**线性回归算法**

**一、实验目的**

1、掌握线性回归算法的相关概念；

2、掌握线性回归算法求解问题的流程；

3、能够编写出线性回归算法预测问题的代码；

4、能够分析实验结果，对算法进行评估。

**二、实验原理**

**1、回归预测模型**

数据预测，简而言之就是基于已有的数据集，归纳出输入变量和输出变量之前的数量关系。基于这种数量关系，一方面，可发现对输出变量产生重要影响的输入变量；另一方面，可以对新数据输出变量取值的预测。对数值型输出变量的预测称为回归。

最常见的回归预测模型为：

𝑦=β0+β1X1+β2X2+⋯+β𝑝X𝑝+𝜀 （1）

该模型被称为一般线性模型。其中，y为数值型的输出变量，X𝑖(𝑖=1,2,⋯,𝑝)为输入变量（这里有P个输入变量）。

公式（1）中，𝛽𝑖(𝑖=1,2,⋯,𝑝)为模型参数：𝛽0为截距项，其他参数称为回归系数，度量了𝑋𝑖取值的单位变化给𝑦带来的数量变动；𝜀为随机误差项，体现了模型之外的其他输入变量对𝑦的影响。

若将公式（1）改写为𝑦=𝑓(𝑋)+𝜀，则𝑓(𝑋)即为输入变量和输出变量间的真实数量关系，是数据无法呈现的。预测建模的目的就是要基于数据集得到𝑓(𝑋)的估计值𝑓̂(x)。

一般线性模型假设𝑓(𝑋)为公式（1）所示的线性关系，并在这样的假设前提下，基于数据集并通过估计策略估计出模型参数，进而最终得到𝑓̂(x)。

**2、预测模型参数估计**

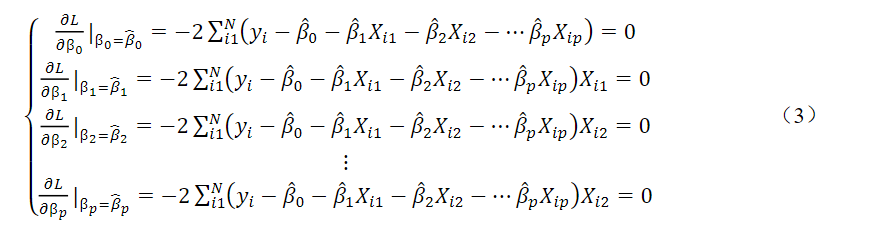
回归预测模型需要基于数据集给出参数的合理估计值，从而使回归平面能够很好地拟合样本观测点。为此，需要设置一个度量指标，测度上述目标尚未到达的程度（也可以是达成程度），一般称之为损失函数（LossFunction），记为L。通常模型的参数估计以损失函数最小为指导目标。

损失函数L是误差e的函数，即为L(e)，与e成正比，用于度量预测模型对数据的拟合误差。回归建模中的误差通过残差来估计，是输出变量实际值𝑦𝑖和预测值𝑦𝑖̂的差，用平方误差的形式，即𝐿(𝑒)=𝐿(𝑦𝑖,𝑦𝑖̂)=(𝑦𝑖−𝑦̂𝑖)2，称为平方损失函数，是回归建模中最常见的损失函数。预测模型中对数据全体的拟合误差也称为总损失为

**3. 参数解空间**

损失函数是关于输出变量实际值和模型参数的函数。例如：回归建模中的平法损失函数可进一步细化为：

∑𝐿𝑁𝑖=1(𝑦𝑖,𝑦𝑖̂)= ∑𝐿𝑁𝑖=1(𝑦𝑖,𝑓̂(𝑋𝑖))=∑[𝑦𝑖−(𝛽̂0+𝛽̂1𝑋𝑖1+𝛽̂2𝑋𝑖2+⋯𝛽̂𝑝𝑋𝑖𝑝)]² (2)

其中，损失函数L是模型参数β𝑖(𝑖=1,2,⋯,𝑝)的二次函数，存在最小值。为求能是损失函数L取的最小值的参数β𝑖(𝑖=1,2,⋯,𝑝)，只需要对参数求偏导，令导数等于0并解方程组得到：

这种参数求解的方法称为最小二乘法。

机器学习常常在预测模型参数的解空间中，采用一定的搜索策略估计参数。预测模型参数的解空间，简单讲就是由所有模型参数解的集合构成的空间，通常是一个高维空间。这里的模型参数不限于一般线性或广义线性模型中的参数，还包括其他任意非线性模型中的参数。

**三、实验环境**

Anaconda（支持Python3.7），使用Jupyternotebook基于网页的交互性文本编辑器，安装Numpy库、Pandas库、matplotlib库和Scikit-learn库。

**四、实验内容及步骤**

**1、实验内容**

(1)**用线性回归模型预测房价**（数据集：housing\_price.csv），分别建立单因子模型（面积为输入量）和多因子模型（以收入、房龄、房间数等为输入变量），输出回归模型，评估模型表现，并可视化模型。

(2)**用线性回归模型预测空气质量**（数据集：北京市空气质量数据.xlsx）,建立PM2.5预测的一元线性回归模型，其中输入变量为CO，输出变量为PM2.5；建立多元线性回归模型对PM2.5进行预测，输入变量为（CO和S02），输出变量为PM2.5，输出回归模型，评估模型表现。（对数据进行预处理，将数据集中的0值替换为缺失值“NaN”；剔除缺失值的样本观测；仅针对PM2.5浓度低于200且SO2浓度低于20的数据集建模）。

**2、实验步骤**

(1)导入Numpy库、Matplotlib库、Pandas库和Scikit-learn库。

(2)导入Scikit-learn提供的线性回归模块sklearn.linear\_model.

(3)使用Pandas库读取数据文件

(4)进行回归分析（单因子模型和多因子模型）

(5)打印输出回归模型参数和回归方程

(6)可视化回归结果，并对新样本点进行预测

(7)计算回归模型MES和R2\_score指标，并分别比较（单因子模型和多因子模型）的指标差异，并作出分析。