机器学习课程报告

学号：23210240272

姓名：石璞

# 任务描述

自行实现k最近邻算法（k-NN）算法。

使用k-NN进行基于MINST数据集的分类。

# 数据集描述

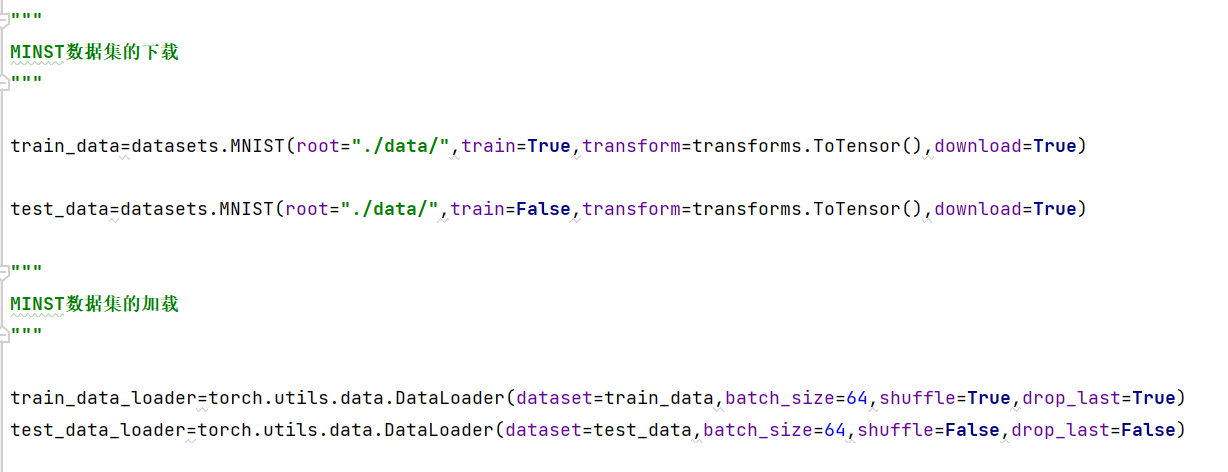
MINST数据集来自美国国家标准与技术研究所，训练集由来自250个不同人手写的数字构成，其中50%是高中学生，50%来自人口普查局的工作人员，测试集也是同样的比例。但测试集和训练集的作者不相交。

MINST数据集一共有7万张图片，其中6万张作为训练集，1万张是测试集。每张图片是28\*28的0-9数字。每张图片是黑底白字的形式，黑底用0表示，白字用0-1之间的浮点数表示，越接近1，颜色越白。

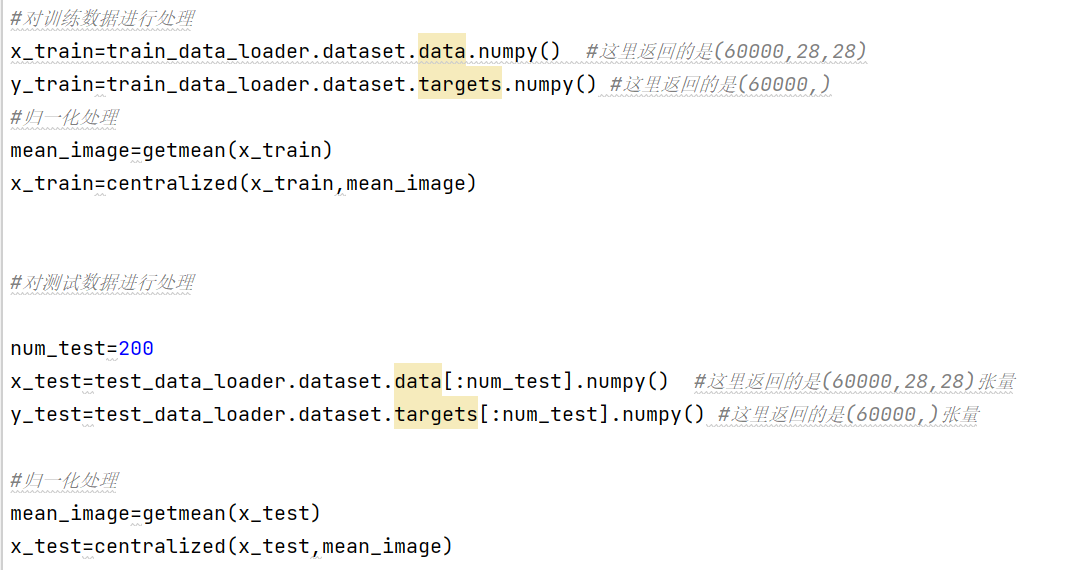
# 3.方法介绍

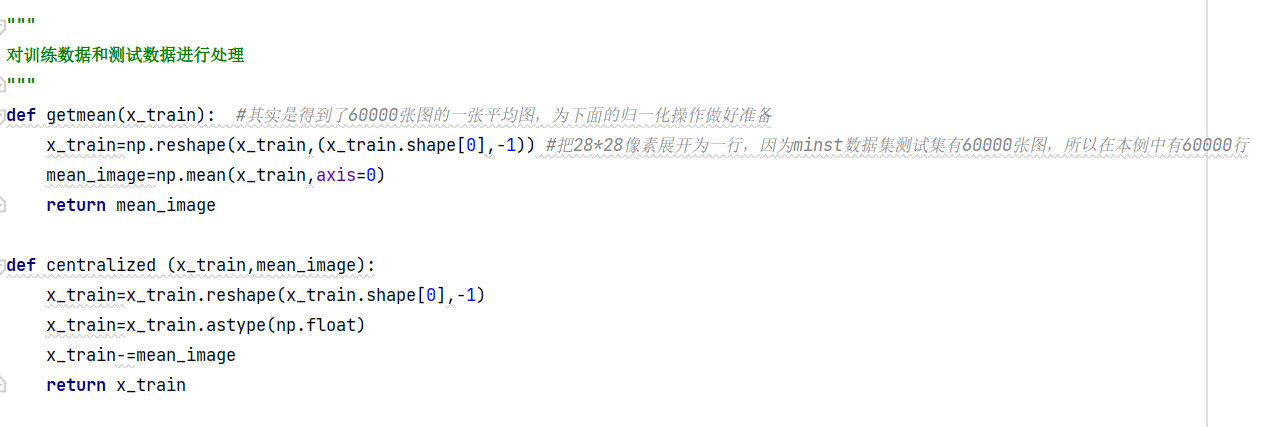
**3.1数据预处理**

**1.MINST数据集的读取和加载**

****

2.对训练数据和测试数据做进一步处理，涉及到getmean()函数以及centralized()函数。Getmean()函数是获取测试集、训练集的“平均图”。在本项目中实际上是分别生成了训练集的平均图和测试集的“平均图”。Centralized()函数是归一化函数，实际上是分别在训练集与测试集中对每张图片-“平均图”完成归一化操作。





**3.2算法描述**

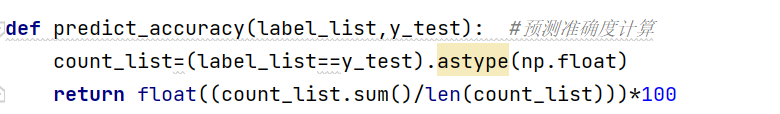
k-NN算法实质上利用距离，给定训练集数据以及标签（类别）。对于测试集中每一个点来言，分别求其到训练集每一个点的距离，选择前k个距离最小的点（k-NN中k的由来，其是一个超参数），考察其所属类别，这k个点中出现次数最多的类别即可认为是测试集中该点的所属类别。具体实现代码如下所示，核心的距离计算采用欧拉距离计算方法：



# 4. 实验结果分析

4.1 评价指标

采用预测准确度作为评价指标，即用预测标签与真实标签相比较，计算预测正确所占的比例。设计predict\_accuracy()函数进行评价。



4.2 定量评价结果

分别取k=1，3，5进行实验，计算分类准确程度



从上述实验结果来言，k-NN算法在简单数据集的分类上能够取得较好效果，同时值得注意的是超参数k的选择就目前实验结果上看k=1的时候效果最好，当k取值变大的时候反而准确性会下降，但总体来言准确性相差不大。

# 5. 总结

本次实验采用了k-NN算法对MINST数据集进行分类，采用的距离计算方法为欧拉计算方法，值得注意的是k-NN算法并没有训练过程。就实验结果来言，面对简单数据集k-NN算法可以取得较好的分类效果。在未来可以尝试采用其他算法分类MINST数据集并与k-NN方法比较。