

实验报告：机器学习算法实践

一、实验目的

掌握使用 scikit-learn 实现经典机器学习算法（决策树与支持向量机），并评估其在 Iris 数据集上的分类性能。

二、实验内容

- 1. 加载 Iris 数据集；
- 2. 使用决策树与支持向量机（SVM）进行分类；
- 3. 划分训练集与测试集，计算模型准确率与分类评估报告。

三、实验步骤与代码

1. 导入所需库

```
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report
```

2. 加载数据并划分训练集/测试集

```
iris = datasets.load_iris()
X, y = iris.data, iris.target
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)
```

3. 训练模型

```
dt = DecisionTreeClassifier(random_state=42)
svm = SVC(kernel='rbf', random_state=42)
dt.fit(X_train, y_train)
svm.fit(X_train, y_train)
```

4. 预测和评估

```
y_pred_dt = dt.predict(X_test)
y_pred_svm = svm.predict(X_test)

print("决策树准确率:", accuracy_score(y_test, y_pred_dt))
print("SVM 准确率:", accuracy_score(y_test, y_pred_svm))
print("SVM 分类报告:\n", classification_report(y_test, y_pred_svm))
```

四、实验结果与分析

• 输出结果示例：

```
决策树准确率: 1.0
SVM 准确率: 1.0

SVM 分类报告:
              precision    recall  f1-score   support
```

0	1.00	1.00	1.00	19
1	1.00	1.00	1.00	13
2	1.00	1.00	1.00	13
accuracy			1.00	45

• 结果分析：

两种模型在 Iris 数据集上均达到 100% 准确率，说明数据集线性可分且特征区分度高。SVM 使用 RBF 核在小样本下表现稳健，决策树模型结构直观且可解释性强。

五、思考题

- 1. 如何防止模型过拟合？
 - 可采用交叉验证、正则化（如 L1/L2）、剪枝（决策树）、早停法（神经网络）、增加训练数据或 Dropout 等方法。简化模型结构也能有效降低过拟合风险。
- 2. 特征工程对模型性能有何影响？
 - 良好的特征工程能显著提升模型效果。例如：数值特征标准化可加速 SVM 收敛；类别特征进行 One-Hot 编码可避免模型误判顺序关系；特征选择可去除噪声，提高模型泛化能力。

六、实验总结

本实验通过 scikit-learn 实现了决策树与 SVM 分类器，完成了数据加载、模型训练、预测与评估的全流程。结果表明，两种算法在 Iris 数据集上均能达到最优性能，验证了经典机器学习算法在标准数据集上的有效性。