Regression Analysis

请于 4 月 10 日前将作业命名为姓名+学号+第一次作业发送到助教邮箱 xinfeiliu@mail.ustc.edu.cn

1、Dataset

Boston house price dataset, which has a dimension of (506,14), including 506 data, and each data contains 14 feature dimensions. The characteristic dimension includes the following 13 dimensions and the corresponding house price (Target).

Columns:

CRIM: Per capita crime rate by town

ZN: Proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS: Proportion of non-retail business acres per town

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX:Nitric oxides concentration (parts per 10 million)

RM: Average number of rooms per dwelling

AGE: Proportion of owner-occupied units built prior to 1940

DIS:Weighted distances to five Boston employment centres

RAD: Index of accessibility to radial highways

TAX: Full-value property-tax rate per $10,000

PTRATIO:Pupil-teacher ratio by town

B: $1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town

LSTAT:Percentage of lower status of the population

MEDV: Median value of owner-occupied homes in $1000's

2、Methods

1）由于 13 个特征全部使用，会过度拟合数据中的噪声。为此可以使用相关性分析或主成分分析等机器学习方法挑选波士顿房价数据集中的主要特征，以下列举了一些可能的相关分析方法：

a) 相关性分析：通过计算各个特征与房价之间的相关性，选择与房价具有较高相关性的特征作为主要特征。

b) 主成分分析：将原始特征进行降维，提取能够解释大部分数据方差的主成分作为主要特征。

2) 推荐的方法：上课时讲到的线性回归或者构建神经网络。

下面以搭建全连接神经网络方法为例阐述具体的操作过程：

a) 数据准备：将波士顿房价数据集的前 450 条作为训练集，后 50 条作为测试集。

b) 数据预处理：对训练集和测试集的特征进行预处理（如：对缺失值进行处理），确保输入数据有相同的维度。

c) 神经网络构建：在 Python 中使用 Pytorch 等神经网络框架构建一个适合回归问题的神经网络模型，包括输入层、隐藏层和输出层。可以选择使用多个隐藏层和不同的激活函数测试效果。

d) 模型训练：使用训练集对神经网络模型进行训练，通过反向传播算法更新权重和偏置。

e) 模型测试和评估：使用测试集对训练好的神经网络模型进行测试，计算预测值与实际值之间的均方误差作为评价指标。