作业 1-1：人工智能著名数据集

图像 MNIST 数据集（手写数字集）学习笔记

一、MNIST 数据集内容

训练集：60,000 张 28×28 像素的灰度图像，每张图像对应一个 0-9 的标签。

测试集：10,000 张 28×28 像素的灰度图像，用于模型验证和测试。

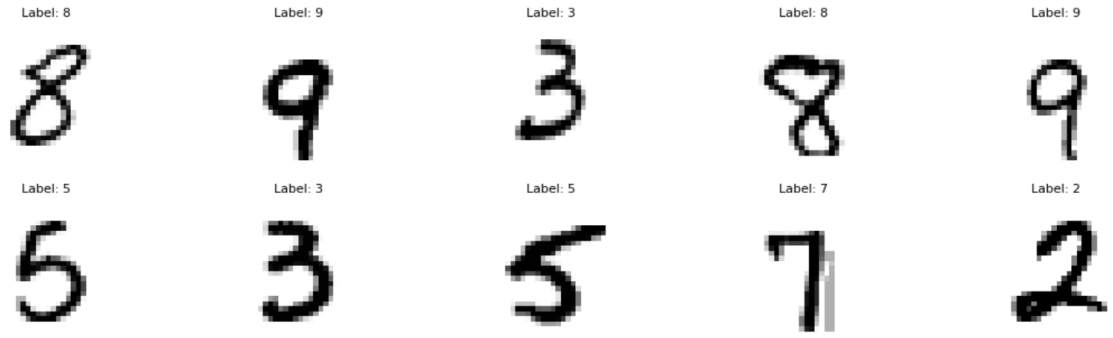
MNIST 又称“手写数字集”。简而言之，就是不同人手写得出的大量数字图，覆盖数字 0 至 9。

图像格式：灰度图像，像素值范围为 0 到 255。

标签格式：每个图像对应一个数字标签（0 到 9）。

灰度图像像素值范围是 0 到 255，相比于彩色图像，数据量减少了。对于“手写数字识别”而言，我们无需关注图像的颜色，只需要识别数字形状即可。因此，可以简化问题、减少计算复杂度。

随机部分内容展示：



下载方式：

https://github.com/geektutu/tensorflow-tutorial-samples/tree/master/mnist/data_set

或者使用 Scikit-learn 加载简化版 MNIST 数据集：

```
# 使用 Scikit-learn 加载 MNIST 简化版（适合快速实验）
from sklearn.datasets import fetch_openml
mnist = fetch_openml('mnist_784', version=1, as_frame=False)
data = mnist.data                    # 图像数据（每张图拉平为 784 维向量）
target = mnist.target.astype(int)    # 标签需转为整数，便于后续计算
```

二、MNIST 数据集用途

1）初学者入门：MNIST 数据集因其规模小且无复杂噪声，非常适合初学者入门深度学习和机器学习。它可以帮助新手快速理解图像识别的基本

概念和流程，例如数据预处理、模型训练、评估和优化。

2）算法对比和优化：MNIST 数据集常用于训练和测试各种图像处理系统，能够对比不同系统的正确率，以选择最优的算法。例如，可以通过对比感知机和 SVM（支持向量机）等算法的性能，了解它们在手写数字识别任务中的优缺点。（将在第三部分进行展示）

三、处理 MNIST 数据集的经典算法

常见的对 MNIST 数据集分类学习的算法有感知机、SVM（支持向量机）、决策树、最大熵模型和贝叶斯朴素法等等。这些算法在 sklearn 库里面都能找到。**接下来，以第一周学的“感知机”作例子，这里由于有从 0 到 9 一共十个例子，所以是“多层感知机”。**

1）回顾上课讲的简单感知机（二分类）基本原理

感知机是一种线性分类器，是一种“模拟人类决策过程”的最简单的人工神经元模型，目标是通过一个线性函数来区分两类数据。

通俗来说，现在有一些点分布在二维平面上，其中一些属于类别 A，另一些属于类别 B。感知机的目标是找一条直线，或者称作决策超平面，把这两类点划分开。核心思想是用权重（w_1、w_2）和阈值（y）把多个输入信号（x_1、x_2）综合成一个“是/否”（比如说y大于或小于 0，分别对应是和否）的决策。

$$y = w_1x_1 + w_2x_2 + b$$

举个简单的例子：“判断这个苹果是否属于优质苹果？”

- 输入：颜色红吗？表面光滑吗？（是=1，否=0）
- 权重：颜色比表面光滑更重要（权重分别为 0.7 和 0.3）
- 偏置：b 可以理解为“门槛的高低”，设为-0.5；y大于 0 时为优质苹果

若苹果颜色红（1），表面不光滑（0），总分$0.7 \times 1 + 0.3 \times 0 - 0.5 = 0.2 > 0$.

这个苹果分类为优质苹果。

2）从“单线思维”到“多层决策”：推广到 10 分类

将 10 分类问题转化为 10 个二分类问题。预测时，选择 10 个感知机中“可能性评分最高”的类别作为最终结果。

3）梯度下降法（SGD）的应用

通俗来说，假设你已经画了一条分界线，但有些小球可能没有被正确分开（比如，有些红色小球跑到了蓝色小球的那一边）。感知机的工作则是：

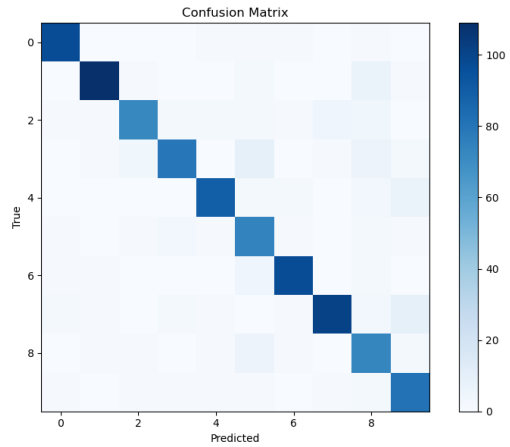
- 检查小球有没有被误分类
- 计算这些误分类小球到分界线的距离，然后调整分界线，使得这些误分类小球到分界线的距离变得更远。

感知机的目标是最小化误分类点到决策超平面的距离，使用梯度下降算法（SGD）对目标函数进行优化。每次随机选取一个误分类样本，计算损失函数的梯度，并更新权重w和偏置b。损失函数定义为：

$$L(w,b) = \sum_{x_i \in M} -y_i(w^T x_i + b)$$

M是误分类点的集合。

4）应用“多层感知机”对 MNIST 数据集进行手写数据识别结果



混淆矩阵：

对角线上的值表示正确分类的样本数，非对角线则为误分类的。比如这个例子中，我们可以看出（5,3）的颜色较深，说明误分类 5 和 3 的样本数较多。手写时，5 和 3 的顶部也确实容易分不清。接下来，我们可以适当增多 5 和 3 的训练样本，强化识别力。可见，MNIST 数据集的用途之一包括训练和测试各种图像处理系统。

接下来，我们同时运用多层感知机和 SVM 支持向量机算法，得出结果如下：

```
感知机准确率：88.10%，训练时间：1.46s
SVM 准确率：91.60%，训练时间：1.55s
```

对比两种算法的准确率和训练时间，可知 SVM 在处理复杂问题时具有更好的分类能力；感知机在计算效率上具有一定优势，但可能牺牲一定的分类精度。

从原理上来看，对于一个二分类问题而言，如果存在红色蓝色两种小球，感知机需要找到一条直线将它们分开，而 SVM 不仅要将它们区分开，还要确保分界线两边的点尽可能远。

MNIST 数据集可以帮助我们了解不同算法的优缺点，并在实际应用中选择更合适的算法。感知机适用于简单的线性可分问题，尤其是在计算资源有限的情况下。SVM 适用于复杂的非线性问题，尤其是在需要高分类精度的场景中。

具体代码实现见附件。