实验九：Bagging 集成学习

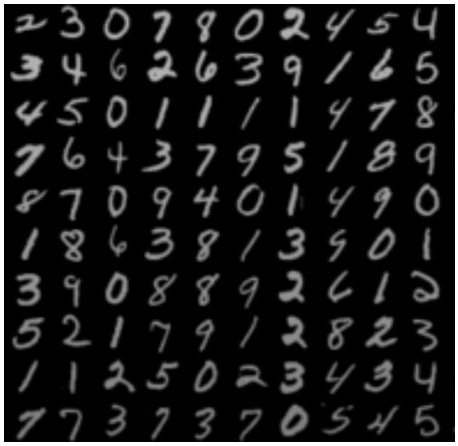
姓名：　　学号：

- 实验目的

参考随机森林，以决策树为基学习器，构建 bagging 集成器用于多分类任务。

- 实验要求

编程实现随机森林模型，对手写数字识别数据集进行分类。基模型采用决策树模型，划分属性指标采用信息熵指标，随机选取属性子集数目为 50。将决策树数量T依次设置为$1,2,\ldots,20$，计算随机森林在测试集上的精度，并绘制随机森林精度随基模型数量增加的变化曲线。



- 实验环境
- Python，numpy，matplotlib，sklearn
- 实验代码

```
import numpy as np
from sklearn import tree
from scipy import stats
from matplotlib import pyplot as plt


# 加载数据
training_data = np.loadtxt('experiment_09_training_set.csv', delimiter=',', skiprows=1)
testing_data = np.loadtxt('experiment_09_testing_set.csv', delimiter=',', skiprows=1)
```

```
# 数据分割及预处理
y_train = training_data[:, 0]
x_train = training_data[:, 1:] / 255.0
y_train = y_train.reshape(-1, 1)
y_test = testing_data[:, 0]
x_test = testing_data[:, 1:] / 255.0
y_test = y_test.reshape(-1, 1)

# 模型训练
model_all = []
for i in range(1, 21):
    choice_array = np.random.choice(x_train.shape[0], x_train.shape[0], replace=True)
    x_train_choice = x_train[choice_array, :]
    y_train_choice = y_train[choice_array]
    model = tree.DecisionTreeClassifier(random_state=1, criterion='entropy', max_features=50)
    model.fit(x_train_choice, y_train_choice)
    model_all.append(model)

# 计算精度
y_pred_all = []
```

```
accuracy_all = []
for i in range(len(model_all)):
    y_pred = model_all[i].predict(x_test)
    y_pred = y_pred.reshape(-1, 1)
    if i == 0:
        y_pred_all = y_pred
    else:
        y_pred_all = np.hstack((y_pred_all, y_pred))
    y_pred_f = stats.mode(y_pred_all, axis=1).mode
    y_pred_f = y_pred_f.reshape(-1, 1)
    accuracy = np.sum((y_pred_f == y_test)) / y_test.size
    accuracy_all.append(accuracy)
    print(f"T:{i+1} accuracy:", accuracy)

# 可视化
plt.rcParams['font.sans-serif'] = ['Microsoft YaHei']
plt.plot(range(1, 21), accuracy_all, marker='*', linewidth=0.8)
plt.xlabel("T")
plt.ylabel("精度")
plt.title("Bagging 集成学习决策树数量 T 与精度曲线")
plt.show()
```

- 结果分析

测试集上精度

T	1	2	3	4	5	6	7	8	9	10
精度	0.8178	0.8207	0.8788	0.9021	0.9149	0.9222	0.9260	0.9313	0.9354	0.9383
T	11	12	13	14	15	16	17	18	19	20
精度	0.9393	0.9418	0.9442	0.9445	0.9460	0.9470	0.9489	0.9493	0.9508	0.9514

精度随T增加的变化曲线