# 题目分析：

## 题目类型

图像分类问题

## 数据类型

png图像，具有RGB三通道。

共有12种图像，要求可以分类这十二种图像。

## 图像分析



### 每一张图片都包含一株绿色植物的叶片

### 可能影响图像分类准确率的因素：

#### 光照条件的明暗差别

#### 背景的石子和盆栽边框

#### 图像清晰图和阴影造成的图像缺失

#### 白色字条的影响

### 可能利用上的图像特征信息：

#### 叶脉的纹理

#### 叶片的形状

#### 叶片的个数

#### 叶片的颜色深浅

#### 叶片的大小等

## 评价方法：

### ：

区别于简单利用分类准确率进行评分。

是一种均衡了精度()和召回率()的综合评价指标。准确率和错误率的缺点在于一旦样本分布不平衡，错误率和准确率两个衡量指标便很难对模型性能进行有效衡量。而利用了精度和召回率这两个指标，弥补了准确率和错误率的缺点。

我们可以先明确几个概念：

#### *TP*（*True Positive*）：预测答案正确

#### *FP*（*False Positive*）：错将其他类预测为本类

#### *FN*（*False Negative*）：本类标签预测为其他标签

#### *TN*（*True Negative*）：被模型分类正确的负样本

根据以上的四个概念，我们可以得到如下的混淆矩阵：

|  |  |  |
| --- | --- | --- |
|  | *Predicted as Positive* | *Predicted as Negative* |
| *Labeled as Positive* | *True Positive* (*TP*) | *False Negative* (*FN*) |
| *Labeled as Negative* | *False Positive* (*FP*) | *True Negative* (*TN*) |

下面来介绍精度（）和召回率（）的概念：

#### 精度（）：指被分类器判定正例中的正样本的比重

#### 其公式为：

#### 召回率（）：指的是被预测为正例的占总的正例的比重

#### 其公式为：

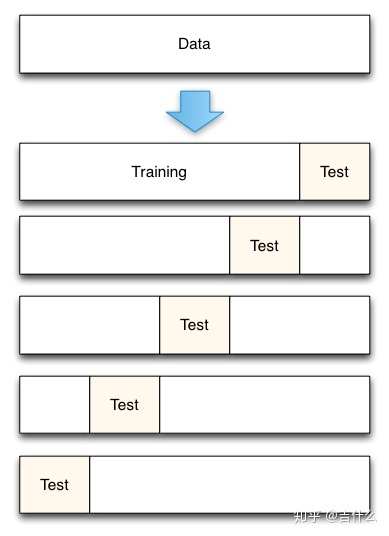
而每个类别下的的计算方式如下：

最后通过求得各类别下的求均值，得到最后的测评结果，计算公式如下：

就是一个精度和召回率的调和平均数。它可以在样本分布不平衡时对模型效果进行更加准确的分析。

#### 折交叉验证：

##### 流程图示：



##### 流程步骤：

###### 将原始数据集划分为相等的部分（“折”）

###### 将第部分作为测试集，其余作为训练集

###### 训练模型，计算模型在测试集上的准确率

###### 每次用不同的部分作为测试集，重复步骤B和C 次

###### 将平均准确率作为最终的模型准确率

##### 选择交叉验证的原因：

###### 传统*train/test split*进行模型评估的缺点：

*train/test split*是将原始数据集划分为训练集和测试集，以避免为了追求高准确率而在训练集上产生过拟合。而由于数据分布的不同，不同的划分会使模型的准确率产生明显的变化。而使得对于模型效果的评估不准确。

###### 的优点：

通过每次训练讲不通的部分作为验证集的方式：

对样本外的数据有更高的准确率。

可以更有效地发挥样本的作用。

从而使得交叉验证对于模型效能的评估更加准确

# 图像特征提取：

## HOG (Histogram Of Gradient)

### 基本概念

方向梯度直方图（*HOG*）特征是一种在计算机视觉和图像处理中用来进行物体检测的特征描述子。它通过计算和统计图像局部区域的梯度方向直方图来构成特征。*HOG*特征结合 *SVM*分类器已经被广泛应用于图像识别中，尤其在行人检测中获得了极大的成功。

*HOG*特征主要利用的思想为:在一副图像中，局部目标的表象和形状（*appearance and shape*）能够被梯度或边缘的方向密度分布很好地描述。

### 计算方法：

##### 某一点梯度的计算方法：

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

我们如果要求A点的梯度，则我们首先需要求其水平梯度和垂直梯度。根据上图我们可以得出，其水平梯度，。总的梯度强度值和梯度方向可以由下面的公式算得：

#### 计算梯度直方图：

##### 图像预处理

##### 我们可以对图像做以下两种变换

###### 伽马校正：减少光度对实验的影响。

###### 灰度化：将图像转化为灰度图像

##### 刚才我们可以看到，在每一个点都可以得到两个值，一个是梯度强度，一个是梯度方向。根据这两个值，我们可以画出一个梯度直方图。

例如，*HOG*是在一个的*cell*里面计算的，这样我们就可以得到一个个值，我们可以通过梯度直方图的方式将这个值变为个值，这样不仅可以减少运算量，还可以对光照条件拥有更多的鲁棒性。

##### 梯度方向的取值范围使度，按照梯度直方图，我们把， … 分为段，根据点梯度方向将梯度强度加入到该*bin*中，最后得到一个梯度直方图。

##### 归一化：

###### 归一化的目的是降低光照的影响。

###### 我们假设按之前的说法，一个*cell*可以得到长度为的向量，假设一个*Block*里有个*cell*，那么我们就可以得到大小为个向量，我们就对它进行归一化，归一化的方法为把向量中的每一个值都除以响亮的模长。

##### 得到*HOG*特征向量：

我们可以看一张图片中有多少的*Block*，每一个*Block*可以提取出的向量长度是多少，这样将他们组合在一起，我们就可以的到最后的*HOG*特征向量了。

#### *HOG*特征的优缺点：

##### 优点：

###### *HOG*表示的是边缘的结构特征，*HOG*能较好捕捉局部形状信息，对几何和光学变化都有很好的不变性。

###### *HOG*特征是在密集的采样图像块中求取的，特征中隐含了该块与监测窗口之间的空间位置关系。

##### 缺点：

###### 无法处理遮挡问题

###### 没有选取主方向，也没有旋转梯度方向直方图，所以没有旋转不变性。*HOG*本身也没有尺度不变性，对噪点也相对敏感。

## LBP (Local Binary Pattern)

### 基本概念

局部二值模式(LBP)是一种用来描述图像局部纹理特征的算子；它具有旋转不变性和灰度不变性等显著的优点，它可以用来进行纹理特征的提取。

### 计算方法：

#### 首先将检测窗口划分为的小区域（*cell*）；

#### 对于每个*cell*中的一个像素，将相邻的个像素的灰度值与其进行比较，若周围像素值大于中心像素值，则该像素点的位置被标记为，否则为。这样，邻域内的个点经比较可产生位二进制数，即得到该窗口中心像素点的*LBP*值；

#### 然后计算每个*cell*的直方图，即每个数字（假定是十进制数*LBP*值）出现的频率；然后对该直方图进行归一化处理。

#### 最后将得到的每个*cell*的统计直方图进行连接成为一个特征向量，也就是整幅图的*LBP*纹理特征向量；

### *LBP*特征的优缺点：

#### 优点：

##### 旋转不变性

##### 灰度不变性

#### 缺点：

##### 在平坦的图像中不够稳键。

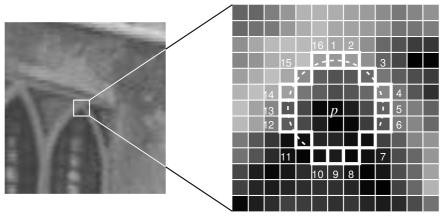
## ORB(Oriented Fast and Rotated Brief)

### 基本概念：

*ORB*是*Oriented Fast and Rotated Brief*的简称，可以用来对图像中的关键点快速创建特征向量。其中*FAST*和*BRIEF*分别是用来特征检测和向量生成的算法。

### 基本原理：

*FAST*是*Features from Accelerated Segments Test*的简称，可以快速选择关键点。给定数组中的像素，*FAST*将的亮度与周围小圆圈中的个像素进行比较。从而将圆圈中的像素分为三类（比亮、比暗或与相似）。如果超过个像素比更暗或更亮，则将其选为关键点。因此，通过快速找到的关键点为我们提供了确定图像边缘位置的信息。

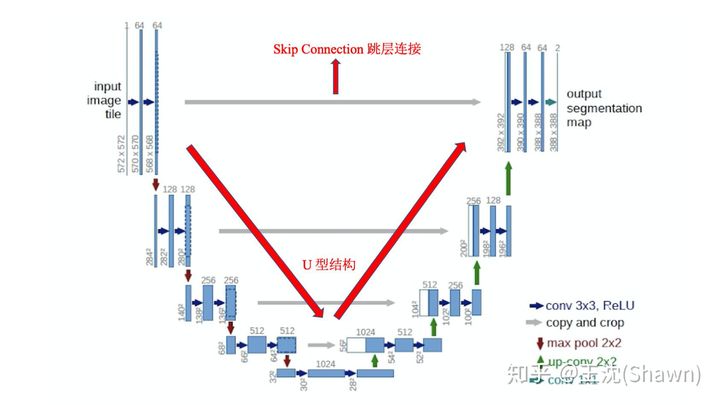


但是，*FAST*特征没有方向分量和多尺度特征。所以*ORB*算法使用多尺度图像金字塔。图像金字塔是由图像序列组成的单个图像的多尺度表示，所有图像序列为不同分辨率的同一图像。金字塔中的每一层都包含图像比前一层的下采样。一旦使用*ORB*算法创建了一个金字塔，它就会使用快速算法来检测图像中的关键点。通过检测每个级别的关键点*ORB*有效地定位不同尺度的关键点。因此，*ORB* 是部分尺度不变的。

*BRIEF*（亦即*Binary Robust Independent Elementary Feature*）将通过快速算法获取到的所有关键点，转换为二进制特征向量，使得它们能够共同表示一个对象。二元特征向量也称为二元特征描述符，是只包含和的特征向量。也就是说，每个关键点都由一个位字符串的特征向量描述。

## GRAY原图像

在这里我们之所以考虑到将原图像放入其中的原因是参照了Unet的思想。



当Unet网络对于图像压缩上有着比较好的效果，我们可以看到在压缩过程中会把压缩前图像作为参数传入，与压缩后的放在一起，可以防止特征丢失。所以我们选择将图像变为灰度图像后放入特征中，最后获得了两个百分点的准确率提升。

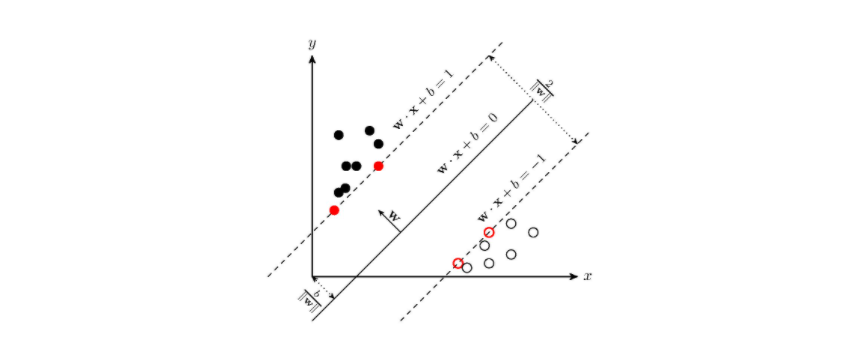
# 模型选择:

## SVM

*SVM*（*Support Vector Machine*）

支持向量机是一种二分类模型，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*算法原理：*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），但是几何间隔最大的分离超平面却是唯一的。

 存在一个最优的超平面，假设这个最优面的公式是：，那么对于所有的点集x，都存在平行于最优超平面，的点集的边界面或,这里可以归一化为。最大化这两个平行超平面的距离，即或者说是最小化，即 。另外一个条件是或.

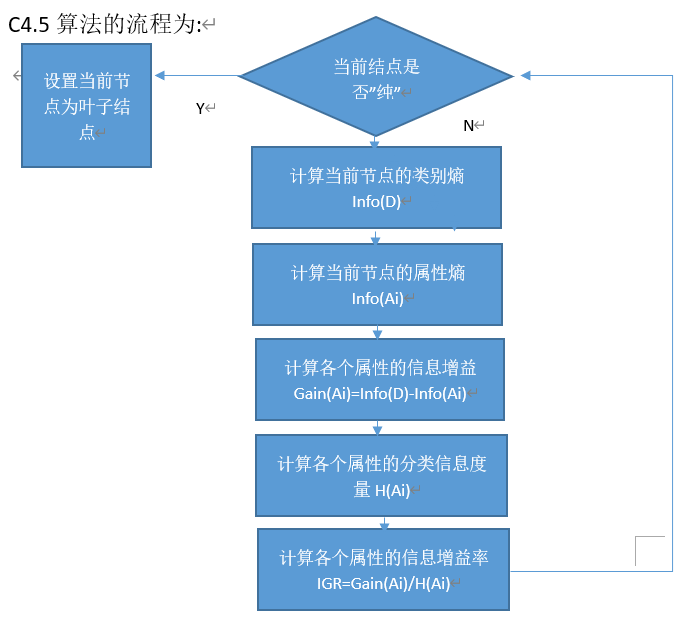
## Random Forest

随机森林指的是利用多棵树对样本进行训练并预测的一种分类器。

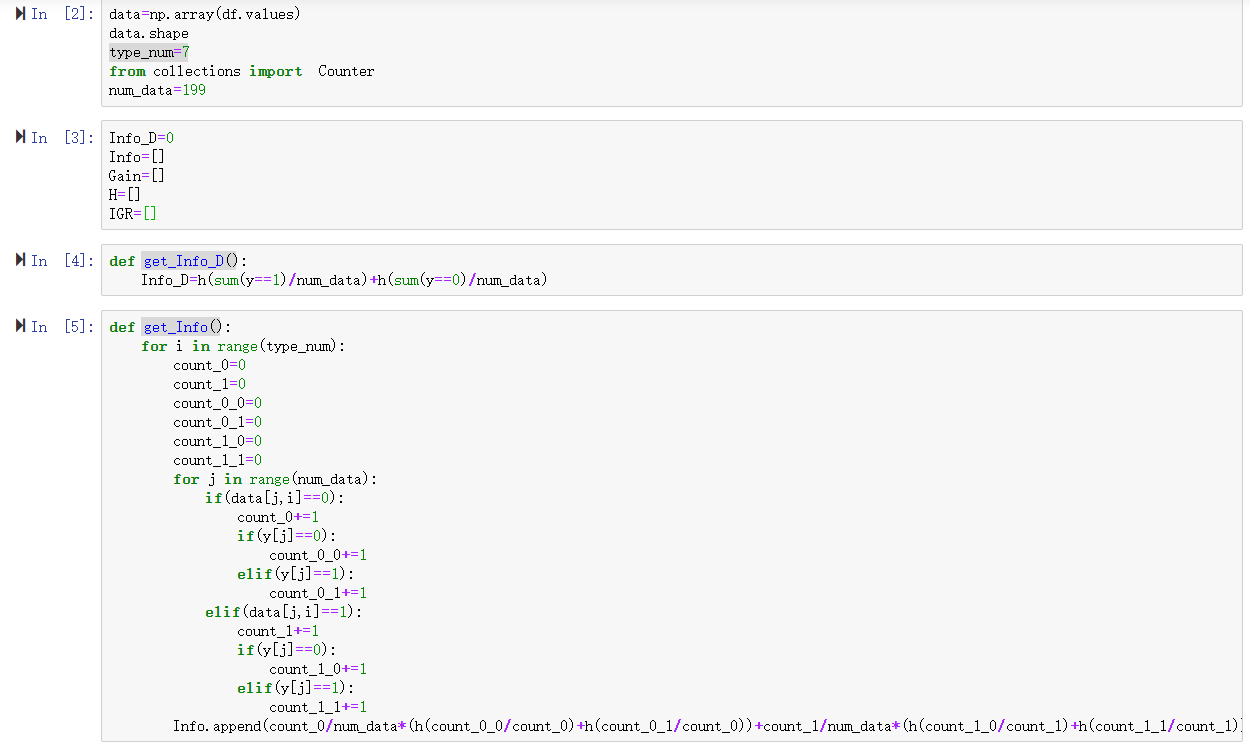
它的随机性主要包括数据的随机性和待选特征的随机性。对于随机森林的每一棵树，其生成时所选择数据集是原数据集的子数据集，而其选用的特征也是原特征集合的子集。这样我们就可以生成许多的决策树。最后通过决策树投票的方式获得分类结果。

对于每一课决策树来说，应用的是决策树分类。

对于结点划分，以决策树算法中*C4.5*为例：结点分裂的流程为



关于分裂属性的选择的代码如下：





如第11个模块的输出流程，我们只需要执行这几个过程，就可以得出当前结点的分裂属性了。

## XGBoost和light XGBoost

*XGBoost*和*light XGBoost*分别是*GBDT*（*Gradient Boost Decision Tree*)的具体实现和轻量化实现。这三者的基本思想类似，我们以*XGBoost*的应用举例。

算法的核心思想：

1. 不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。
2. 当我们训练完成得到棵树，我们要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数
3. 最后只需要将每棵树对应的分数加起来就是该样本的预测值。

我们最后的实际输出为每一棵树上的输出的累积。

为第个决策树的输出

我们需要输出的预测值 尽可能接近于真实值，同时提高其泛化性。

关于 的选择上，我们需要使目标函数损失尽可能地小，在*GBDT*中，目标函数*j*可以表示为

我们需要在损失函数下，找到最小化函数的值。

*XGBoost*中树的生长：

原作者给出了一个简单粗暴的方法：利用贪心法枚举出不同树的结构。不断地枚举不同树的结构，利用打分函数找到一个最有结构的树加入到模型中，并不断重复这一操作。而且利用贪心算法每次找到的都是*loss function*的最小值。为了限制树的过拟合，还可以加入阈值，只有当分裂后的增益大于阈值才会分裂。

树停止循环生成的条件：

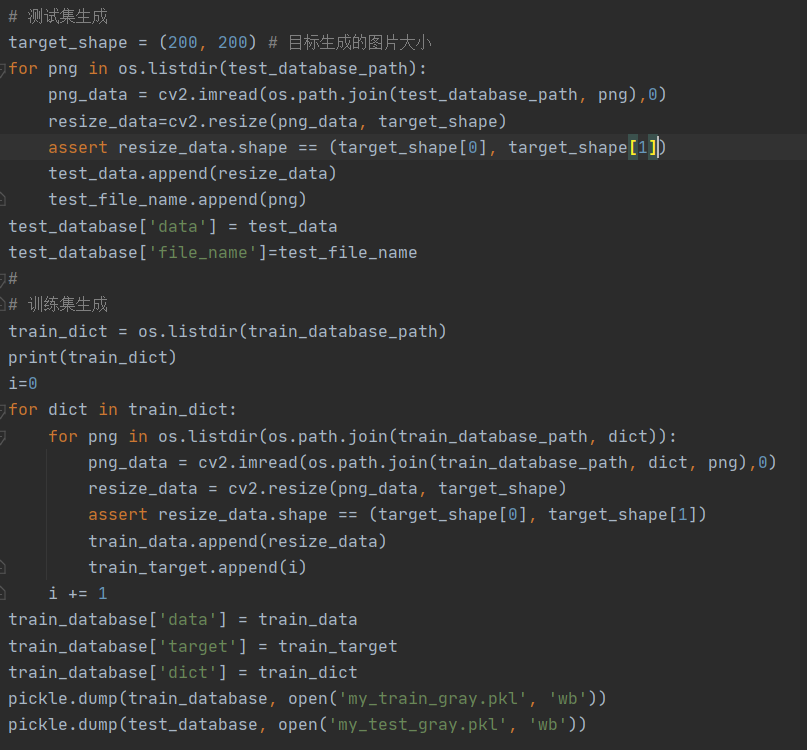
1. 分裂增益小于阈值
2. 树的深度达到最大值，防止过拟合
3. 样本权重和小于阈值（叶子节点的样本太少，防止过拟合等）

*XGBoost*的主要优点：

1. *XGBoos*t中的基学习器除了可以是*CART*（*GBTree*）也可以是线性分类器（*GBLinear*）
2. 正则化项​​​​​​​防止过拟合
3. *XGBoost*不仅使用到了一阶导数，还使用二阶导数，损失更精确，还可以自定义损失
4. 考虑了训练数据为稀疏值的情况，可以为缺失值或者指定的值指定分支的默认方向，这能大大提升算法的效率

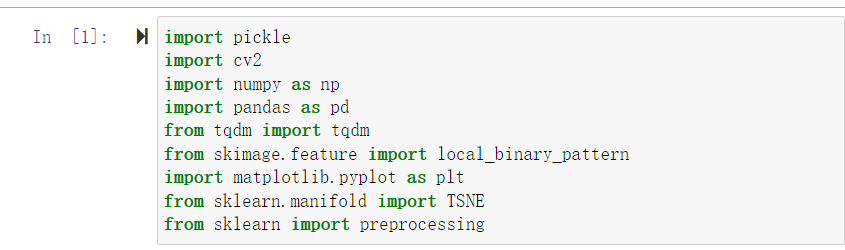
# 数据处理和分类流程：

## 数据打包



为了后期方便导入数据，我们将全体训练集和测试集打包。

## 导入模块



## 数据库导入

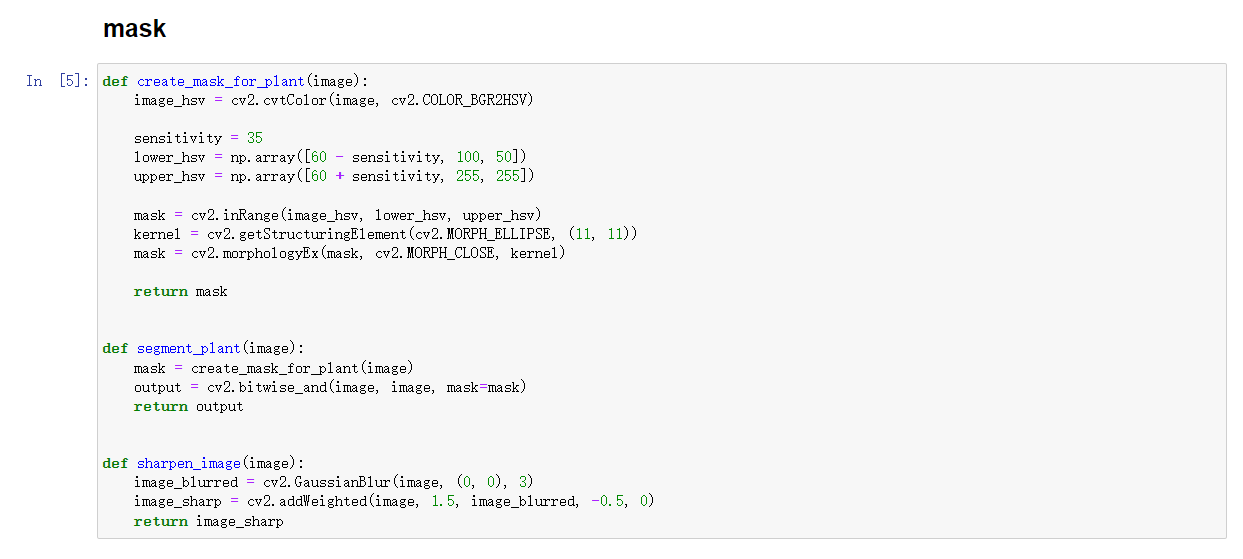


## 数据增强



因为我们采取的部分特征没有旋转不变性和尺度不变性，所以在代码中我们进行图像翻转、图像旋转、图像放缩的方式对数据进行拓展。这样会对准确率有提升的作用。

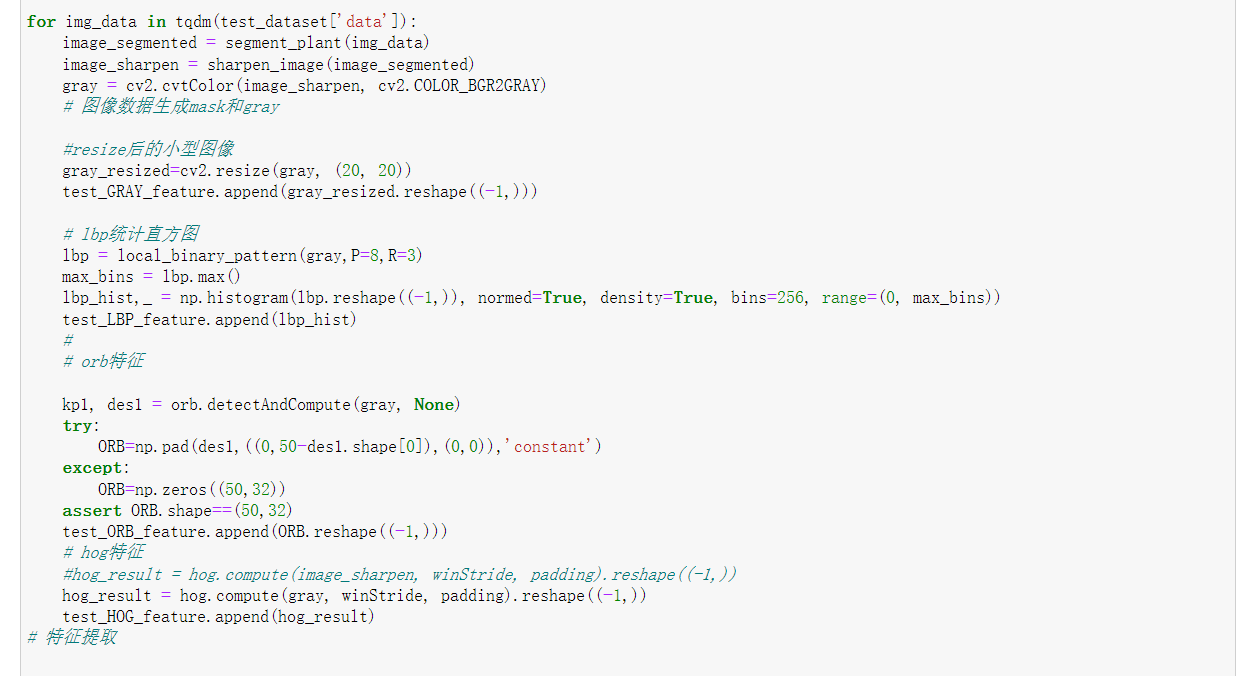
## 生成mask



## 特征提取

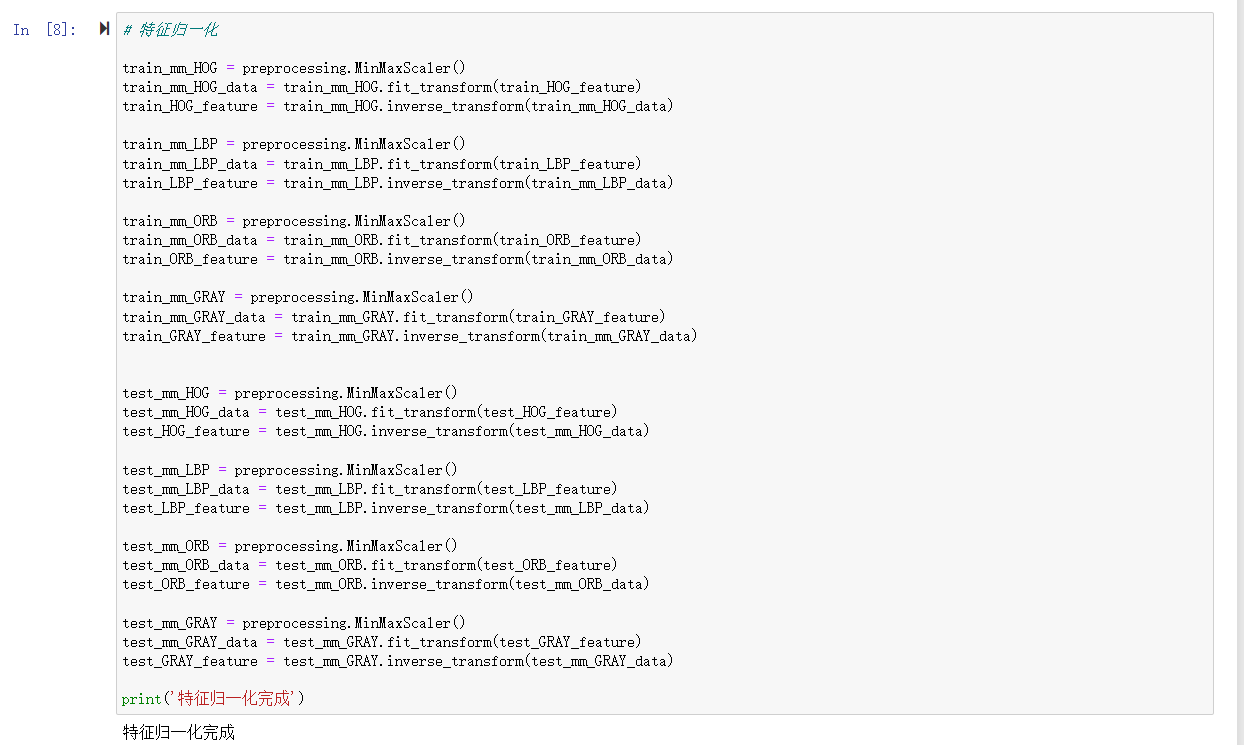






我们按照之前的思路，提取了HOG、LBP、ORB和GRAY原图像信息。

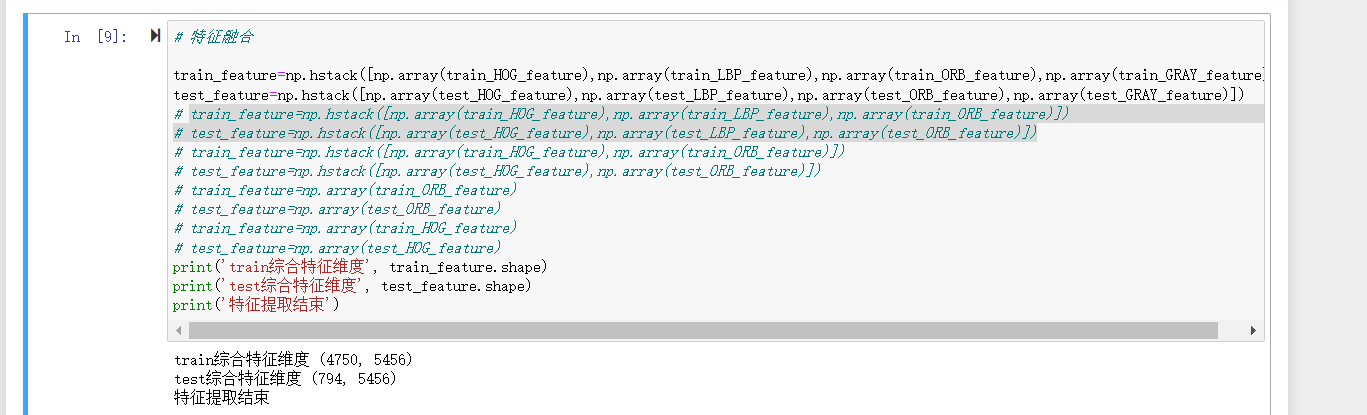
## 特征归一化



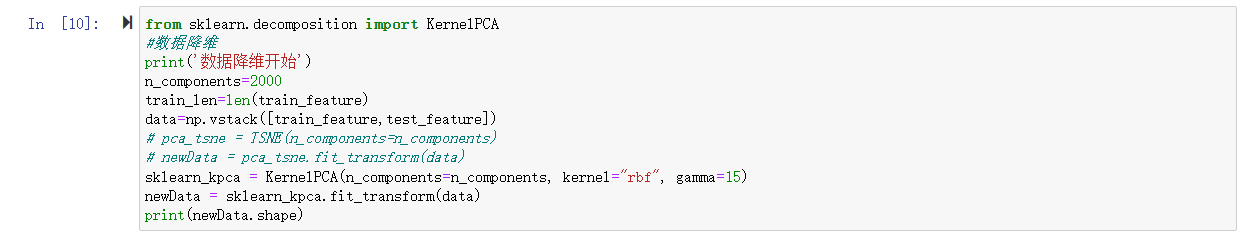
特征归一化的意义：

在机器学习领域中，不同评价指标往往具有不同的量纲和量纲单位，这样的情况会影响到数据分析的结果，为了消除指标之间的量纲影响，需要进行数据标准化处理，以解决数据指标之间的可比性。原始数据经过数据标准化处理后，各指标处于同一数量级，适合进行综合对比评价。

## 特征融合



## 特征降维



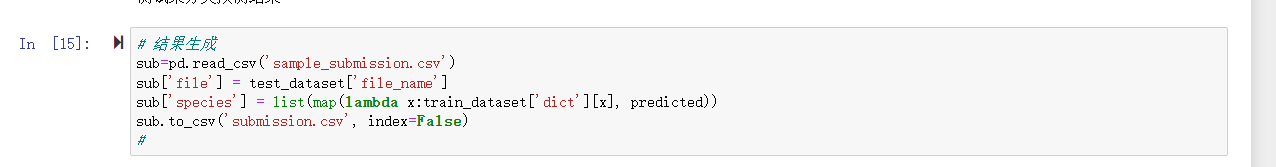
在这里特征降维我们尝试了PCA和tSNE降维两种方法，但是并没有获得良好的效果。

## 模型训练



我们尝试了多种模型，其中我们发现XGBoost模型的效果最好，之后我们在XGBoost上进行调参，最后获得比较好的效果。

## 结果存储



最后我们获得了78.589%的准确率：

