



# 燕山大学

## Python 机器学习实验指导书

---

Python machine learning Experiment Instruction Book

### 实验一：线性回归

教 务 处

2021 年 2 月

# 实验一 线性回归

## 一、实验目的

1. 理解并掌握经典的线性回归模型。
2. 熟悉并掌握 AI Studio 实践平台的账户创建与实践基本操作。
3. 能够基于线性回归模型进行数据分析与预测。

## 二、实验原理

### （一）线性回归模型

#### 1. 线性回归

线性回归：目标值预期是输入变量的线性组合。线性模型形式简单、易于建模，但却蕴含着机器学习中一些重要的基本思想。简单来说，就是选择一条线性函数来很好的拟合已知数据并预测未知数据。

经典的线性回归模型主要用来预测一些存在着线性关系的数据集。回归模型可以理解为：存在一个点集，用一条曲线去拟合它分布的过程。如果拟合曲线是一条直线，则称为线性回归。如果是一条二次曲线，则被称为二次回归。如果包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归。线性回归是回归模型中最简单的一种。

#### 2. 损失函数

要根据已知数据集，在假设空间中，选出最合适的线性回归模型，就要找出使损失函数最小的向量。

线性回归中，损失函数用均方误差表示，即最小二乘法。

最小二乘法：有很多的给定点，需要找出一条线去拟合它。先假设这个线的方程，然后把数据点代入假设的方程得到观测值，求使得实际值与观测值相减的平方和最小的参数。

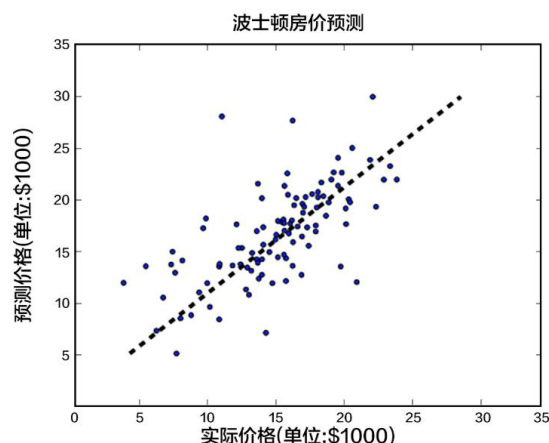
### （二）数据集

#### 1. 波士顿房价数据集（uci-housing）

波士顿数据集共 506 行，每行 14 列。前 13 列用来描述房屋的各种信息，最后一列为该类房屋价格中位数。

注：中位数  $\neq$  平均值。把一组样本的变量值按升或降序排列，处于中间位置的那个数就是这组样本的中位数。

属性名	解释
CRIM	该镇的人均犯罪率
ZN	占地面积超过25,000平方呎的住宅用地比例
INDUS	非零售商业用地比例
CHAS	是否邻近 Charles River
NOX	一氧化氮浓度
RM	每栋房屋的平均客房数
AGE	1940年之前建成的自用单位比例
DIS	到波士顿5个就业中心的加权距离
RAD	到径向公路的可达性指数
TAX	全值财产税率
PTRATIO	学生与教师的比例
B	$1000(BK - 0.63)^2$ , 其中BK为黑人占比
LSTAT	低收入人群占比
MEDV	同类房屋价格的中位数



散点图展示了使用模型对部分房屋价格进行的预测。

- 每点的横坐标：同一类房屋真实价格的中位数。
  - 每点的纵坐标：线性回归模型根据特征预测的结果
- 当二者值完全相等的时候就会落在虚线上。  
模型预测得越准确，则点离虚线越近。

## 2. 糖尿病情数据集 (diabetes)


diabetes 是一个关于糖尿病的数据集,该数据集包括 442 个病人的生理数据及一年以后的病情发展情况。 对 442 例糖尿病患者,分别获得了 10 个基线变量、年龄、性别、体重指数、平均血压和 6 个血清测量值,以及兴趣反应(基线后一年疾病进展的定量测量)。


属性数: 前 10 列是数值预测值。


目标: 第 11 列是基线检查后一年疾病进展的定量测量。




该数据集共 442 条信息,特征值总共 10 项,如下:

- age: 年龄
- sex: 性别
- bmi (body mass index): 身体质量指数,是衡量是否肥胖和标准体重的重要指标,理想 BMI  $(18.5 \sim 23.9) = \text{体重(单位 Kg)} \div \text{身高的平方(单位 m)}$
- bp (blood pressure): 血压(平均血压)
- s1, s2, s3, s4, s5, s6: 六种血清的化验数据,是血液中各种疾病级数指针的 6 的属性值。

 s1——tc, T 细胞

 s2——ldl, 低密度脂蛋白

 s3——hdl, 高密度脂蛋白

-  s4——tch, 促甲状腺激素
-  s5——ltg, 拉莫三嗪
-  s6——glu, 血糖水平

【注意】：以上的数据是经过特殊处理，10 个数据中的每个都做了均值中心化处理，然后又用标准差乘以个体数量调整了数值范围。验证就会发现任何一列的所有数值平方和为 1。这 10 个特征变量中的每一个都以平均值为中心，并按标准差乘以“n\_samples”（即每列的平方和总计为 1）进行缩放。

### 三、实验环境

计算机；网络环境。

### 四、实验内容及步骤

#### （一）实验内容

1. 在 AI Studio 实践平台创建账户，加入本课程。
2. 熟悉 AI Studio 实践平台的基本操作。
3. 对于给定的 3 个例题，基于线性回归模型进行波士顿房价预测、疫情预测等练习。
4. 对于给定的 2 个项目，自行编写程序，基于线性回归模型对糖尿病、影厅观影人数进行回归分析与预测。

#### （二）实验步骤

1. 注册并登录百度 AI Studio 实践平台，加入本课程。
2. 参考教师给定的资料，熟悉 AI Studio 实践平台的基本操作。
3. 进入指定实验课程，进行“实验一：线性回归”的实践操作练习。
  - 例题 1——波士顿房屋价格的拟合与预测（单变量线性回归）
  - 例题 2——波士顿房价预测（三种回归模型对比）
  - 例题 3——基于线性回归模型的疫情预测
4. 在步骤 3 例题练习的基础上，对给定的 2 个实操项目，自行编写程序，对糖尿病、影厅观影人数的线性回归分析与预测。
  - 实操项目 1——糖尿病病情预测
  - 实操项目 2——影厅观影人数预测（多变量线性回归）
5. 完成实验报告。

请严格基于实验报告的模板撰写实验报告。

本课程所有实验全部结束后再统一打印。

## 附：实操项目要求

### ➤ 实操项目 1——糖尿病病情预测

#### 【实验要求】

#### 一、加载糖尿病数据集 `diabetes`，观察数据

1. 载入糖尿病病情数据库 `diabetes`，查看数据。
2. 切分数据，组合成 `DataFrame` 数据，并输出数据集前几行，观察数据。

#### 二、基于线性回归对数据集进行分析

3. 查看数据集信息，从数据集中抽取训练集和测试集。
4. 建立线性回归模型，训练数据，评估模型。

#### 三、考察每个特征值与结果之间的关联性，观察得出最相关的特征

5. 考察每个特征值与结果之间的关系，分别以散点图展示。
- 思考：根据散点图结果对比，哪个特征值与结果之间的相关性最高？

#### 四、使用回归分析找出 XX 特征值与糖尿病的关联性，并预测出相关结果

6. 把 5 中相关性最高的特征值提取，然后进行数据切分。
8. 创建线性回归模型，进行线性回归模型训练。
9. 对测试集进行预测，求出权重系数。
10. 对预测结果进行评价，结果可视化。

### ➤ 实操项目 2——影厅观影人数预测（多变量线性回归）

#### 【实验要求】

1. 读取给定文件中数据。（数据集路径：`data/data72160/1_film.csv`）
2. 绘制影厅观影人数（`filmnum`）与影厅面积（`filmsize`）的散点图。
3. 绘制影厅人数数据集的散点图矩阵。
4. 选取特征变量与相应变量，并进行数据划分。
5. 进行线性回归模型训练。
6. 根据求出的参数对测试集进行预测。
7. 绘制测试集相应变量实际值与预测值的比较。
8. 对预测结果进行评价。

封面设计： 贾丽

地 址： 中国河北省秦皇岛市河北大街 438 号

邮 编： 066004

电 话： 0335-8057068

传 真： 0335-8057068

网 址： <http://jwc.ysu.edu.cn>