集。
3. 根据聚类结果，重新计算质心(计算方法为子集中所有元素各个维度的算术平均数)
4. 将集合A中全部元素按照新的质心然后再重新聚类。
5. 重复第4步，直到聚类结果不再发生变化。

### 5. 主成分分析方法（**P**rincipal **C**omponent **A**nalysis）

PCA（principal components analysis）即主成分分析技术，又称主分量分析。其最主要的用途在于**“降维”**，通过析取主成分显出的最大的个别差异，发现更便于人类理解的特征。也可以用来削减回归分析和聚类分析中变量的数目。

在很多场景中需要对多变量数据进行观测，在一定程度上增加了数据采集的工作量。更重要的是：多变量之间可能存在相关性，从而增加了问题分析的复杂性。

如果对每个指标进行单独分析，其分析结果往往是孤立的，不能完全利用数据中的信息，因此盲目减少指标会损失很多有用的信息，从而产生错误的结论。

因此需要找到一种合理的方法，在减少需要分析的指标同时，尽量减少原指标包含信息的损失，以达到对所收集数据进行全面分析的目的。由于各变量之间存在一定的相关关系，因此可以考虑将关系紧密的变量变成尽可能少的新变量，使这些新变量是两两不相关的，那么就可以用较少的综合指标分别代表存在于各个变量中的各类信息。主成分分析与因子分析就属于这类降维算法。

PCA降维的主要步骤如下：

1. **标准化数据**：由于PCA对数据的尺度敏感，因此在进行PCA之前，通常需要对数据进行标准化处理，使得每个特征的均值为0，标准差为1。
2. **计算协方差矩阵**：计算数据的协方差矩阵，以确定各变量间的相关性。协方差矩阵是一个方阵，其元素表示变量间的协方差。
3. **计算协方差矩阵的特征值和特征向量**：对协方差矩阵进行特征分解，得到一组特征值和对应的特征向量。特征值表示每个特征向量的方向在数据中的重要性，而特征向量则表示PCA转换的方向。
4. **选择主成分**：根据特征值的大小对特征向量进行排序，选择前k个最大的特征值对应的特征向量，这些特征向量构成了新的特征空间，即主成分。
5. **构造投影矩阵**：将选定的特征向量作为列向量组成一个矩阵，这个矩阵称为投影矩阵。
6. **数据转换**：将原始数据通过投影矩阵进行转换，得到新的数据集，这个新数据集的维度就是选择的主成分的数量。

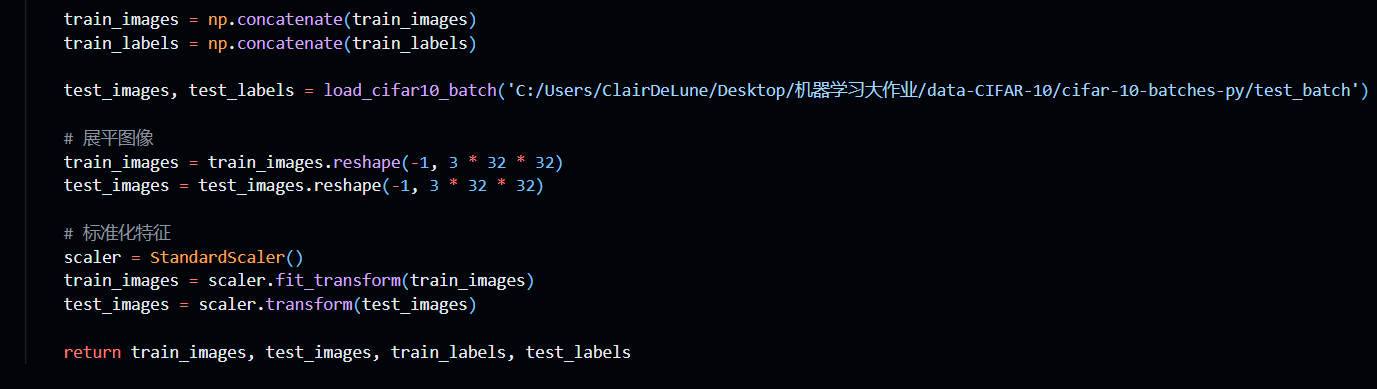
本实验采用PCA方法对数据进行降维以去除冗余信息，进而实现更好的分类效果。

## 三、结果展示与总结

在运用上述算法分类之前，首先要对数据集进行读入及简单处理。

相关代码如下：

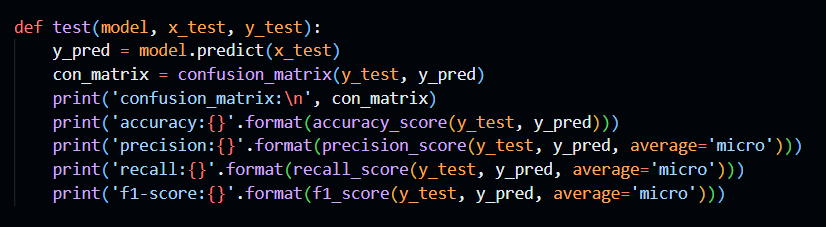




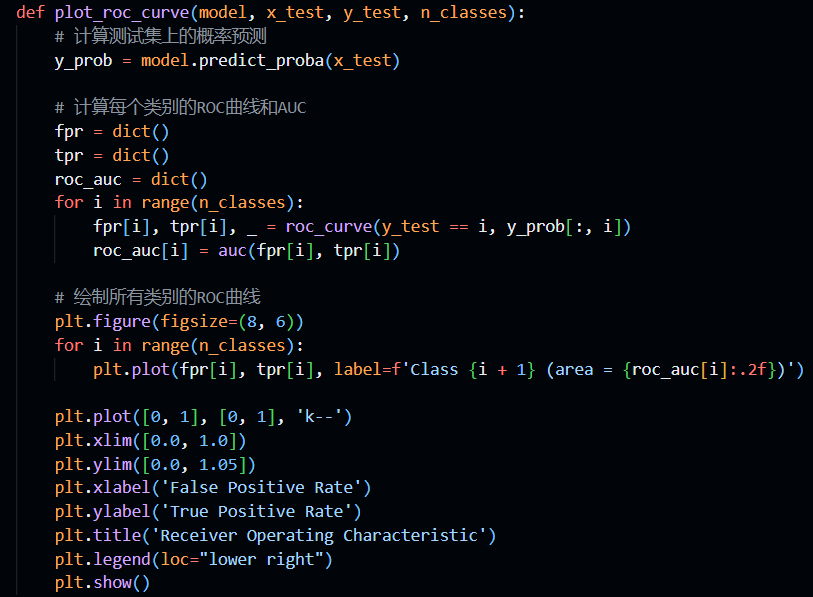
定义了一个loadDataSet函数，该函数可将原数据集中batch1至batch5分别读入训练数据、训练数据对应标签列表，将test\_batch读入测试数据。

为方便后续运用各类机器学习方法，首先预定义几个函数：

定义一个预测函数test，该函数需要三个参数——训练好的模型，测试数据集及测试数据对应的label，test可将训练好的模型用于预测测试数据的label并与真实值做对比，输出相应混淆矩阵confusion\_matrix，精度accuracy等评估指标。

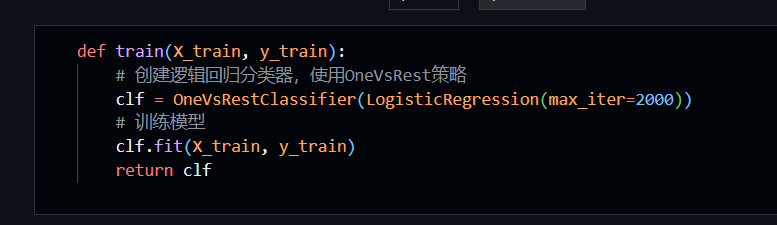


定义一个函数plot\_roc\_curve，实现ROC曲线的可视化。

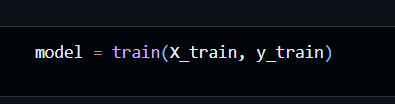


### 1. 基于逻辑回归的CIFAR-10数据集分类

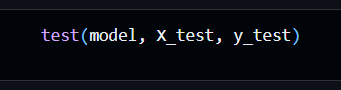
首先调用sklearn库中LogisticRegression函数创建逻辑回归分类器，使用OneVsRest策略实现多类分类。



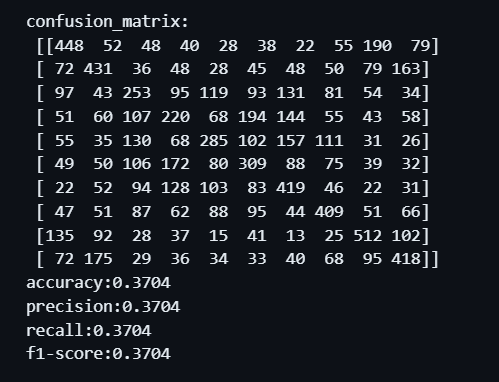
训练模型：



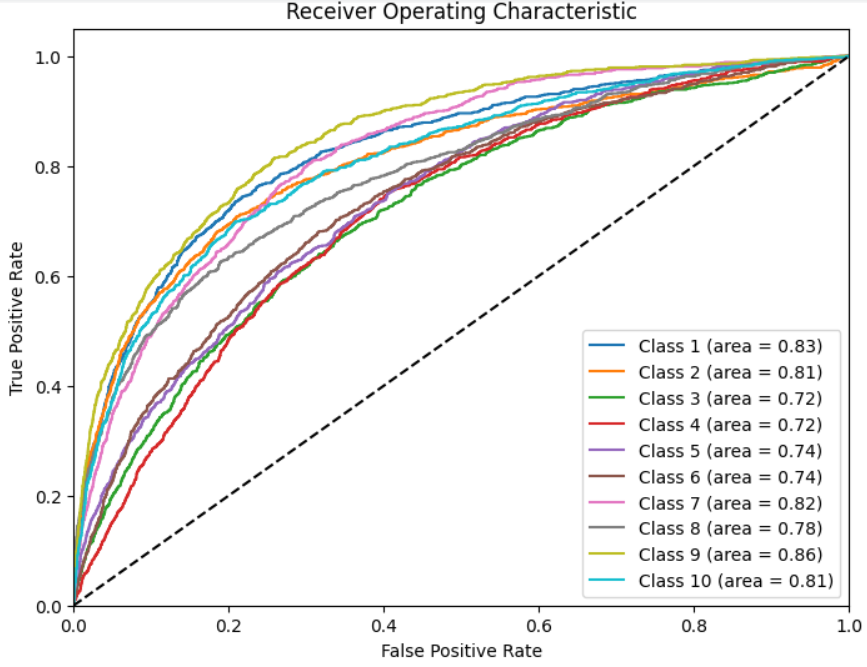
调用预定义的test函数进行预测：



模型分类结果如下：



绘制出相应ROC曲线图：

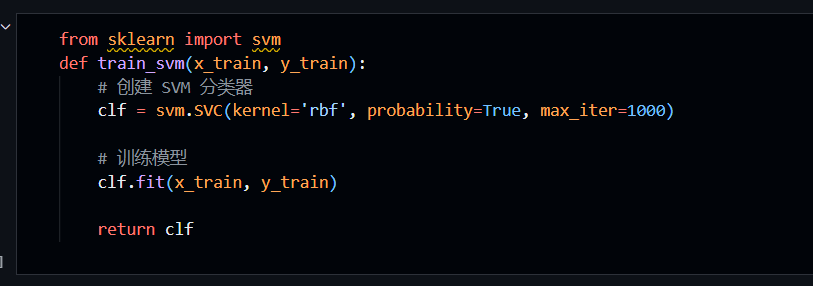


从预测结果可明显看出，模型发生了欠拟合，这是因为CIFAR-10中是较为复杂的图片数据，维度高且特征不易区分，逻辑回归在它上面很难表现出良好性能。

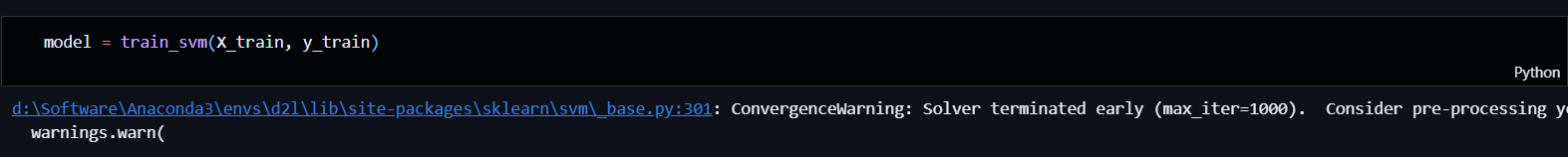
但仍可以通过调节相应参数提高精度，比如**将最大迭代次数max\_iter上调**，经过实验将其调节为5000，模型精度略有提升。

### 2. 基于SVM的CIFAR-10数据集分类

定义训练函数，调用sklearn库中的svm.SVC函数，使用rbf核函数，设置最大迭代次数max\_iter为1000

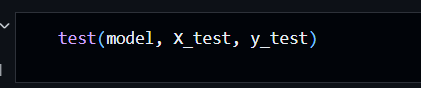


训练模型：

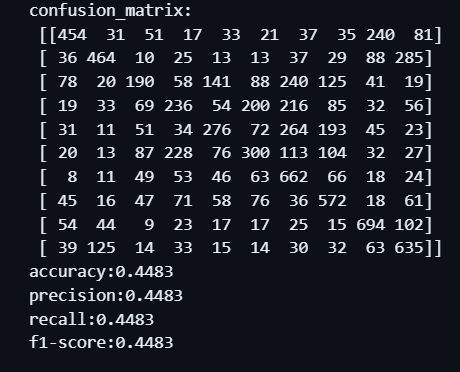


可以看到，SVM在训练过程中并没有收敛，只是因为达到了最大迭代次数才最终停止。

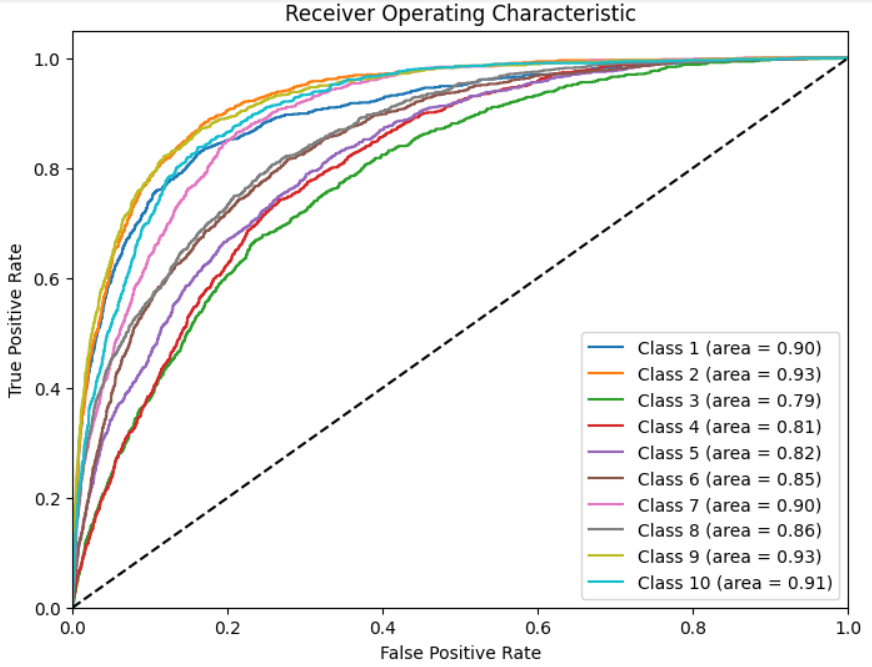
使用训练好的模型对测试数据进行预测：



分类性能如下：



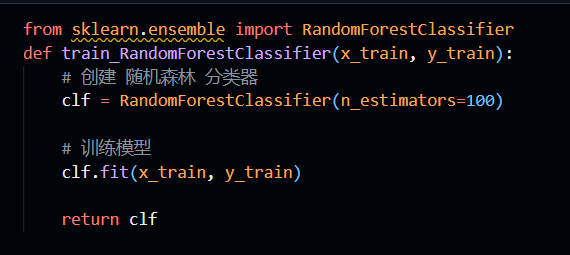
绘制出相应ROC曲线如下：



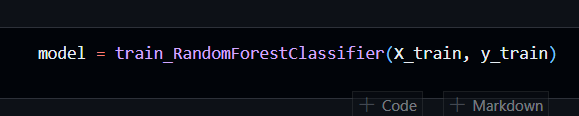
可以看出，模型最终分类精度为0.4483，略优于逻辑回归方法，但仍出现了欠拟合现象，在**上调max\_ite**r后精度略有提升。

### 3. 基于随机森林的CIFAR-10数据集分类

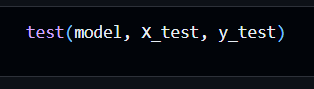
定义训练函数，调用sklearn库中的RandomForestClassifier函数，使用rbf核函数，设置基学习器数量为100



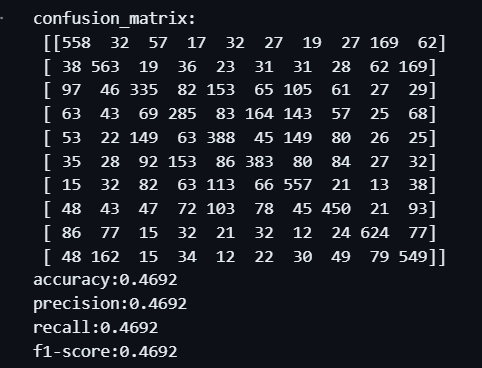
训练模型：



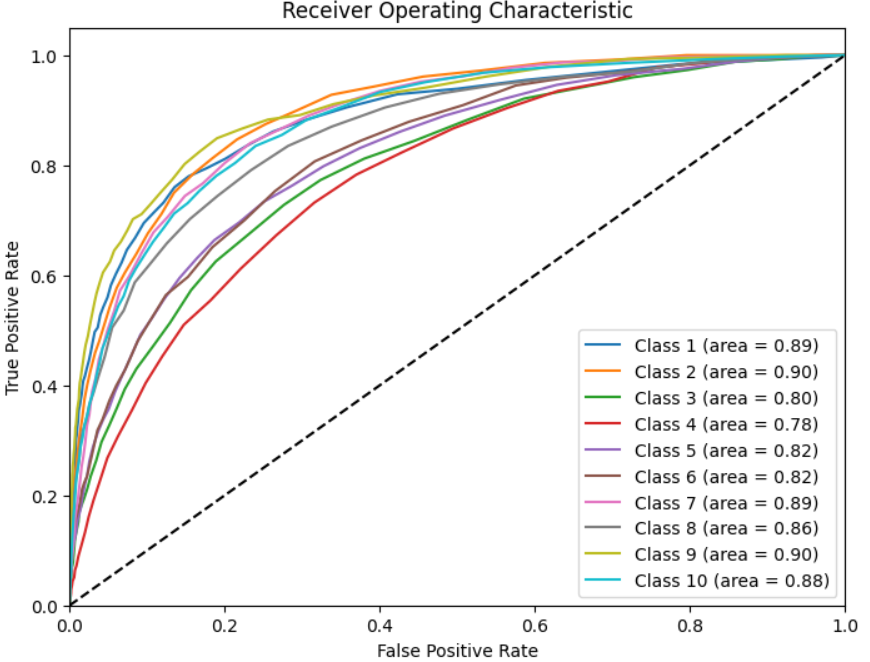
使用训练好的模型对测试数据进行预测：



分类结果如下：



绘制ROC曲线如下：



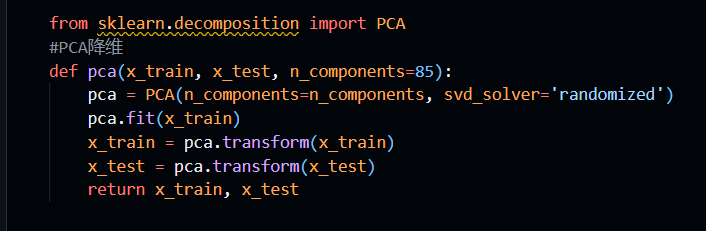
可以看出，分类精度为0.4692，仍然出现了欠拟合现象，但随机森林算法是目前精度最高的算法，且相比于运行了三个小时的SVM，随机森林仅仅只用了十分钟，所用时间成本也较小。

因此，接下来对随机森林算法进行改进，**引入PCA降维方法同时调节相应参数提升其分类性能。**

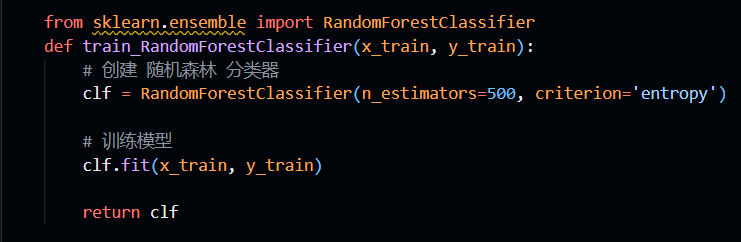
### 4. 基于改进的随机森林算法的CIFAR-10数据集分类

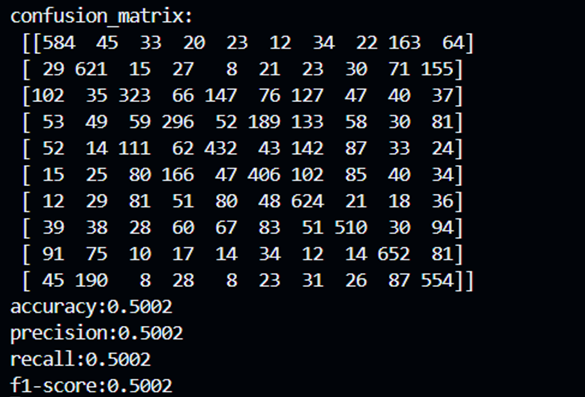
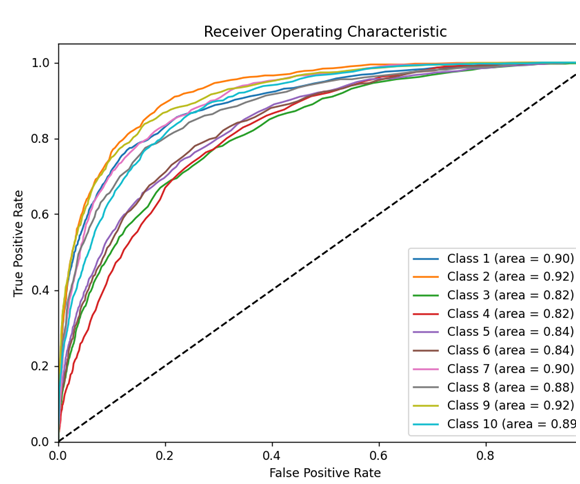
在上述算法中，随机森林以0.4692的分类精度取得了最好成绩，因此对其进行优化提升。

引入PCA降维方法对数据进行预处理，经过调节，将保留特征数设置为85左右时模型性能最佳。



同时重新设置随机森林分类器，将基学习器数量设为500，同时采用交叉熵损失函数作为评价指标：



对模型进行训练及测试，分类结果如下：

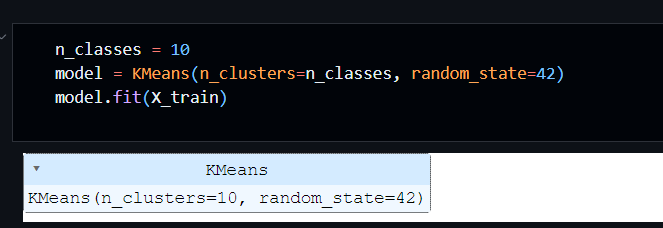
可以看到，改进后的随机森林算法分类精度为0.5002，相较于改进之前提升了3.1个百分点。

因此证明，PCA方法可以有效去除信息较少的特征以降低分类难度，且基学习器数量的多少对随机森林算法的性能有较大的影响。

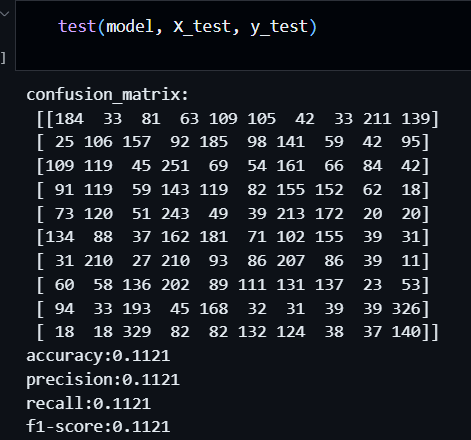
### 5. 基于K-means的CIFAR-10数据集分类

以上所用方法皆为有监督学习方法，接下来引入一种无监督方法——K-means聚类作为对比。

使用sklearn库中函数定义训练模型，需提前给出类别数：



使用测试数据集作预测：



最终分类精度为0.1121，同有监督方法相比性能大幅下降。这是因为CIFAR-10数据集本身便是一个复杂的多维图像多分类任务，单纯依靠K均值聚类很难得到良好的分类结果。、

### 总结

在本次实验当中，分别选用了有监督机器学习方法和无监督机器学习方法对CIFAR-10数据集进行分类。

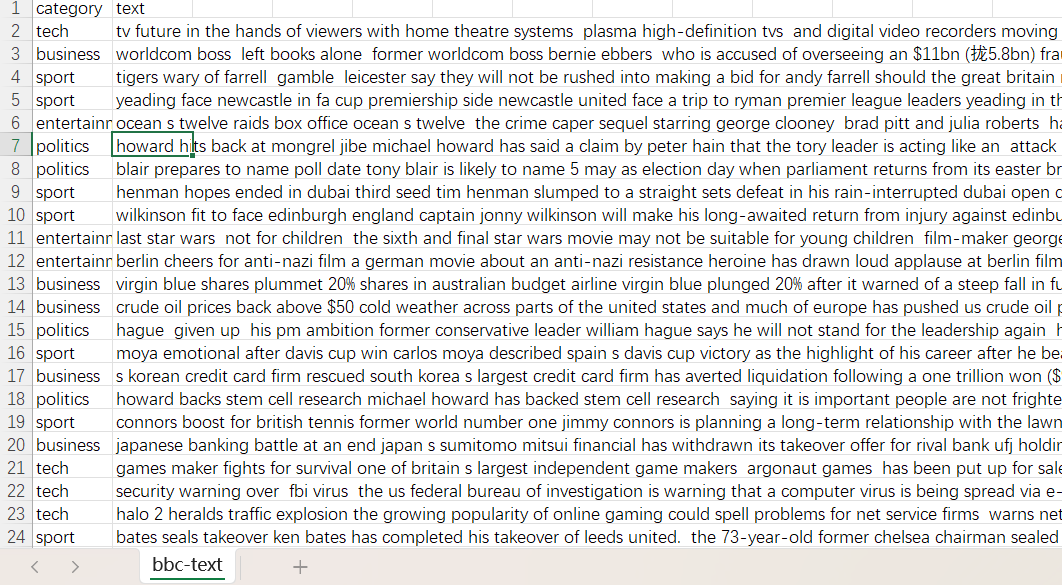
针对无监督学习方法，以K-means方法进行分类任务的训练及测试，最终分类精度仅为0.1121，出现了明显的欠拟合问题，而无论怎么调节相关参数或者对数据特征进行降维，模型性能都很难有大的提升。

针对有监督学习方法，分别使用了逻辑回归、SVM、随机森林方法。在这些方法中，迭代次数、选用的损失函数类型及随机森林中基学习器数量的大小都会对模型性能产生影响。最终随机森林的分类精度最高，因此对随机森林模型作了进一步优化，引入了PCA方法对数据集进行降维处理，同时调整了合适的基学习器数量大小，选用交叉熵损失函数，使得模型性能得到提高，最终达到0.5002的分类精度。

# 基于深度学习的数据集分类与预测实践

## 一、BBC文本数据集背景及介绍

BBC News数据集源自英国广播公司（BBC），该机构自1922年成立以来，一直是全球新闻报道的重要力量。BBC News数据集汇集了自2004年至2005年间的新闻文章，涵盖了五个主要类别：商业、娱乐、政治、体育和技术。这一数据集的创建旨在为自然语言处理（NLP）领域的研究提供丰富的文本资源，特别是在文本分类和情感分析方面。通过提供多样化和高质量的新闻内容，BBC News数据集极大地推动了相关研究的发展，成为NLP领域的重要基准数据集之一。



BBC News数据集在构建过程中面临了多个挑战。首先，新闻文本的多样性和复杂性使得数据预处理变得尤为重要，包括文本清洗、去重和标准化处理。其次，新闻内容的时效性和动态变化要求数据集的更新和维护必须及时，以确保其持续的相关性和有效性。此外，新闻文本中可能存在的偏见和主观性也是一大挑战，需要在数据分析和模型训练中加以考虑和处理。最后，如何有效地标注和分类新闻文章，以确保分类的准确性和一致性，也是该数据集面临的重要问题。

*数据集下载地址：*[*http://mlg.ucd.ie/datasets/bbc.html*](http://mlg.ucd.ie/datasets/bbc.html)

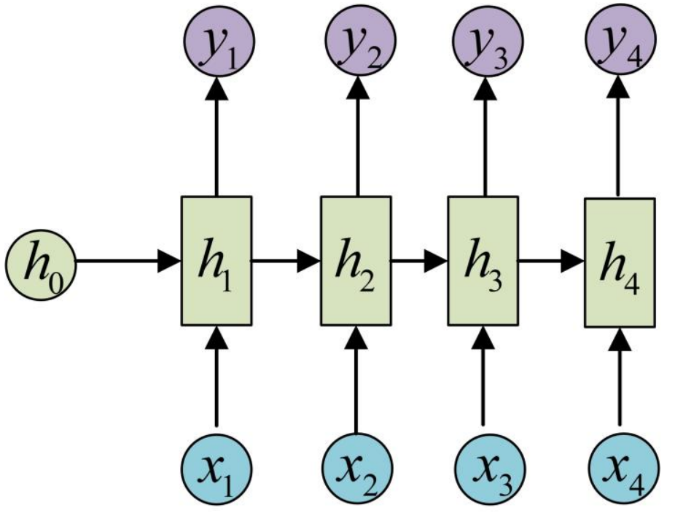
## 二、RNN(Recurrent Neural Network)介绍

循环神经网络（Recurrent Neural Network, RNN）是一类以序列（sequence）数据为输入，在序列的演进方向进行递归（recursion）且所有节点（循环单元）按链式连接的递归神经网络（recursive neural network）。

对循环神经网络的研究始于二十世纪80-90年代，并在二十一世纪初发展为深度学习（deep learning）算法之一，其中双向循环神经网络（Bidirectional RNN, Bi-RNN）和长短期记忆网络（Long Short-Term Memory networks，LSTM）是常见的循环神经网络。

循环神经网络具有记忆性、参数共享并且图灵完备（Turing completeness），因此在对序列的非线性特征进行学习时具有一定优势。循环神经网络在自然语言处理（Natural Language Processing, NLP），例如语音识别、语言建模、机器翻译等领域有应用，也被用于各类时间序列预报。引入了卷积神经网络（Convolutional Neural Network,CNN）构筑的循环神经网络可以处理包含序列输入的计算机视觉问题。

RNN用于处理序列数据。在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词，因为一个句子中前后单词并不是独立的。RNN之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。理论上，RNN能够对任何长度的序列数据进行处理。但是在实践中，为了降低复杂性往往假设当前的状态只与前面的几个状态相关。

下图是RNN 的一种基本结构：

参数更新公式如下：

这里的、和分别是权重矩阵，是激活函数，和是偏置项。

从公式可以看到，隐藏状态会不断地循环传递，这也是RNN能够“记忆”过去信息的原因。

## 三、结果展示与总结

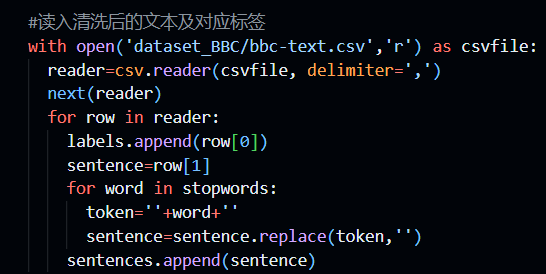
### 实验过程

在使用RNN对BBC文本数据集进行分类之前，首先要进行一定的数据预处理。

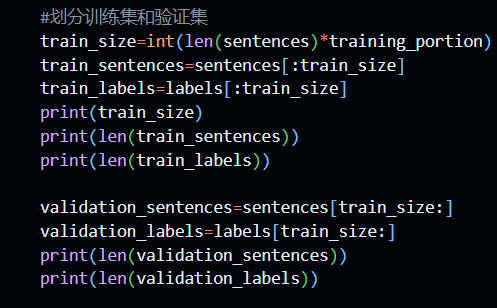
预定义词汇表大小及词汇表外单词占位符、词嵌入层维度、将一些不包含分类信息的词汇视为停用词以进行数据清洗，划分训练集、测试集占比为8：2。



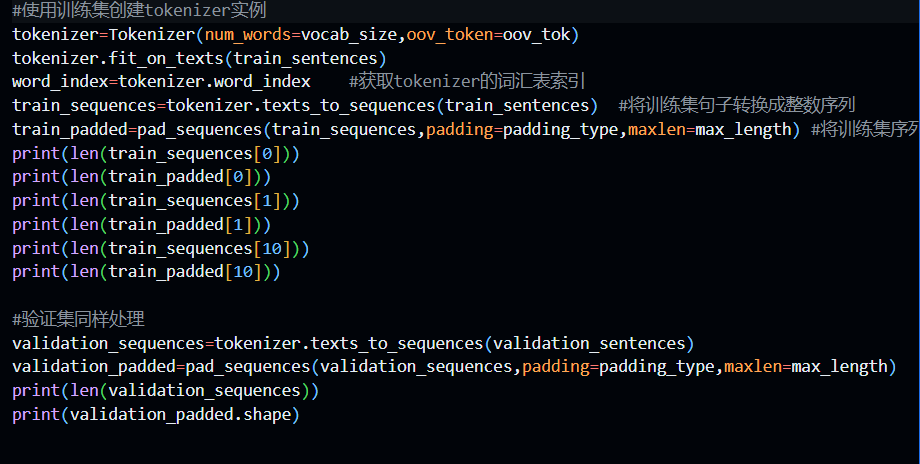
读入处理之后的文本及相应标签：

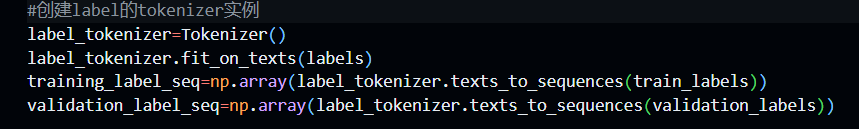


划分训练集、验证集：



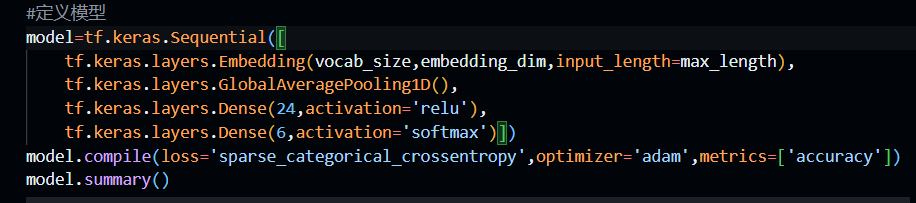
将文本数据集转换成Tokenizer实例，即将所有词汇映射成整数索引，再将所有文本映射至向量空间。



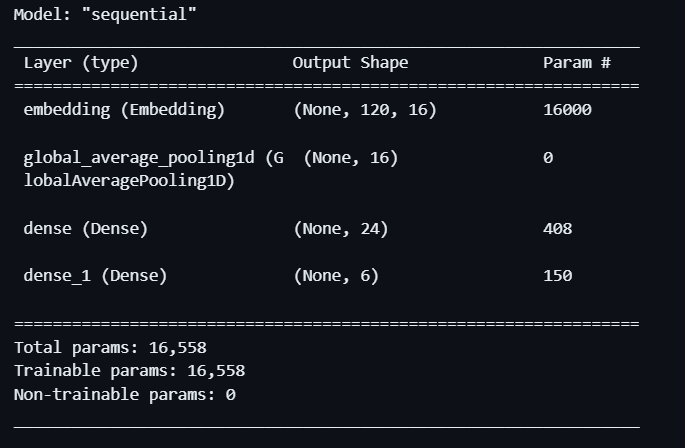


经过以上文本预处理之后，就可以使用RNN在训练数据集上进行训练并在验证数据集作预测。

定义RNN模型：

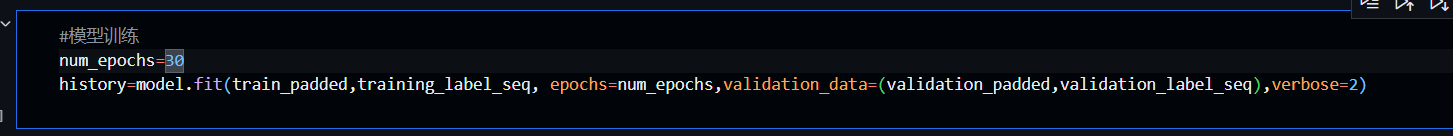


输出模型各层参数情况：

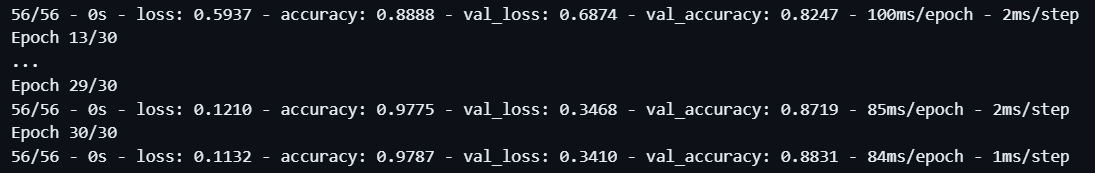


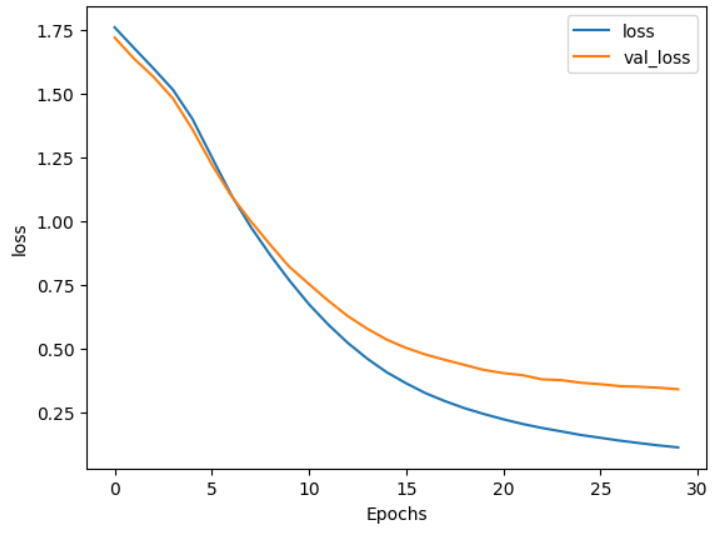
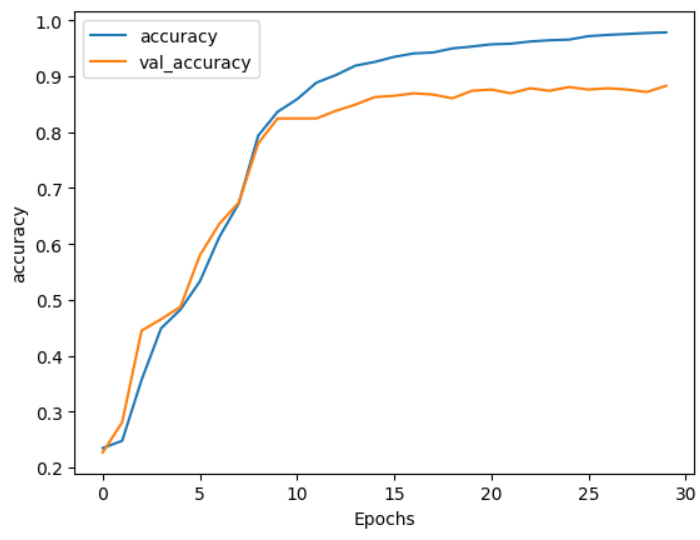
可以看到，使用的RNN分别由一个词嵌入层、一个全局池化层和两个全连接层组成，总参数有16558个。

使用划分好的训练数据集对模型进行训练，并使用验证数据集查看模型性能：



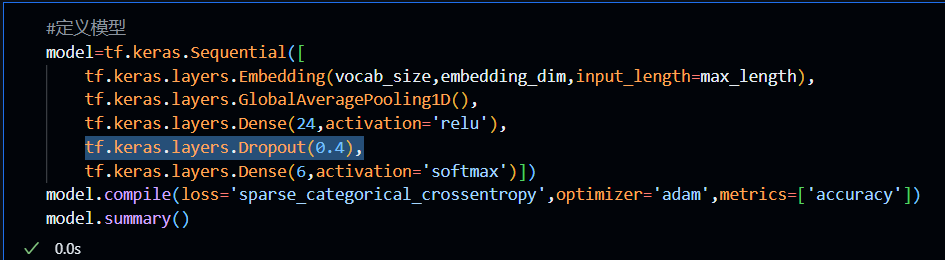
迭代三十轮后训练误差和精度分别为0.1132、0.9787，验证误差和精度分别为0.3412、0.8831。

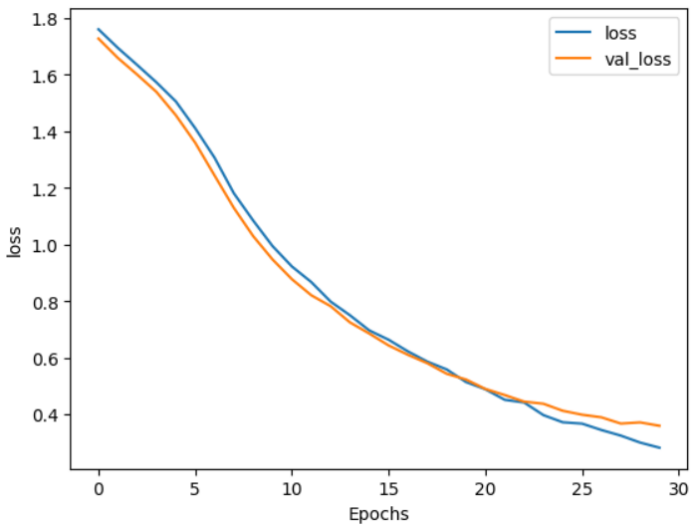
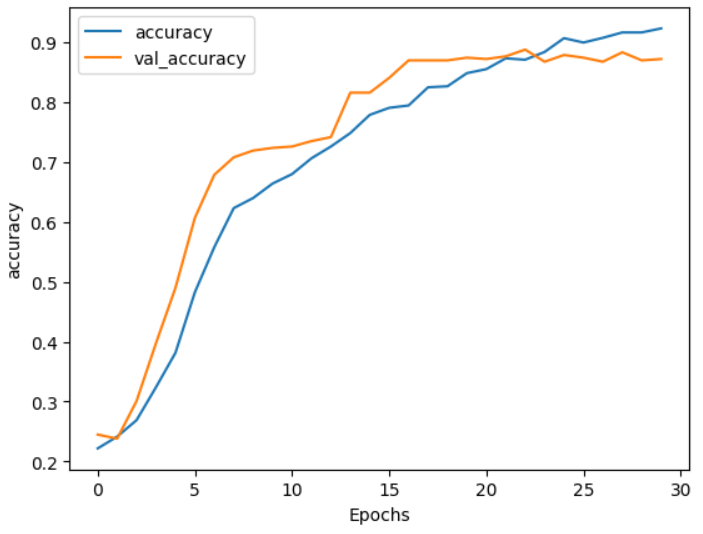


可视化loss和acc曲线：

可以看出，模型出现了过拟合现象，这可能是由于模型参数过多造成。

针对该问题，在模型中引入Dropout层，该层可以使权重矩阵变得稀疏：



最后可以看到，过拟合现象得到缓解：

### 总结

深度学习方法相比较于传统的机器学习方法，更强调于模型结构的深度及特征学习的重要性。它通常有5层、6层、甚至10层以上的隐层节点，同时它通过逐层特征变换，将样本在原空间的特征表示变换到一个新特征空间，从而使得分类或预测更加容易。

RNN，即循环神经网络，是一个在序列数据上表现良好的网络。在本实验中，选用了RNN作了BBC文本数据集的分类和预测任务，但由于模型参数过多出现了过拟合现象，通过引入Dropout层缓解了这一问题，最终模型的分类精度达到0.8831，表现出了良好的性能。