



**数据挖掘**

**上机实验报告**

|  |  |
| --- | --- |
| 专 业： |  |
| 班 级： |  |
| 学 号： |  |
| 姓 名： |  |

2024年10月14日

## 一、实验目的

1. 掌握使用python进行数据清洗、数据规约、数据规范化和数据离散化。

2. 回顾线性回归用于预测任务。

## 二、实验内容

#### 1. 加利福尼亚房价预测实验

（1）明确数据集中每个属性的含义；

（2）选择合适的方法处理数据集中的缺失值；

（3）判断哪些属性需要规范化，选择合适的方法对这些属性进行数据规范化；

（4）判断哪些属性可以离散化，选择合适的方法对这些属性进行数据离散化；

（5）将数据集的前70%作为训练集，剩余的30%作为测试集（去掉包含缺失值的样本）。对训练集分别随机抽取10%，30%，50%和80%的样本作为训练子集，利用训练子集训练线性回归模型，然后在测试集上预测房价属性。以MSE作为评估指标，观察不同采样率下模型性能的变化；

（6）结合数据集的现实背景，思考能否优化以上步骤，使这些步骤更符合实际情况？

**提交入口：**

<https://send2me.cn/2Y9cMJ4g/Si2jXtvZGC9YGA>

## 三、实验报告评分标准

1. 完成实验（1）（5分）

2. 完成实验（2）（10分）

3. 完成实验（3）（10分）

4. 完成实验（4）（15分）

5. 完成实验（5）（30分）

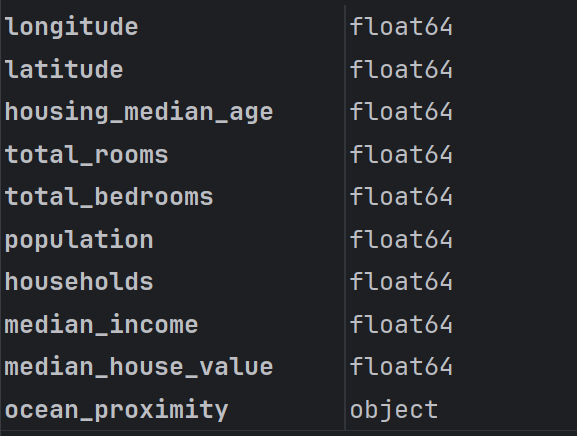
6. 实验过程与实验结论的记录完整（10分）

7. 表述逻辑清晰（10分）

8. 排版工整自洽，图表标题准确（10分）

## 四、实验过程

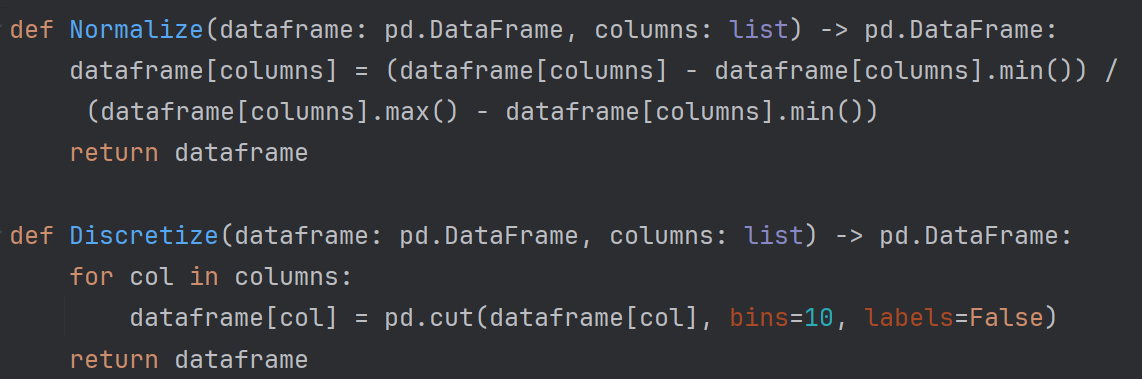
观察数据集



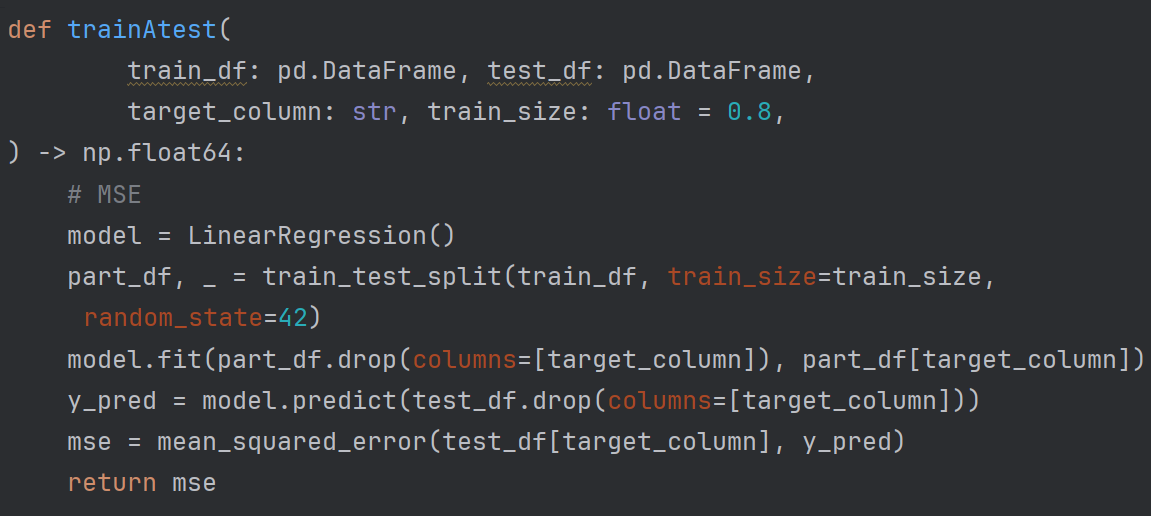
缺失值处理：dropna舍弃

ocean\_proximity: 由于是文本，所以使用LabelEncoder编码

规范化和离散化



训练并测试模型性能



## 五、实验结论

1. 各个属性的含义

'longitude', 精度

'latitude', 纬度

'housing\_median\_age', 中位数年龄

'total\_rooms', 房间数

'total\_bedrooms', 卧室数

'population', 人口

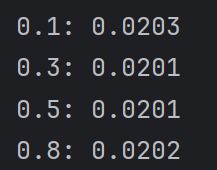
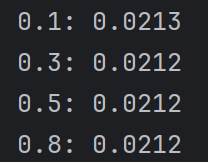
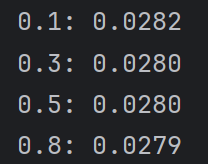
'households', 家庭

'median\_income', 中位数收入

'ocean\_proximity', 与海洋的距离（文本型）

'median\_house\_value', 房价中位数（目标）

1. 缺失值使用dropna舍弃
2. 由于使用的模型是线性模型，因此规范化属性的数量并不会对模型的性能产生显著影响，本文对所有属性都进行了规范化
3. 在线性模型中，离散化实际上会使得样本的噪声更大，主要用于防止过拟合。但是本文使用的数据集较好，因此没有离散化的必要，这在下文中会给出结果作为佐证。从实际角度出发，年龄(housing\_median\_age)和收入(median\_income)两项可以适当离散化。
4. 对所有属性都不进行离散化 对4中两项离散化 全部离散化

1. 缺失值可以使用插值手段填充，模型可以使用随机森林等复杂模型，采样可以使用交叉验证。
2. 代码可以通过 github 查看