单元5

回归算法

任务 01

波士顿房价预测问题

知识目标

- 学习线性回归模型参数求解的原理
- 掌握Sklearn中回归算法的调用

能力目标

- 能够调用Sklearn中的算法解决回归问题

职业素养目标

- 培养学生对所学理论知识的实际运用能力

1

任务1：波士顿房价预测问题

1 任务描述和目标

任务描述

波士顿房价数据集是Sklearn工具包中内置的数据集，共506个样本数据点，涵盖了波士顿不同地区房屋的房价信息，每个样本包括13个特征属性和它的房价。我们的任务是利用回归算法找到一个模型，能对数据集中的样本进行拟合，并对新样本的房价进行预测。

任务目标

- 学习线性回归、多项式回归、Lasso回归、岭回归模型算法的原理
- 了解过拟合和欠拟合的概念和处理方法
- 掌握使用Sklearn中回归算法模型的解决房价预测等回归问题的方法

1 一、单变量线性回归

1.模型定义

一般用来表示参数集合，即通用公式为：

$$y=h_\theta(x)=\theta_0+\theta_1 x$$

其中的x是数据的特征属性值，例如上面例子中房子的面积；y是目标值，即房子的价格。这就是我们常说的线性回归方程。和是模型的参数，参数的学习过程就是根据训练集来确定，学习任务就是用一条直线来拟合训练数据，也就是通过学习找到合适的参数值。

1 一、单变量线性回归

2.损失函数

一般采用损失函数来衡量模型与数据点的接近程度。所有m个样本点的误差项的均方差我们称之为损失函数（或目标函数、代价函数）

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$$

显然，J越小，模型就能越好的描述样本数据，我们的目标就是找到使J取到最小值时的参数

1 一、单变量线性回归

3.最小二乘法求解

根据微积分的知识，在连续函数的最小值处，函数关于参数的偏导数为0。对J分别求偏导数，如图。然后令两式等于0，得到：

$$\frac{\partial J}{\partial \theta_0}=\frac{\partial}{\partial \theta_0}\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2=\frac{\partial}{\partial \theta_0}\frac{1}{m}\sum_{i=1}^{m}(\theta_0+\theta_1x^{(i)}-y^{(i)})^2$$

$$=\frac{2}{m}\sum_{i=1}^{m}(\theta_0+\theta_1x^{(i)}-y^{(i)})=2\theta_0+\frac{2}{m}\sum_{i=1}^{m}(\theta_1x^{(i)}-y^{(i)})$$

$$\frac{\partial J}{\partial \theta_1}=\frac{\partial}{\partial \theta_1}\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2=\frac{\partial}{\partial \theta_1}\frac{1}{m}\sum_{i=1}^{m}(\theta_0+\theta_1x^{(i)}-y^{(i)})^2$$

$$=\frac{2}{m}\sum_{i=1}^{m}(\theta_0+\theta_1x^{(i)}-y^{(i)})x^{(i)}=\frac{2\theta_0}{m}\sum_{i=1}^{m}x^{(i)}+\frac{2\theta_1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)}-\frac{2}{m}\sum_{i=1}^{m}x^{(i)}y^{(i)}$$

$$=2(\theta_0\bar{x}+\theta_1\overline{x^2}-\overline{xy})$$

$$\theta_0=\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}-\theta_1x^{(i)})=\bar{y}-\theta_1\bar{x}$$

$$\theta_1=\frac{\overline{xy}-\bar{x}\bar{y}}{\overline{x^2}-(\bar{x})^2}$$

1 一、单变量线性回归

4.回归效果评价

1）决定系数：也称为判定系数或者拟合优度

总平方和（total sum of squares，SST）定义为$SST = \sum_{i=1}^{n} (y_i - \text{mean}(y))^2$。

回归平方和（regression sum of squares，SSR）定义为$SSR = \sum_{i=1}^{n} (f_i - \text{mean}(y))^2$。

误差平方和（error sum of squares，SSE）定义为$SSE = \sum_{i=1}^{n} (y_i - f_i)^2$。

决定系数定义为R^2 = SSR/SST = 1-SSE/SST，可以看出R^2越大则拟合得越好。

2）剩余标准差（Root mean squared error，RMSE）：也称为均方根误差、标准误差、残差平方和，定义为

$$S = \sqrt{SSE/n}$$

其中，n为样本数量，因此也可以将S看成是平均残差平方和的算术根，其值越小，则拟合得更好。

1

二、多变量线性回归

1.回归方程

对于多变量回归，例如房价高低跟房子面积、楼层位置、装修情况有关，通用的模型为：

$$y = h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

假设我们的问题中有m个样本（例如m个房子），每个样本有n个特征和一个对应的真实结果y（即真实房价），那么可以将模型用矩阵形式表达为：

$$y = h_\theta(X) = X\theta = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \cdots \\ X^{(m)} \end{bmatrix} \theta$$

1

二、多变量线性回归

2.损失函数

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2=\frac{1}{m}(X\theta-y)^T(X\theta-y)$$

3.最小二乘法求解

对损失函数求偏导并令其等于0，得到的θ就是模型参数的值。

$$J(\theta)=\frac{1}{m}(X\theta-y)^T(X\theta-y)=\frac{1}{m}(\theta^TX^T-y^T)(X\theta-y)$$

$$=\frac{1}{m}(\theta^TX^TX\theta-\theta^TX^Ty-y^TX\theta+y^Ty)$$

$$\nabla_\theta J(\theta)=\frac{2}{m}(X^TX\theta-X^Ty)$$

$$\theta=(X^TX)^{-1}X^Ty$$

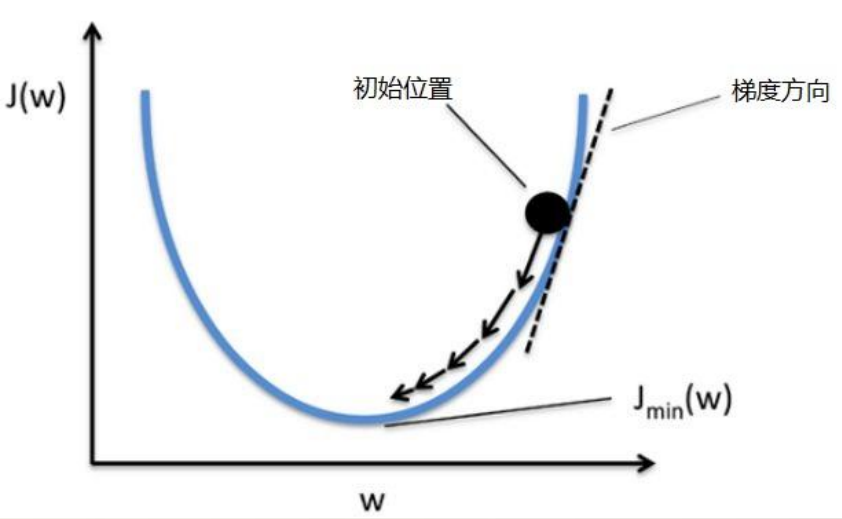
1

二、多变量线性回归

4.梯度下降法求解

梯度下降法的基本思想可以类比为一个下山的过程。一个人需要从山上快速准确地下山到达山谷，首先以他当前的所处的位置为基准，寻找这个位置最陡峭的地方，然后朝着下降方向走一步，然后又继续以当前位置为基准，再找最陡峭的地方，再走直到最后到达最低处。

首先对参数θ取一个随机初始值，然后不断迭代改变θ的值使损失函数J(θ)根据梯度下降的方向减小，直到收敛求出某θ值使J(θ)达到最小或局部最小。θ更新规则为：



$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

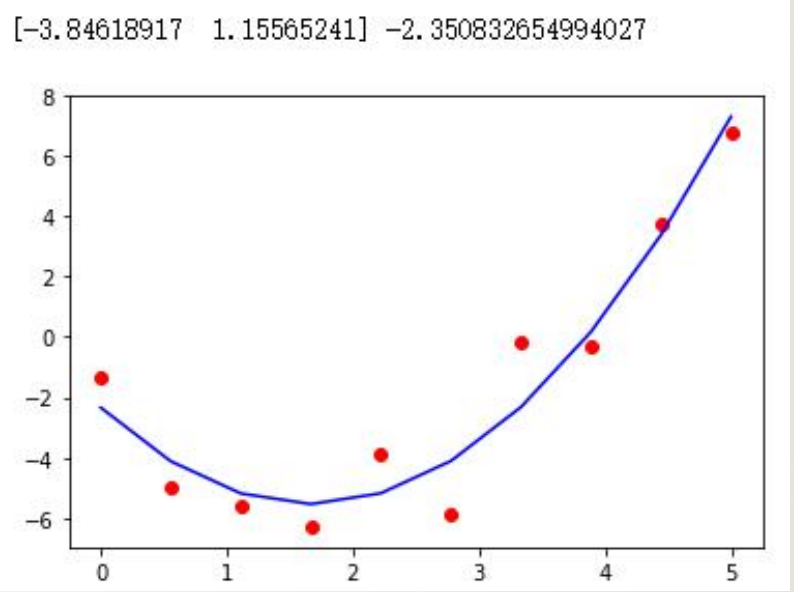
1

三、多项式回归与正则化

1.多项式回归

如果训练集的散点图没有呈现出明显的线性关系，而是类似于一条曲线的样子，就像图中这样。我们尝试用多项式回归对它进行拟合。

我们先看一下二次曲线的方程y=ax^2+bx+c，如果将x^2理解为一个特征，将x理解为另外一个特征，那么就可以看成有两个特征的一个数据集，这个式子依旧是一个线性回归的式子。这样就将多项式回归问题，转化为多变量线性回归模型。



1

三、多项式回归与正则化

2.正则化处理

损失函数J也称为经验风险，通过使经验风险最小化而学到的模型经常会出现过拟合，因此就需要增加正则化项，此时损失函数就变成：

$$J=\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}-\theta^{T}x^{(i)})^{2}+\lambda R(\theta)$$

$$R(\theta)=\frac{1}{m}\sum_{j=0}^{n}\theta_{j}^{2}$$

称为结构风险。其中λ为正则化参数，其作用是控制拟合训练数据的目标和保持参数值较小的目标之间的平衡关系。

1

三、多项式回归与正则化

3.非线性回归

因变量和和自变量的关系是非线性的，这种情况下的回归问题就是非线性回归。常见的非线性模型有双曲线模型、幂函数模型、指数模型和对数模型

双曲线模型形式是：$\frac{1}{h(x)} = w_0 + \frac{w_1}{x_1} + \frac{w_2}{x_2} + \ldots + \frac{w_n}{x_n}$

幂函数模型形式是：$h(x) = w_0 x_1^{w_1} x_2^{w_2} x_3^{w_3} \ldots x_n^{w_n}$

指数模型形式是：$h(x) = w_0 e^{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}$

对数模型形式是：$h(x) = w_0 + w_1 \ln x_1 + w_2 \ln x_2 + \ldots + w_n x_n$

任何一种函数都可以用多项式来逼近，因此非线性回归可以转换为多项式回归。

1

四、过拟合和欠拟合

1.欠拟合

欠拟合指的是模型在训练和预测时表现都不好的情况，需要继续学习。

解决欠拟合的方法：

- 添加其他特征项，一般可以通过组合、抽取等得到新的特征。
- 添加多项式特征，例如将线性模型通过添加二次项或者三次项使模型泛化能力更强。
- 减少正则化参数，正则化的目的是用来防止过拟合的，但是模型出现欠拟合，则需要减少正则化参数。

1

四、过拟合和欠拟合

2.过拟合

过拟合指的是模型对于训练数据拟合程度过当，造成训练集效果很好，而测试集效果较差的情况。解决过拟合的方法：

- 重新清洗数据，出现过拟合有可能是数据不纯导致的。
- 增大数据的训练量，出现过拟合也有可能是训练的数据量太小导致的，训练数据占总数据的比例过小。
- 减少样本的特征维度，有些特征可能对结果的影响非常小，可以忽略。
- 采用正则化方法，正则化方法包括L0正则、L1正则和L2正则。
- 采用dropout方法，这个方法在神经网络里面很常用，通俗一点讲就是在训练的时候让神经元以一定的概率不工作。

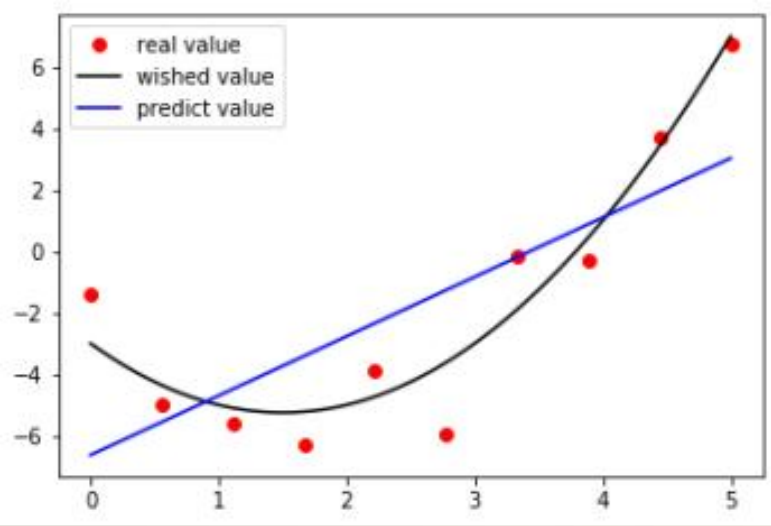
1

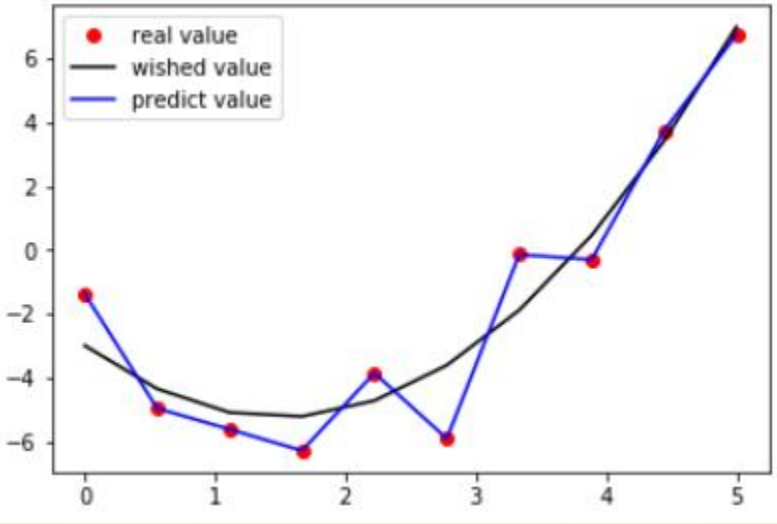
四、过拟合和欠拟合

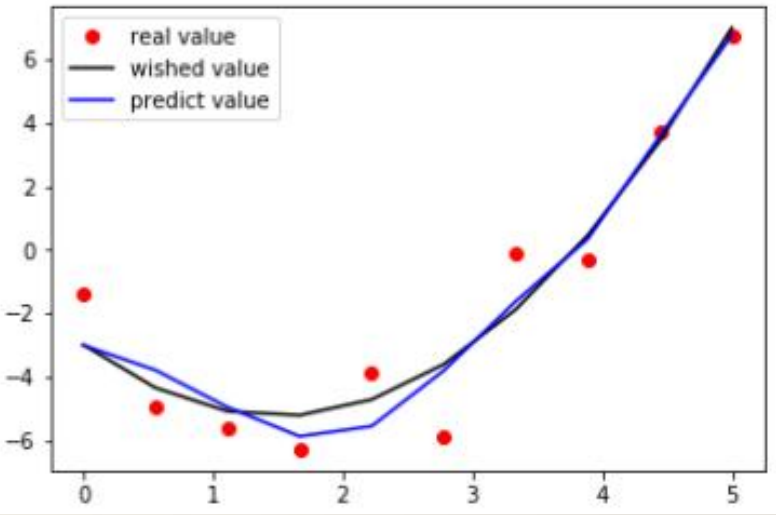
3.较好的拟合

理想上，一个好的模型是一个正好介于欠拟合和过拟合之间的模型。

下面分别是对测试数据进行欠拟合、过拟合、较好拟合的图像。从拟合结果看到，测试数据最好使用二阶多项式进行拟合，虽然该参数和真正的生成函数有一定差距，但是发现拟合效果不错。







1 任务实施

我们采用了几种不同的回归算法对波士顿房价问题进行拟合：

- 首先，使用sklearn.model_selection模块的train_test_split()方法来划分测试集和训练集。
- 接着，我们需要定义R2函数，用于衡量回归模型对观测值的拟合程度。它的定义是R2 = 1-“回归平方和在总平方和中所占的比率”，可见回归平方和（即预测值与实际值的误差平方和）越小，则R2越接近于1，预测越准确，模型的拟合效果越好。
- 然后，我们分别使用sklearn中的线性回归模型、多项式回归模型、Lasso模型和Ridge模型对波士顿房价问题进行训练和测试，并比较各回归模型拟合结果的R2误差值。

作业

1.简述调用Sklearn中线性回归模型解决波士顿房价预测问题的步骤。

2.简述过拟合、欠拟合的现象及处理方法。

3.本集数据划分函数train_test_split的用法。

4.房价拟合案例中，调用Sklearn的多变量线性回归模型和多项式回归模型进行拟合的主要步骤和方法。

Thank YOU!