**南京邮电大学**

**自动化学院、人工智能学院**

**实 验 报 告**

**课程名称：** 模式识别（双语）A

**实验名称：** 线性分类器实验

**指导教师：** 谢九成

**所在专业：**  人工智能

**学生姓名：** 单家俊

**班级学号：** B210416 B21080526

**2023 /2024 学年第 一 学期**

**实验名称：线性分类器实验**

1. **实验目的**
2. 掌握SVM原理
3. 研究训练样本数量和分类器识别精度的关系
4. **实验任务**

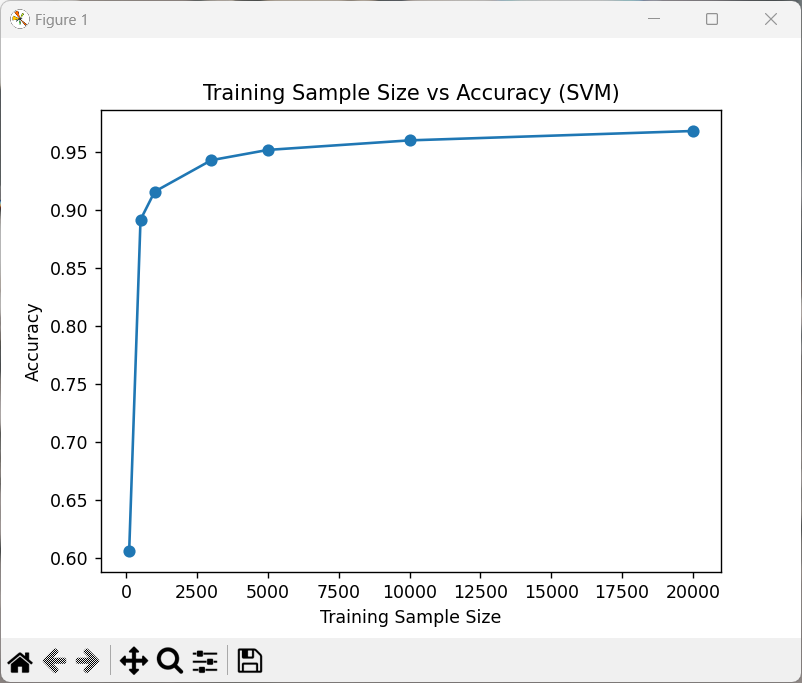
以手写数字识别为任务，基于 MNISTS 数据集，看看SVM分类器在数字识别任务上的性能如何。研究下训练样本数量和分类器识别精度的关系，画出变化曲线图。编程语言不限。

1. **实验内容**

import numpy as np  
import matplotlib.pyplot as plt  
from sklearn.model\_selection import train\_test\_split  
from sklearn.svm import SVC  
from sklearn import datasets  
  
*# 加载MNIST数据集*mnist = datasets.fetch\_openml('mnist\_784',parser='auto',data\_home='local\_directory')  
X = np.array(mnist.data.astype('float32'))  
y = np.array(mnist.target.astype('int'))  
  
*# 将像素值缩放到 [0, 1] 范围*X /= 255.0  
  
*# 划分数据集为训练集和测试集*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  
  
*# 初始化SVM分类器*classifier = SVC()  
  
*# 训练分类器*classifier.fit(X\_train, y\_train)  
  
*# 在测试集上进行预测*accuracy = classifier.score(X\_test, y\_test)  
print(f"Accuracy: {accuracy}")  
  
*# 绘制训练样本数量与分类器识别精度的关系曲线图*train\_sizes = [100, 500, 1000, 3000, 5000, 10000, 20000]  
accuracies = []  
  
for size in train\_sizes:  
 X\_subset, \_, y\_subset, \_ = train\_test\_split(X\_train, y\_train, train\_size=size, random\_state=42)  
 classifier.fit(X\_subset, y\_subset)  
 accuracy\_subset = classifier.score(X\_test, y\_test)  
 accuracies.append(accuracy\_subset)  
  
*# 绘制曲线图*plt.plot(train\_sizes, accuracies, marker='o')  
plt.title('Training Sample Size vs Accuracy (SVM)')  
plt.xlabel('Training Sample Size')  
plt.ylabel('Accuracy')  
plt.show()

1. **实验过程描述**

**实验结果分析：**

****

通过准确度评估，我们可以了解SVM在MNIST数据集上的整体性能。而通过关系图，我们能观察到在不同规模的训练数据集上，分类器的性能如何变化，从而更好地理解训练样本数量对模型性能的影响。当训练样本较少时，增加样本的数量会显著提升分类器的性能，但是当训练样本达到一定数量之后，随着样本的增加，分类器性能提升幅度很小。

**实验过程遇到的问题：**

在实验过程中，可能会遇到过拟合和欠拟合的问题。随着训练样本数量的变化，可能观察到模型在训练集和测试集上的性能差异。解决这些问题需要对模型进行调优，确保在不同规模的数据集上都能表现良好。同时，处理大规模数据集可能涉及到计算资源的限制，需要权衡计算效率和模型性能之间的关系。

**体会：**

通过这个实验，我深入了解了机器学习模型在不同规模数据集上的表现，并认识到训练样本数量对模型性能的重要性。了解模型在不同参数和数据规模下的表现，是调优机器学习模型的重要一步。实验还强调了可视化在分析和沟通结果方面的重要性，通过绘制关系图，我更直观地理解了实验结果。这个经验对我在将来的机器学习项目中选择合适的数据规模和优化模型性能提供了有益的指导。