深圳大学考试答题纸

(以论文、报告等形式考核专用)  
二○ 23 ～二○ 24 学年度第 一 学期

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程编号 |  | | 课序号 | |  | | 课程名称 | |  | | 主讲教师 |  | | 评分 |  |
| 学 号 | 2020271040 | | | 姓名 | | 王伟钊 | | 专业年级 | | 2020级信息与计算科学01 | | | | | |
|  | | | | | | | | | | | | | | | |
| 教师评语： | | | | | | | | | | | | | | | |
| 题目： | | 基于机器学习算法的手写数字识别 | | | | | | | | | | |  | | |

摘要

关键词：手写数字识别、CNN神经网络、KNN算法

1. 背景介绍

手写数字识别是人工智能发展中面临的一个重要的问题，

1. 方法描述
2. 数据集MNIST

MNIST数据集来自美国国家标准与技术研究所。训练集由来自250个不同人手写的数组构成，其中50%是高中学生，50%来自人口普查局的工作人员。测试集也是同意比例的手写数字数据，但保证了测试集和训练集的作者集不相交。

本次实验的MNIST训练集有42000张图片，测试集有25张图片，每张图片的大小是28\*28的0-9的手写数字图片组成，每个图片是的像素点是[0,255]区间的数，越靠近255像素点的灰度越高。本次实验中的MNIST数据集的存储方式为，将第二行开始的数据点都接到上一行的末尾，使其变成一个长度为28\*28=784的向量，作为一个图像。

以下为随机抽取其中3张使用plt打印出来的图片：

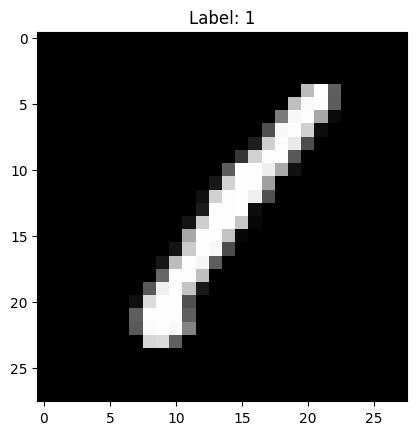
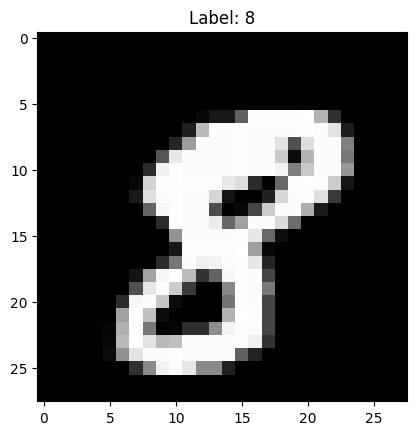
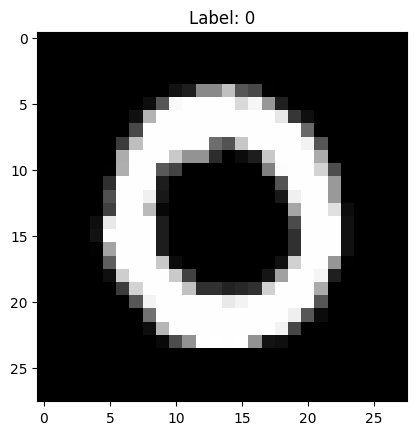


图1 MNIST数据集中的数据抽样

图2为这个数据集经过验证集和训练集按照比例4：1的大小划分后，数字0-9的出现频次的分布，可以发现，每个数字出现的频次相对平均，这样的数据使得样本方差较小，训练起来不易出现因样本数量不足而导致的模型过拟合或欠拟合，即对数据量多的数字判断准确而对数据量少的判断不准。

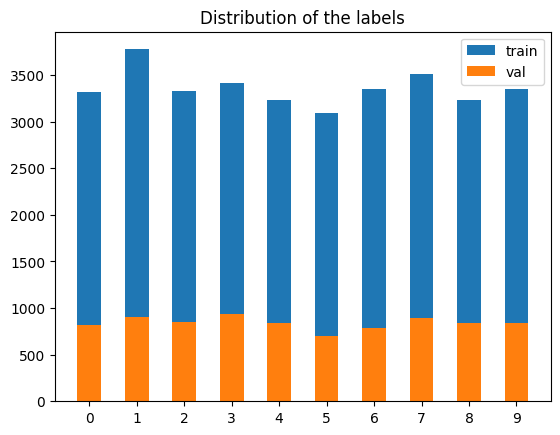


图2 数字0-9经验证集划分后的分布

1. 一般处理方法

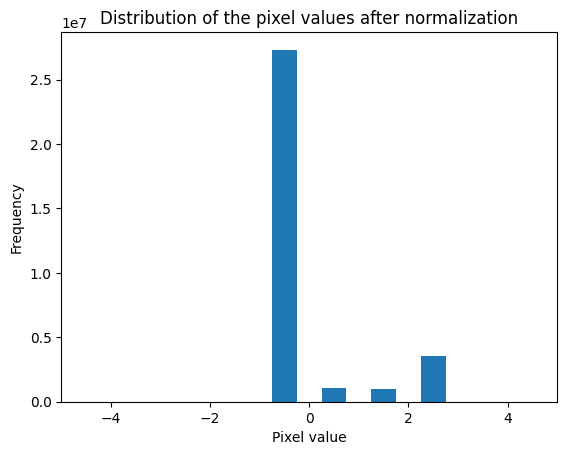
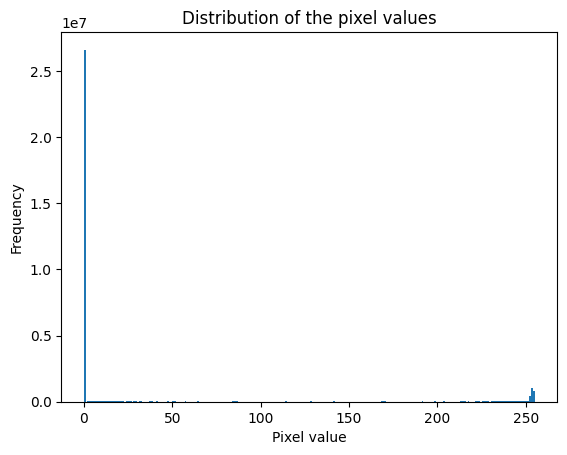
将训练数据分成两部分，一部分为训练集，另一部分为验证集，本次实验训练集与验证集的比例为8:2。使用sklearn中的train\_test\_split函数。

1. k折交叉验证

k折交叉验证将所有数据分成k份大小相等的子集，对于模型训练的过程中，模型将被训练k次，每次使用不同的折作为验证集，其余的k-1折作为训练集，每次训练后，模型都会在验证集上进行评估，记录性能指标（如准确率，召回率等），最终的性能评估取k次评估的平均值。k折交叉验证可以有效降低数据突变值导致的模型误差，通过不同的训练集进行训练也可以减少模型过拟合的风险。本次实验使用这种验证方法，将k值设定为5，则每份数据子集的大小为8400。每一次训练使用8400张作为验证集，另外33600张作为训练集，完成一次训练后，使用另外8400作为验证集，如此反复5次作为一轮训练。

1. 归一化

注意到数据集的分布，它由784个像素点组成一个图片，每一个像素点的取值范围为[0,255]，像素点有值与空白之间的跨度相当的大，若通过梯度下降的方法进行模型训练，模型需要更多的迭代次数才能得到最优的结果。若在神经网络中进行训练，归一化可以减少输入神经元的方差，加快网络的收敛。归一化的处理也能提高KNN算法的训练效果，减少了每一个像素点中间的尺度差异，提高KNN算法的速度。



归一化的方法有两种：

a. 我们注意到，像素点的取值范围为[0,255]，对每一个像素点都除以255，这样可以将数据限制在[0,1]的范围内，这种方法不会改变数据结构上的差异，只会改变尺度上的差异。

b. 另一种方法为计算某一个像素点在所有样本中出现的均值和标准差，用每一个点减去均值后除以标准差得到一个新的归一化后的数据，这样可以使样本的像素点的均值和标准差固定在0和1.但在计算的过程中，我们能发现，对于每一张图片的边边角角可能全部都是0，这使得在计算标准差时可能出现为0的结果，归一化后会出现除0的现象，我们需要在计算后将数据中的nan转换为0才可进行接下来的运算。

1. PCA降维数据

一张图像的像素点有784个，相当于图片的输入的维度为784，使用PCA的方法将它降为低维，使用低维数据作为神经网络的输入，这样可以保证低维向量含有图片的主要特征，同时可以快速提高模型的训练速度。

1. 前馈神经网络

前馈神经网络的含义是，输入m个元素，经过若干个全连接层，得到n个输出。全连接层是一个权重矩阵，其中的每一个节点都是由上一层的输入乘上对应点的权重相加，得到了输出。下图简单描述了全链接的工作原理。

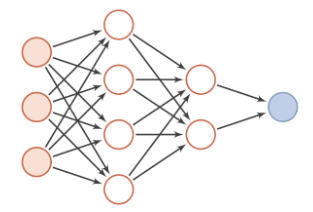


图3 全连接神经网络的工作原理

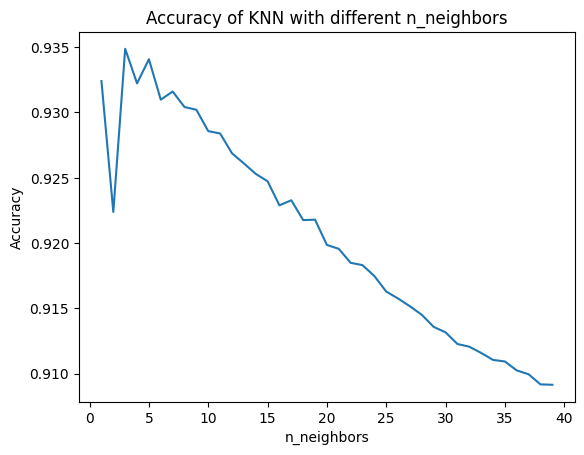
本次实验，我选用了一个最简单的全连接神经网络，一个只有一层全连接层，没有激活函数，输入为784，输出为10的全连接层。

1. 卷积神经网络
2. k近邻算法

k近邻算法（KNN）是有监督学习中的分类算法，其思想主要是对输入的点进行在空间上的对应，例如输入一个向量，其维度为784，则他对应了空间上的一个点，784个值为它784个方向的坐标，通过计算他与其他点的距离进行排序，最靠近这个点的k个点中，出现类别最多的点的类别作为预测输出。点距离的计算默认选择欧氏距离，其计算公式为：



k值的选择：k值是k近邻算法中非常重要的超参数，如何选择k值决定了算法的最终性能，k值过小会导致模型对其他数据的参考性不足，k值过大会导致结果过于以来其他数据，增加计算量同时会导致模型过拟合。下图为选择不同的k值影响的模型的准确率的变化，经过对比本次实验选择k值为5.

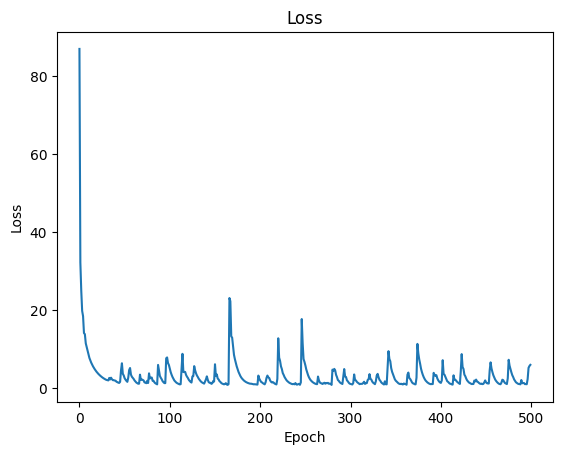
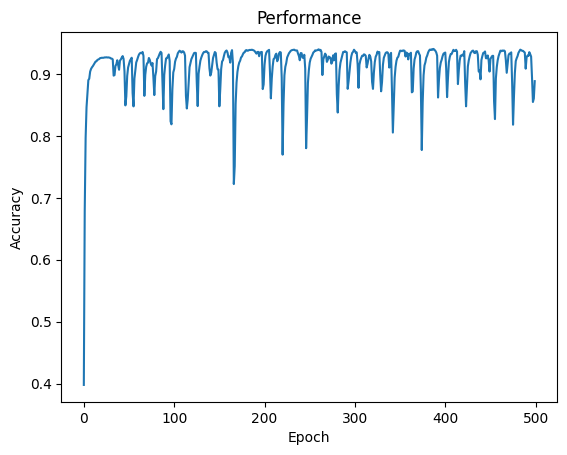


1. 模型评判标准

以上三种算法前馈神经网络、卷积神经网络、k近邻算法都是使用k折交叉验证的验证集的平均准确率作为模型的学习效果评判标准。

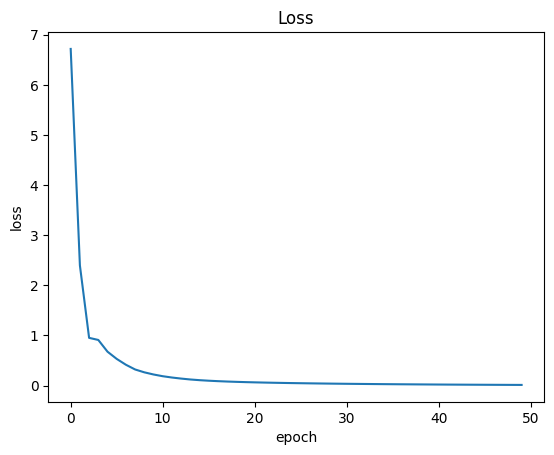
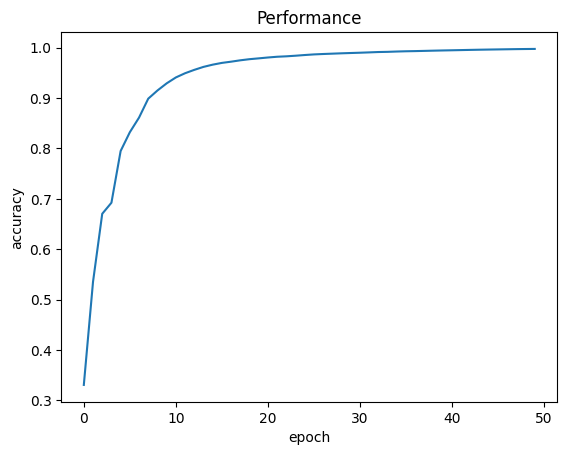
1. 实验内容和实验结果

前馈神经网络的训练效果

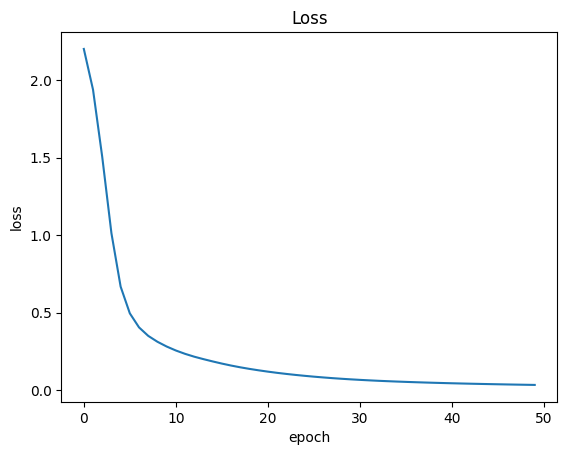
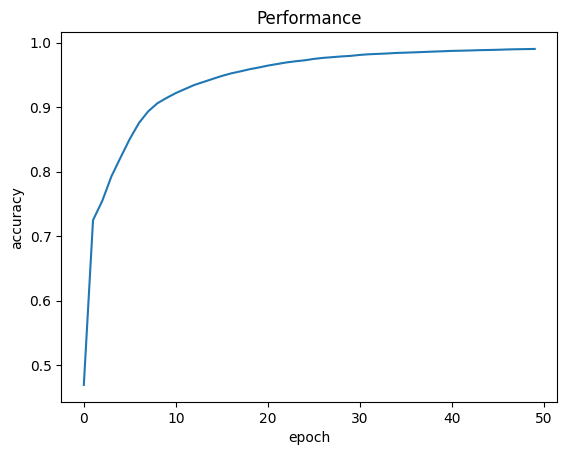


卷积神经网络的训练效果

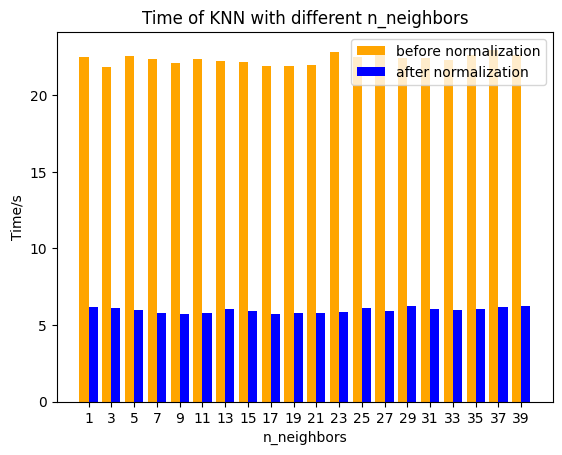
归一化前



归一化后



归一化对KNN算法的速度提升效果显著



1. 分析与结论