机器学习课程报告

学号：23210240272

姓名：石璞

# 任务描述

利用KNN,SVM,mlp-nn等不同的分类算法对MINST数据集进行分类。

# 数据集描述

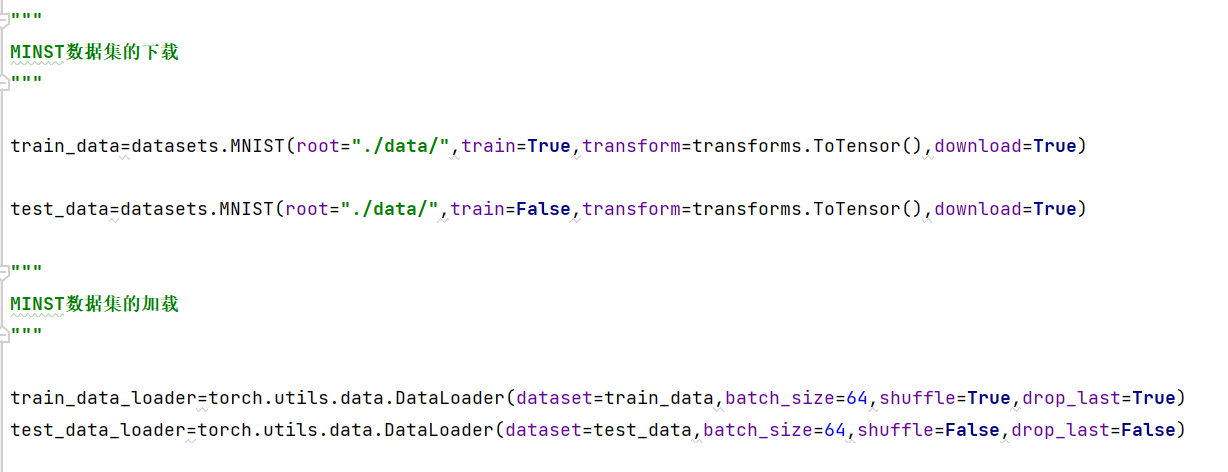
MINST数据集来自美国国家标准与技术研究所，训练集由来自250个不同人手写的数字构成，其中50%是高中学生，50%来自人口普查局的工作人员，测试集也是同样的比例。但测试集和训练集的作者不相交。

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。

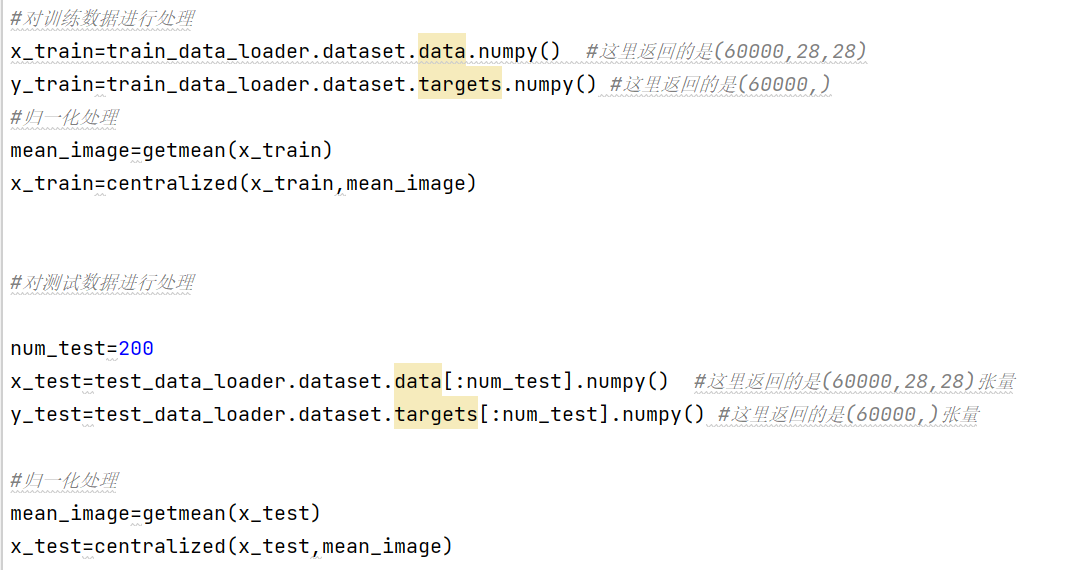
# 3.KNN方法介绍

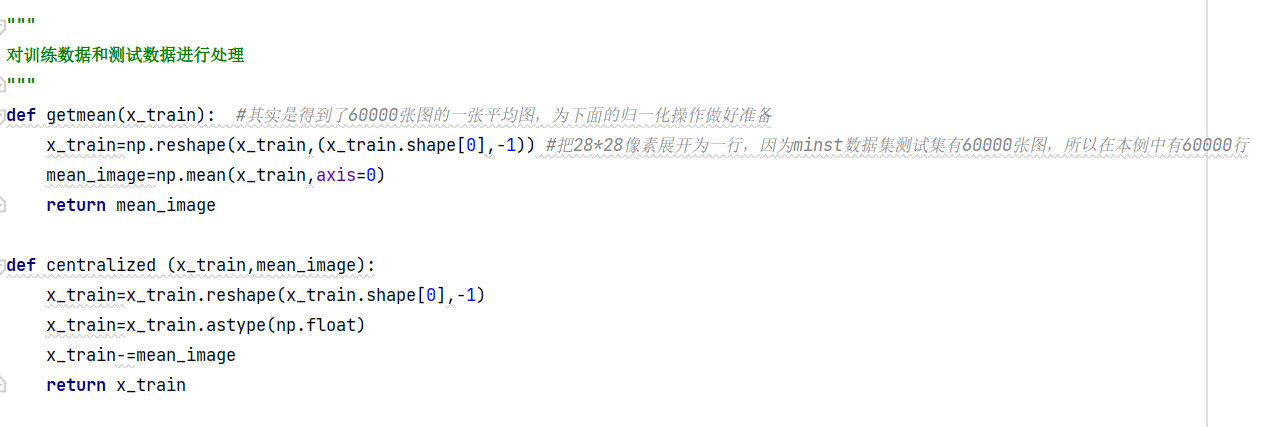
**3.1数据预处理**

**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**

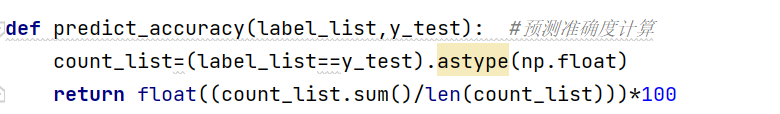
k-NN算法实质上利用距离，给定训练集数据以及标签（类别）。对于测试集中每一个点来言，分别求其到训练集每一个点的距离，选择前k个距离最小的点（k-NN中k的由来，其是一个超参数），考察其所属类别，这k个点中出现次数最多的类别即可认为是测试集中该点的所属类别。具体实现代码如下所示，核心的距离计算采用欧拉距离计算方法：



# 4.KNN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.SVM方法介绍

**5.1数据预处理**

**MINST数据集的读取和加载**



**5.2算法描述**

**SVM的基本思想：找到集合边缘上的若干数据（称为支持向量（Support Vector）），用这些点找出一个平面（称为决策面），使得支持向量到该平面的距离最大。**

**在实践过程中利用sklearn库中的svm包进行实现。SVM参数含义可以查阅**[SVM简介及sklearn参数 - Solong1989 - 博客园 (cnblogs.com)](https://www.cnblogs.com/solong1989/p/9620170.html)**获取。模型定义和训练代码如下所示。**

****

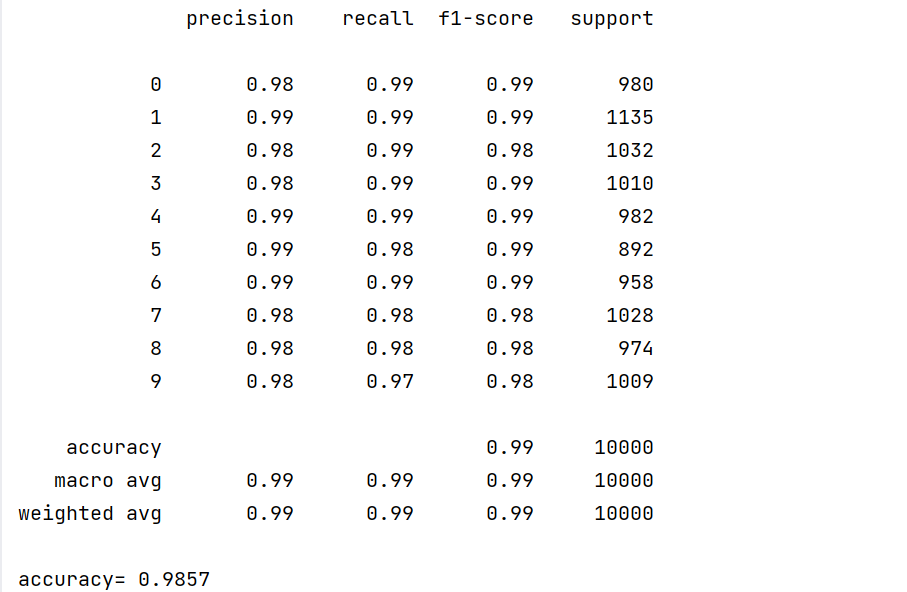
# 6.SVM实验结果分析

6.1评价指标

采用准确率（Accuracy），精确率（Precision），召回率（Recall）和F1分数（F1-Score）作为评价指标。关于分类指标的具体计算公式，指标含义的理解以及评估函数的使用可以参考：[【精选】准确度(accuracy)、精确率（precision)、召回率（recall）、F1值 谈谈我的看法\_精确度和召回率-CSDN博客](https://blog.csdn.net/lhxez6868/article/details/108150777)。

6.2定量评价结果

实验结果如下所示。

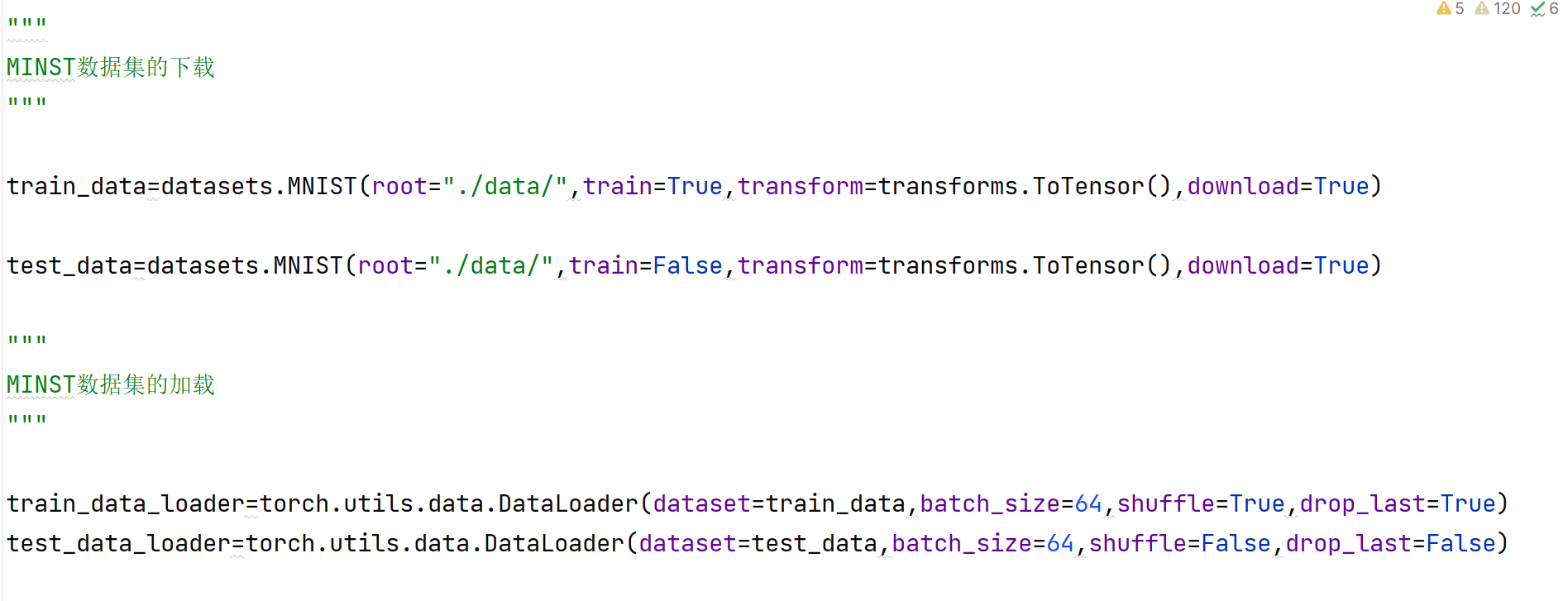


从上图实验结果来看，SVM对每个标签都有较高的分类精确率，召回率和F1分数，总体分类有较高的准确取得了较好的分类效果。

# 7.mlp-nn方法介绍

**7.1数据预处理**

**MINST数据集的读取和加载**



**7.2算法描述**

利用多层感知机并设计目标函数，对目标函数进行优化以达到较好的分类效果，网络定义如下。



采用了含有一个隐藏层的MLP作为我们要训练的模型，同时为了避免过拟合，采用丢弃法来缓解过拟合的情况。

在设计损失函数是，采用了两种方案：一种是交叉熵损失函数（CrossEntropyLoss），一种是均方误差损失函数（MSELoss），在实验过程中观察不同的损失函数对实验结果的影响。

# 8.mlp-nn实验结果分析

8.1评价指标

采用准确率作为评价指标

8.2定量结果分析

采用MSELoss的训练结果如下：



采用CrossEntropyLoss的训练结果如下：



比较上述实验结果可知：

不同的损失函数对训练结果影响很大，分类问题选用CrossEntropyLoss更为合适。在选用CrossEntropyLoss作为损失函数时，mlp-nn取得了较好的分类效果，分类时也没有出现严重的过拟合现象。

# 9.总结

在上述各实验中，采用了KNN，SVM，并采用两种不同的损失函数设计mlp-nn模型，除了采用MSELoss的mlp-nn模型，其他模型均取得了较好的分类效果，这也体现了损失函数在神经网络中的重要性。同时在本次实践中，将mlp-nn网络从cpu迁移到了pgu上进行训练，学会如何将网络放到gpu上训练也是本次的收获之一。