小A创新创业团队寒假训练营任务二

学习要求

- 了解机器学习中简单的分类问题
- 训练集(train)、验证集(val)和测试集(test)的划分
- 模型的评估
- 了解图像基本信息

> 本次的小任务大家可以使用任何的方法，但是基本的数据划分以及评估做好即可

任务2

1. 数据集的划分

数据集存放在marvel文件夹中，该文件夹中包含以下8个子文件夹：

- black widow
- captain america
- doctor strange
- hulk
- ironman
- loki
- spider-man

每个文件夹中含有若干张jpg图片，文件夹的名字即为图片里包含的人物。
现在，请将marvel文件夹里的照片按照8:1:1的比例随机划分成训练集、验证集和测试集。

- 训练集存放在train文件夹中，验证集存放在val文件夹中，测试集存放在test文件夹中
- train、val和test文件夹存放在data文件夹中 【!】注意：train、val和test文件夹中仍然包括上面所述的8个子文件夹

> 提示：数据集划分的方法有很多，这里可以引入一些python的基本库会简单很多

2. 灰度图像的转换

使用numpy等相关库读数据集中的任意一张照片数据，并计算其灰度图像和灰度值的方差(建议手敲锻炼numpy熟练程度)

- 一文详解图像中通道相关知识
- 彩色图像转化为灰度图像的方法

> 提示：可以使用opencv，Image等包进行图像像素点的提取

3. 分类任务

3.1 数据集的介绍

- 数据集存放在data文件夹中并且以npz格式进行储存，其中数据为一个二维矩阵。这个矩阵拥有150个单位，同时每个单位包含了4个特征信息以及一个标签信息(前4列为特征信息，最后一列为标签信息)。
- 其中标签信息含有3个类别，也就是说你需要训练出一个3分类器

3.2 任务要求

- 划分数据集，将数据集随机划分为训练集和测试集(由于数据量比较小，只用划分两个数据集即可，验证集可以不用做)
- 使用训练集的数据训练出一个3分类器，可以使用你想到的任何机器学习的算法。
- 最后使用测试集验证你使用训练集训练出来的模型，要求准确率达到90%以上。
- (选做)可以尝试使用不同的分类算法对数据集操作，然后进行对比，书写到提交报告中。

> 提示：至于机器学习算法，可以使用支持向量机，决策树，随机森林，线性回归等一系列方法，大家可以在网上进行搜索，python也有sklearn的包支持很多分类算法。如果大家打算实现某个算法，我十分推荐正规矩阵法(本质上就是高中所学的最小二乘法的矩阵形式)。

任务资料提示：正规矩阵法

4. 提交要求

上次任务中并没有要求大家提交，这次说一下提交要求，会要求大家在下学期面试之前提交即可（提交方式这边暂时还没商讨）

- 请使用markdown或者Latex编辑的pdf文档或者jupyter notebook编辑的ipynb文档(word排版很难看)
- 在文档中可以书写你遇到的任何问题、算法学习的过程与心得、自己的思考与理解以及使用某种算法解决上述问题的具体过程(会有具体的参考文档)
- 对于数据集划分任务和灰度图像转化，只需要提交具体代码，不需要提供数据集(师兄会将你代码放在自己的电脑上跑，请确保文件的相对路径正确)。
- 对于分类任务，需要上传代码以及pdf文档(到时候可以压缩文件，一次性提交上来即可)
- 小任务只是给大家练手的，看完具体学习路线给大家实践的机会，和下学期考核无关，大家放心大胆的提问和学习。