

人工智能课程实验一

**姓名： 叶宏庭**

**学号： 71118415**

**邮箱： 213182964@seu.edu.cn**

2021 年 1 月 4日

**目录**

[一、学习任务 3](#_Toc61008687)

[1 数据集一 3](#_Toc61008688)

[2 数据集二 3](#_Toc61008689)

[3 数据集三 3](#_Toc61008690)

[4 数据集四 3](#_Toc61008691)

[二、性能指标度量 3](#_Toc61008692)

[1 错误率与精度 3](#_Toc61008693)

[三、数据集信息 4](#_Toc61008694)

[四、数据集预处理 4](#_Toc61008695)

[1 数据集一 4](#_Toc61008696)

[2 数据集二 5](#_Toc61008697)

[3 数据集三 5](#_Toc61008698)

[4 数据集四 6](#_Toc61008699)

[五、对比算法 6](#_Toc61008700)

[1 对比算法 6](#_Toc61008701)

[2 参数设定 6](#_Toc61008702)

[2.1决策树算法 6](#_Toc61008703)

[2.2 MLP多层感知机网络算法 7](#_Toc61008704)

[2.3 SVM算法 7](#_Toc61008705)

[3 算法来源 7](#_Toc61008706)

[六、实验方法 7](#_Toc61008707)

[1 实验方法 7](#_Toc61008708)

[2 假设检验方法 7](#_Toc61008709)

[七、实验结果 8](#_Toc61008710)

[1 精度结果图表 8](#_Toc61008711)

[2 Friedman 检验 9](#_Toc61008712)

# 一、学习任务

## 1 数据集一

数据集一为车辆信息数据集，提供特征属性包括购买价格、维修费用、容量等，要求根据提供的六个特征属性，对车辆进行分类，是一个四分类任务(unacc、acc、good、vgood)。

## 2 数据集二

数据集二为sklearn自带的数据集iris（鸢尾花），提供特征属性包括花萼长度、花萼宽度、花瓣长度、花瓣宽度，特征值均为浮点数，这是一个三分类任务(Iris Setosa、Iris Versicolour、Iris Virginica)。

## 3 数据集三

数据集三为sklearn自带的数据集digits（手写数字识别），提供的数字图像为8\*8大小，因此共有64的特征属性，每个特征属性的取值均是0-16的Integer，这是一个十分类的分类任务（[0-9]十个类别的数据）。

## 4 数据集四

数据集四为sklearn自带的数据集Wine(红酒分类)，提供的特征属性包括红酒中的13种化学成分的含量，根据这些特征属性进行红酒的档次分类，这是一个三分类的任务。

# 二、性能指标度量

## 1 错误率与精度

错误率是分类错误的样本数占样本总数的比例，精度则是分类正确的样本数占样本总数的比例。对于样本集D，分类错误率定义为

精度定义则为

# 三、数据集信息

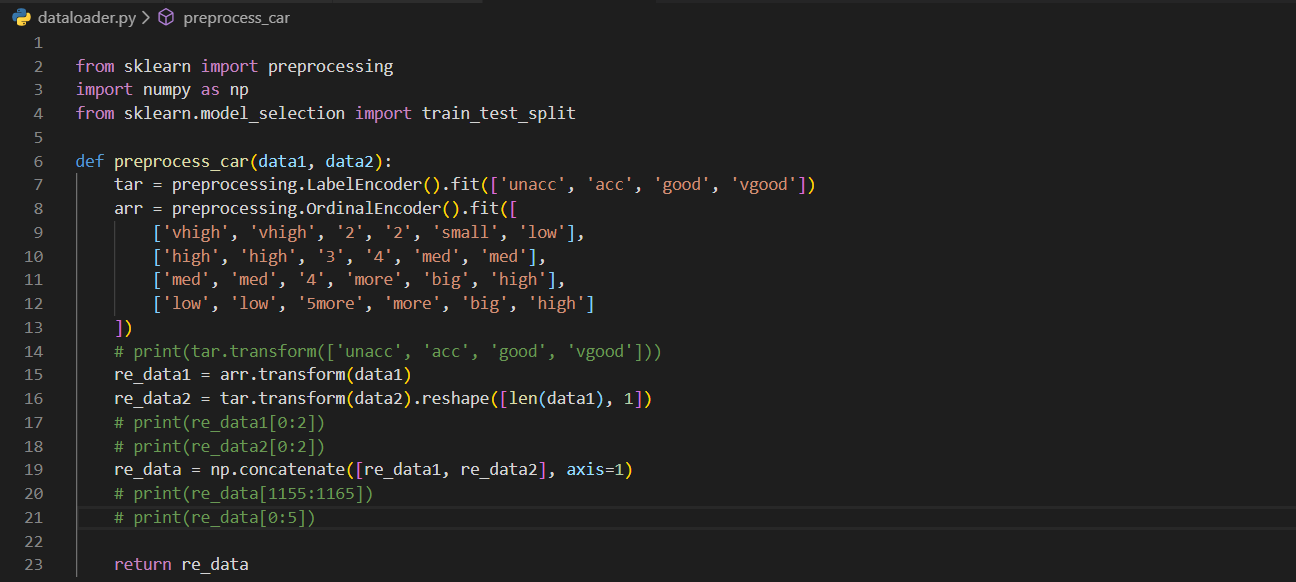
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集名称 | 样本数 | 属性数 | 类别数 | 各类别样本数 | | |
| Car Evaluation | 1728 | 6 | 4 | unacc | | 1210 |
| acc | | 384 |
| good | | 69 |
| vgood | | 65 |
| Iris | 150 | 4 | 3 | Setosa | | 50 |
| Versicolour | | 50 |
| Virginica | | 50 |
| Digits | 1797 | 64 | 10 | [0-9]各类别均≈180个样本 | | |
| Wine | 178 | 13 | 3 | class\_1 | 59 | |
| class\_2 | 71 | |
| class\_3 | 48 | |

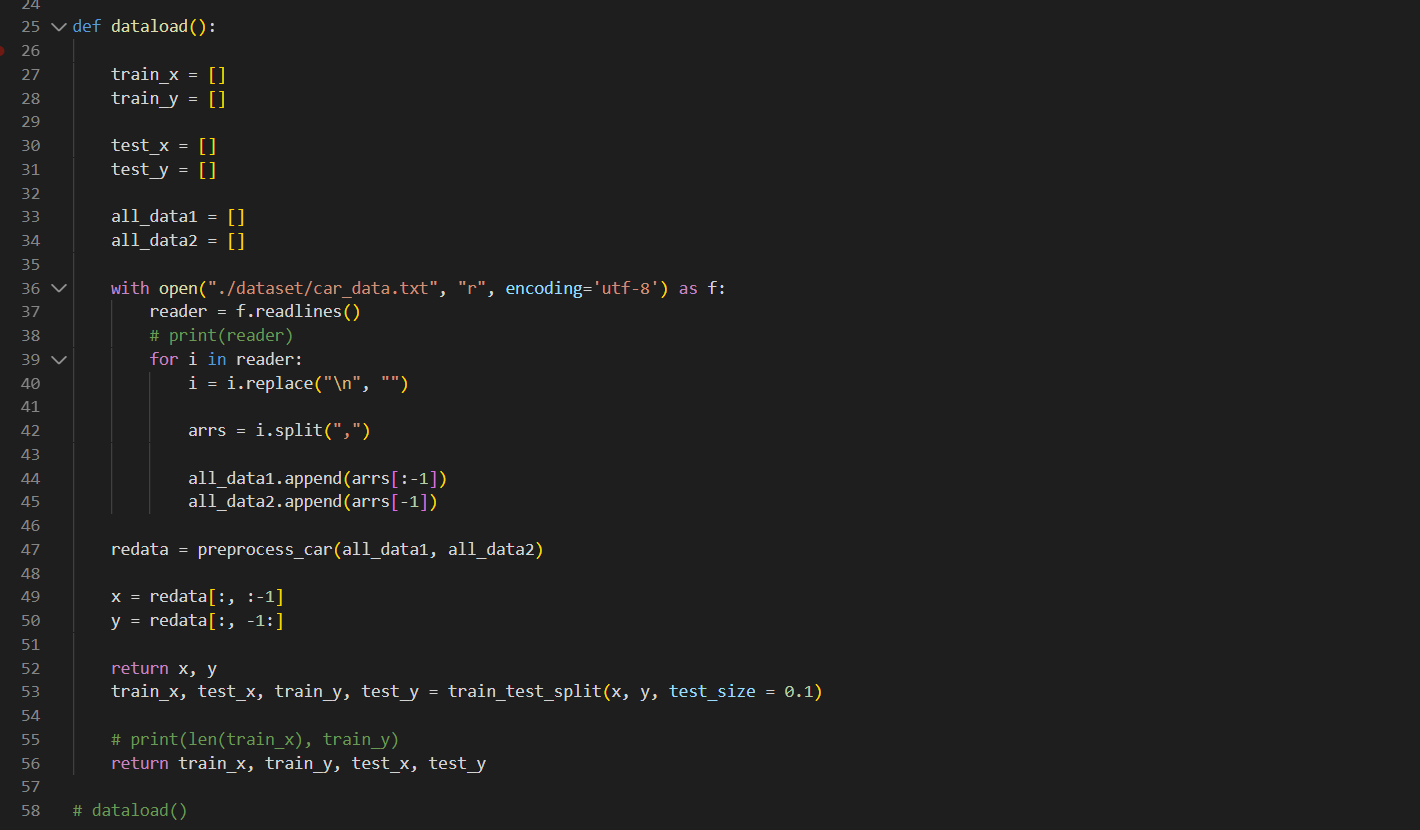
# 四、数据集预处理

## 1 数据集一

针对Car Evaluation数据集，采用python自行编码，实现dataloader，根据属性标签进行划分，采用train\_test\_split()进行随机划分，设置test\_size属性为0.1。

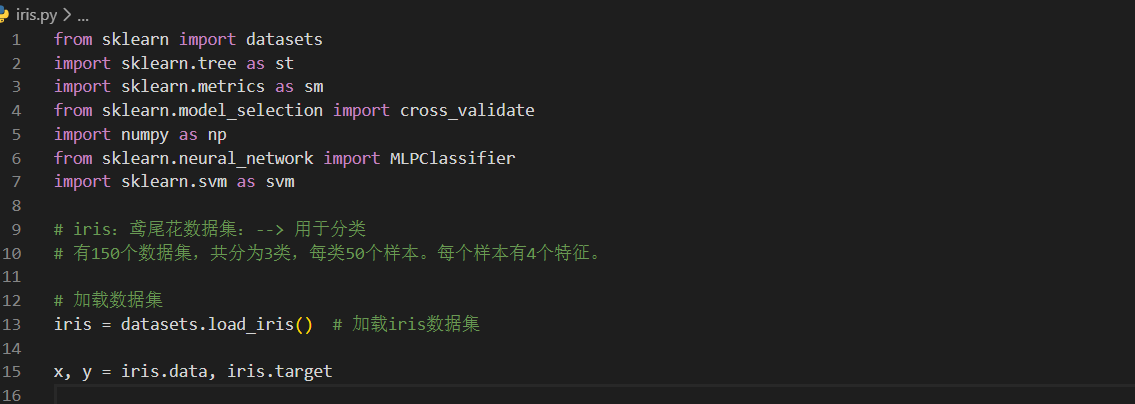
采用分类编码器，将文本类型的特征值转划为机器学习算法可处理的数字类型。





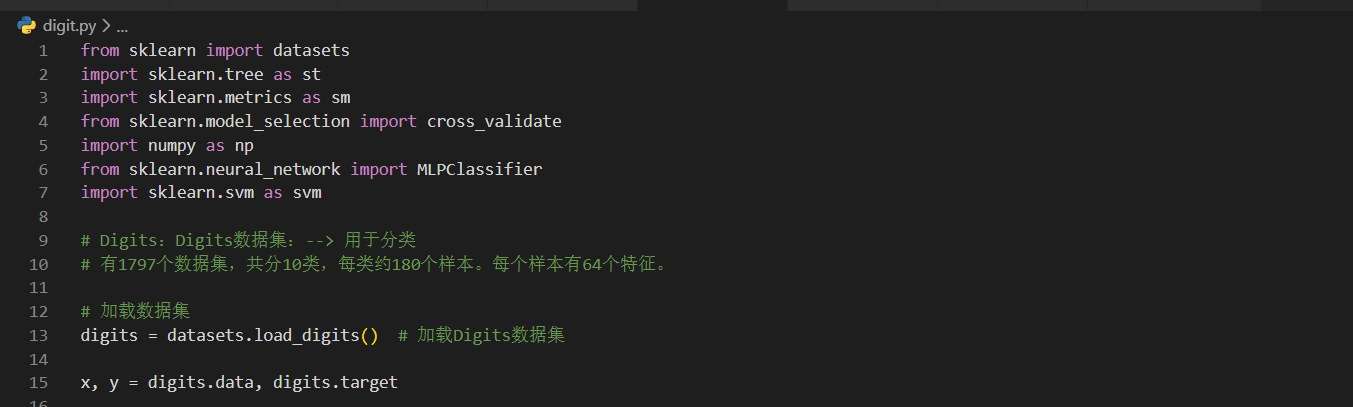
## 2 数据集二

该数据集为sklearn内置数据集，已经经过预处理操作，无缺损值等情况。下面给出数据集加载代码。



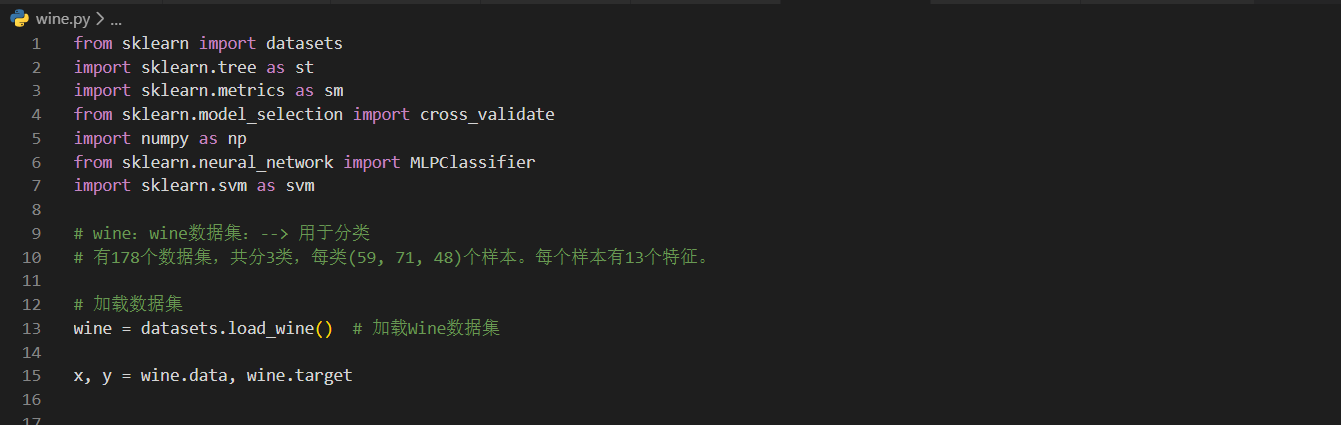
## 3 数据集三

该数据集为sklearn内置数据集，已经结果预处理操作，无缺损值等情况。下面给出数据集加载代码。



## 4 数据集四

该数据集为sklearn内置数据集，已经结果预处理操作，无缺损值等情况。下面给出数据集加载代码。



# 五、对比算法

## 1 对比算法

本实验采用三个算法进行实验对比，分别是决策树算法、MLP多层感知机网络、SVM算法。

## 2 参数设定

### 2.1决策树算法

DecisionTreeClassifier()参数设定如下：

|  |  |  |
| --- | --- | --- |
| 参数名 | 参数设定值 | 说明 |
| criterion | gini | 使用基尼系数作为评判标准 |
| splitter | best | 在所有特征中寻找最好切分点 |
| max\_features | None | 采用所有特征 |
| max\_depth | 7 | 最大深度为7层，经过测试4-10层深度，最终采取7层使得算法效果最好 |
| 其他参数 | default | 采用默认参数 |

### 2.2 MLP多层感知机网络算法

MLPClassifier()参数设定如下：

|  |  |  |
| --- | --- | --- |
| 参数名 | 参数设定值 | 说明 |
| hidden\_​​layer\_sizes | （35，20） | 第i个数表示第i层的神经元个数 |
| activation | logistic | 采用逻辑激活函数，即sigmod()函数 |
| solver | lbfgs | 准牛顿方法族的优化器 |
| alpha | 0.0001 | L2正则化参数 |
| 其他参数 | default | 采用默认参数 |

### 2.3 SVM算法

SVC（C-Support Vector Classification）参数设定如下：

|  |  |  |
| --- | --- | --- |
| 参数名 | 参数设定值 | 说明 |
| kernel | rbf | 算法采用核函数类型 |
| C | 600 | 错误项的惩罚系数 |
| 其他参数 | default | 采用默认参数 |

## 3 算法来源

本实验算法均采用sklearn机器学习包，采用python语言进行实验。

# 六、实验方法

## 1 实验方法

本实验采用交叉验证法进行，设置参数k = 10，即采取10折交叉验证法进行实验，最后取十次的平均准确率作为最终的性能指标。

## 2 假设检验方法

本实验采用Friedman test方法，对实验算法结果进行假设检验。

弗里德曼(Friedman test)检验即 “弗里德曼双向秩方差分析”，是多个(相关)样本齐一性的统计检验。 该方法是弗里德曼 (M.Friedman) 1973年提出的。弗里德曼检验前提要求：1.顺序级数据；2.三个或更多组；3.相关的小组；4.从搭配的数值中随机地抽取样本。

具体检验过程参照《机器学习》(周志华) P42 – P43。

# 七、实验结果

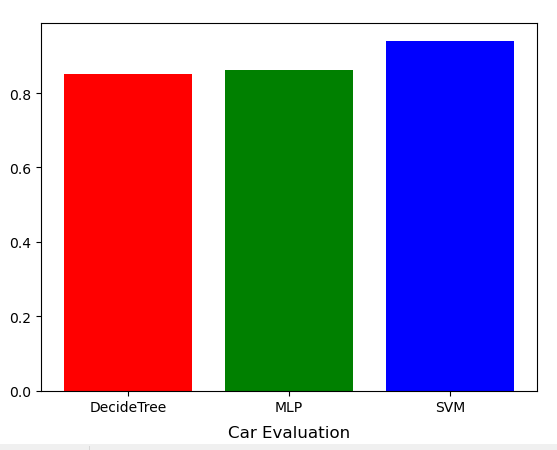
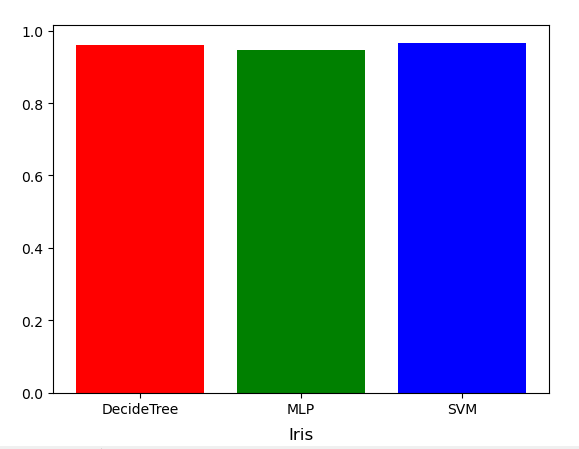
## 1 精度结果图表

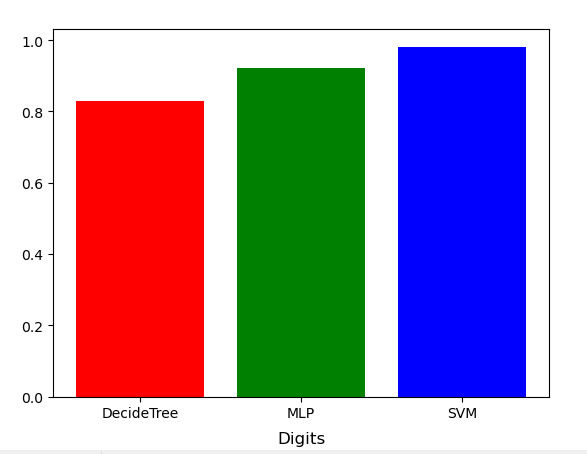
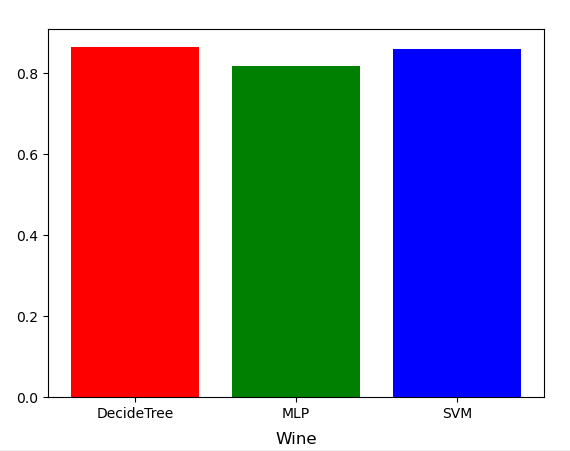
（因部分结果相接近，所以结果保留5位小数）

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集 | 决策树 | MLP多层感知机网络 | SVM算法 |
| Car Evaluation | 0.85196 | 0.86174 | 0.94102 |
| Iris | 0.96000 | 0.94667 | 0.96667 |
| Digits | 0.82916 | 0.92207 | 0.98163 |
| Wine | 0.86470 | 0.81601 | 0.86013 |
| 部分统计量 | | | |
| 均值 | 0.87646 | 0.88662 | 0.93736 |
| 方差 | 0.00249 | 0.00262 | 0.00219 |

从精度结果来看，SVM算法在多个数据集中都表现出了最好的分类精度结果，并且平均精度也取得了最高分数，其次是MLP多层感知机网络。

从精度结果的方差来看，SVM算法最为稳定，方差最小，MLP多层感知机网络的方差最大，结果最不稳定。

## 2 Friedman 检验

根据实验精度结果表，采用Friedman检验方法，构造出下面的检验表。

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集 | 决策树 | MLP多层感知机网络 | SVM算法 |
| Car Evaluation | 3 | 2 | 1 |
| Iris | 2 | 3 | 1 |
| Digits | 3 | 2 | 1 |
| Wine | 1 | 3 | 2 |
| 平均值 | 2.25 | 2.5 | 1.25 |

从表中可以看出，SVM的平均序值最小，MLP的平均序值最大，由平均序值，下面进行假设检验。

下面进行假设检验部分，假设 “所有的算法性能均相同”。

其中，表示数据集个数，表示算法个数，表示第个算法的平均序值。

在本实验中，经计算， = 2.333 。

因为， 服从自由度为和的分布。根据查表可得，，所以，因此可认为三个算法的性能大致相同。