上海理工大学光电信息与计算机工程学院

**《数据挖掘》实验报告**

****

**专　　业 智能科学与技术**

**姓 名　　高浩琦**

**学　 号 2035060413**

**年　　级 2020级**

**指导教师 孙占全**

**成 绩：**

**教师签字：**

# 实验一 数据预处理与可视化

一、实验目的和要求

1. 掌握数据预处理方法，如缺失值修复、数据标准化等。

2. 掌握数据挖掘中的分类聚类方法和分类性能的评价估计

3. 熟悉ROC和PR曲线评估结果的计算过程

4. 以鸢尾花数据集为例（'http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data），学会使用python做去除缺失数据记录或用均值补全缺失数据、做数据标准化、萼片和花瓣长宽属性10等分的直方图、萼片和花瓣长宽属性箱式图、萼片和花瓣长宽属性两两之间散点图

5. 基于鸢尾花数据集，利用决策数分类器进行分类，画出测试结果的ROC和PR曲线。

二、实验算法

Pd.read\_csv()

Data.describe()

Data.iloc()

Data.isnull()

Plt.hist()

Plt.boxplot()

Plt.scatter()

Plt.subplot()

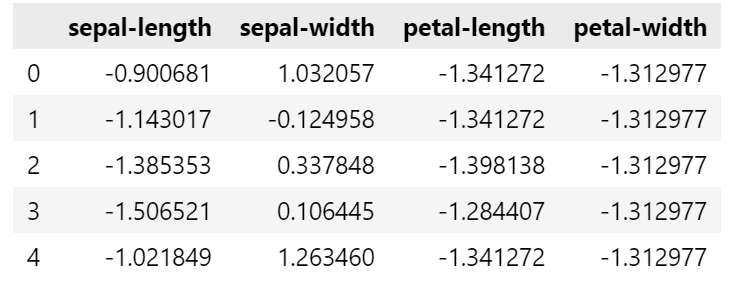
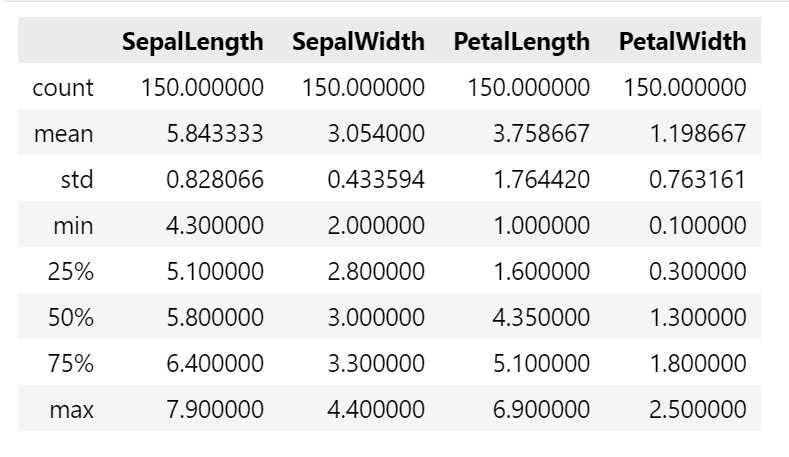
Plt.xlable()

Plt.ylable()

StandardScaler()

三、实验过程记录

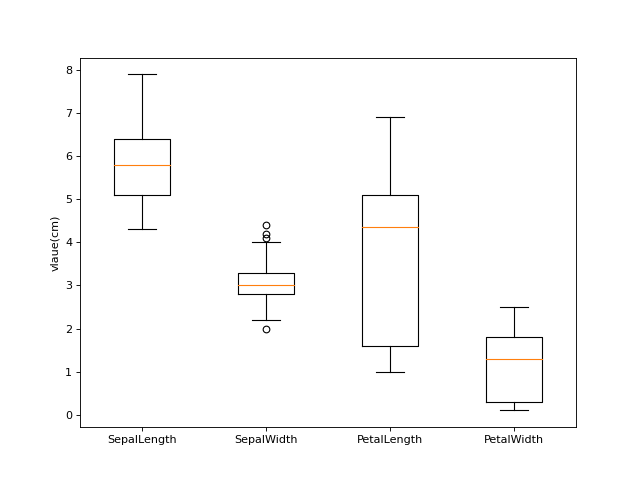
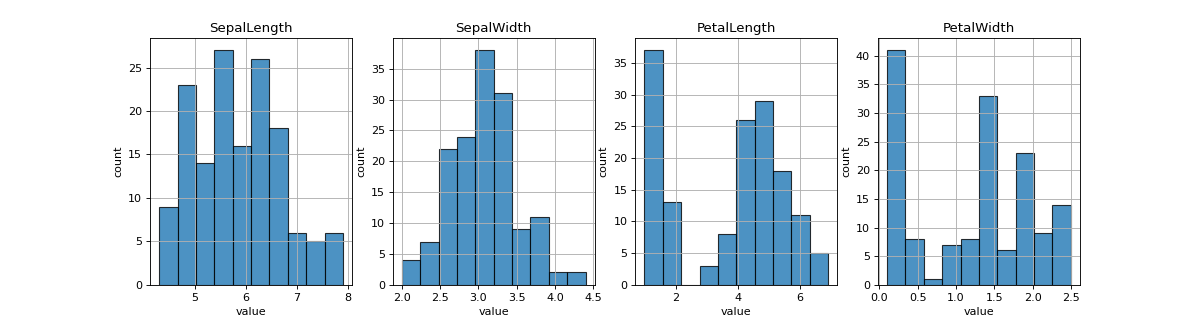
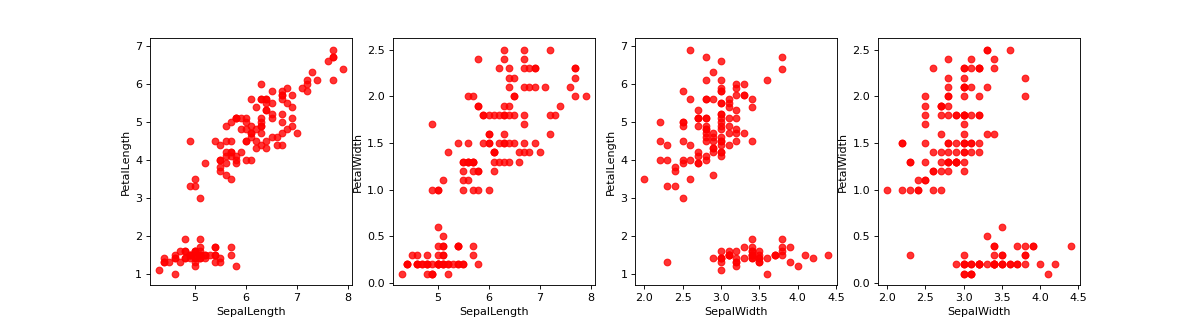
1.导入鸢尾花数据集（'http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data），对数据进行标准化处理

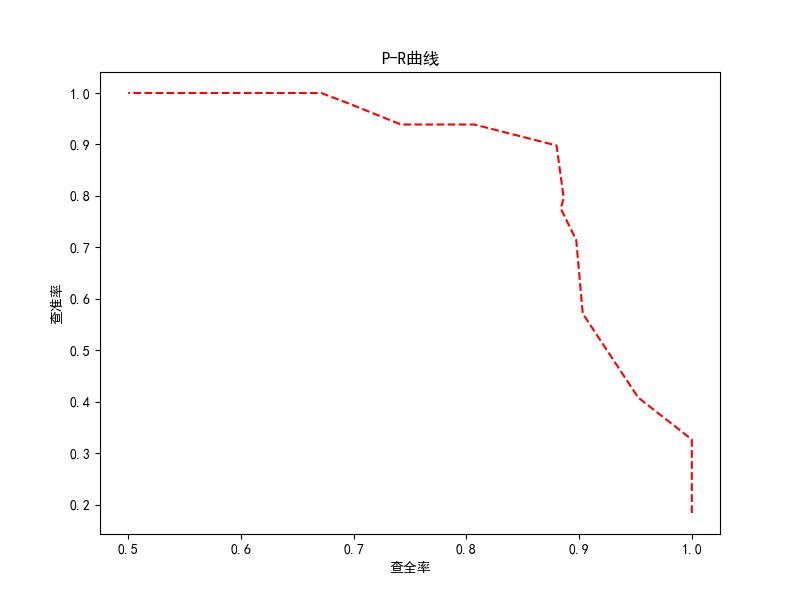
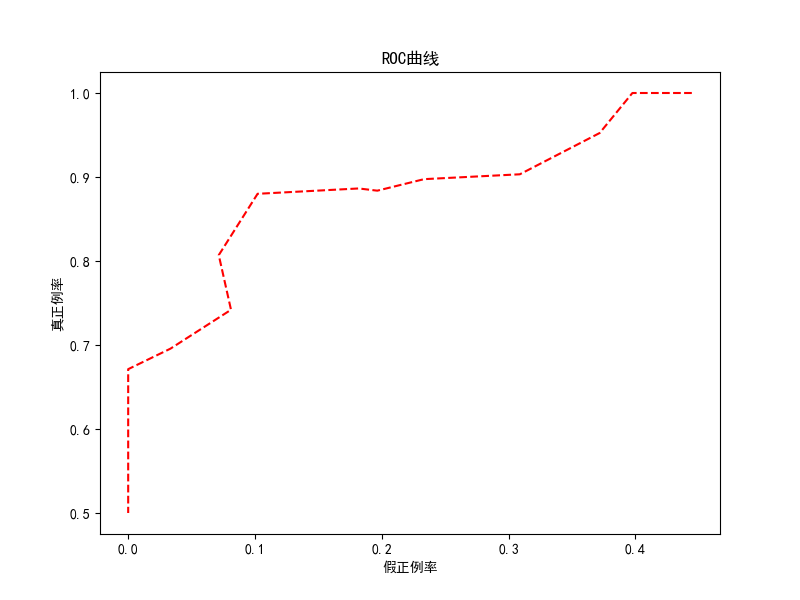


2.分别进行直方图、散点图、箱式图的绘制

3.从4.5的标准开始，通过for语句和if语句的循环嵌套，绘制pr和roc曲线

四、结果与分析

 1. 直方图、散点图、箱式图

 2. PR and ROC

五、自评

经历几天几夜的耕耘后，数据预处理方法、分类聚类方法、分类性能的评价估计、ROC和PR曲线评估结果的计算过程、决策数分类器等等均已经大致掌握。

但是由于自身的python 编程能力还不是很高，在编程方面还是有些许吃力，还是希望自己之后再接再厉，争取更上一层楼。