**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：** 基于贝叶斯理论实验

**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：基于贝叶斯理论实验**

1. **实验目的**

本实验的目的是通过对MNIST手写数字数据集的贝叶斯分类器的训练，观察在不同规模的训练样本数量下，模型的性能如何变化。通过逐步增加训练样本的方式，探讨训练集规模对分类器准确性的影响，以及模型是否在逐渐学习到更多的样本时能够提高在测试集上的泛化性能。

1. **实验任务**

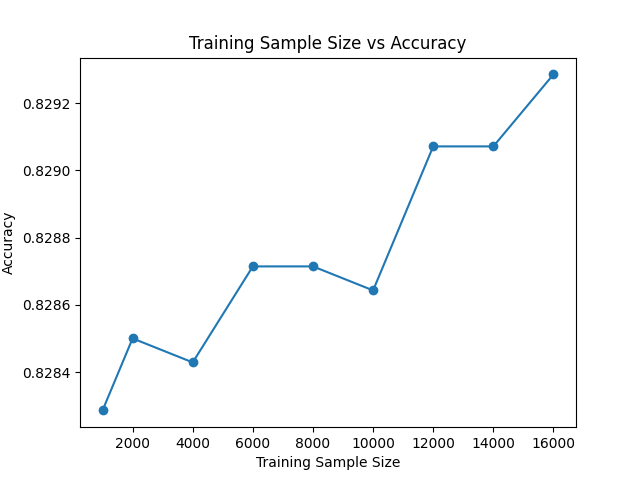
以手写数字识别为任务，看看基于 MNISTS 数据集，贝叶斯分类器在数字识别任务上的性能如何。研究下训练样本数量和分类器识别精度的关系，画出变化曲线图。编程语言不限。

1. **实验内容**

import numpy as np  
import matplotlib.pyplot as plt  
from sklearn.datasets import fetch\_openml  
from sklearn.model\_selection import train\_test\_split  
from sklearn.naive\_bayes import MultinomialNB  
from sklearn.metrics import accuracy\_score  
 *# 加载数据集*mnist = fetch\_openml('mnist\_784', parser='auto', data\_home='local\_directory')  
X, y = mnist.data.astype('float32') / 255.0, mnist.target.astype('int')  
  
*# 拆分数据集*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  
  
*# 训练贝叶斯分类器*clf = MultinomialNB()  
clf.fit(X\_train, y\_train)  
  
*# 步骤6：调整训练样本数量并评估性能*train\_sizes = [1000, 2000, 4000, 6000, 8000, 10000, 12000, 14000, 16000]  
accuracies = []  
  
for size in train\_sizes:  
 clf.partial\_fit(X\_train[:size], y\_train[:size], classes=np.unique(y))  
  
 y\_pred = clf.predict(X\_test)  
 accuracy = accuracy\_score(y\_test, y\_pred)  
 accuracies.append(accuracy)  
  
*# 绘制曲线图*plt.plot(train\_sizes, accuracies, marker='o')  
plt.title('Training Sample Size vs Accuracy')  
plt.xlabel('Training Sample Size')  
plt.ylabel('Accuracy')  
plt.show()

1. **实验过程描述**

**实验结果分析：**



通过实验中的迭代，绘制了训练样本数量与准确度之间的曲线图。从曲线图中可以观察到，随着训练样本数量的增加，模型在测试集上的准确性逐渐提高，，但他们的关系并不是单调的。

**实验过程遇到的问题：**

部分训练的正确调用：

在实现逐步增量式训练时，需要确保partial\_fit函数被正确调用，且参数传递正确。可能出现的问题包括参数类型不匹配、类别信息不正确等。确保正确调用部分训练是保证模型增量学习的重要一步。

模型性能评估：

在迭代中评估模型性能时，需要使用测试集对模型进行评估，但评估方法的选择也是一个重要的问题。在这里采用了准确度（accuracy\_score）作为评估指标，确保模型的整体性能得以衡量。

**体会：**

透过实验结果，深刻认识到训练样本数量对模型性能的重要性。随着样本数量的增加，模型更能学到数据的分布和特征，提升了对新样本的泛化能力。此外，选择合适的模型评估指标对于了解模型性能至关重要。准确度是一个直观而有效的指标，但在一些情况下可能需要考虑其他指标，如精确度、召回率等，以更全面地评估模型。

通过解决实验中遇到的问题和仔细观察实验结果，对机器学习模型训练的关键环节和影响因素有了更深入的理解。这对于在实际应用中选择合适的训练策略和评估方法具有指导意义。