5 线性回归算法

目录

CONTENTS

1.1线性回归

Linear Regression

基本形式

属于监督学习

给定d个属性描述的示例 $\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$ 以及对应 $\boldsymbol{y}$

我们希望习得属性的线性组合来进行预测的函数

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} \omega_i x_i + b$$

向量形式

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{x} + b$$

如预测房价：

Size(feet^2)	Num of bedrooms	Num of floors	Age of home(years)	Price($1000)
2104	5	1	45	460
1416	2	2	40	232
15340	3	2	30	315

单属性线性回归

给定数据集 $D = \{(x_i, y_i)\}_{i=1}^{m}$

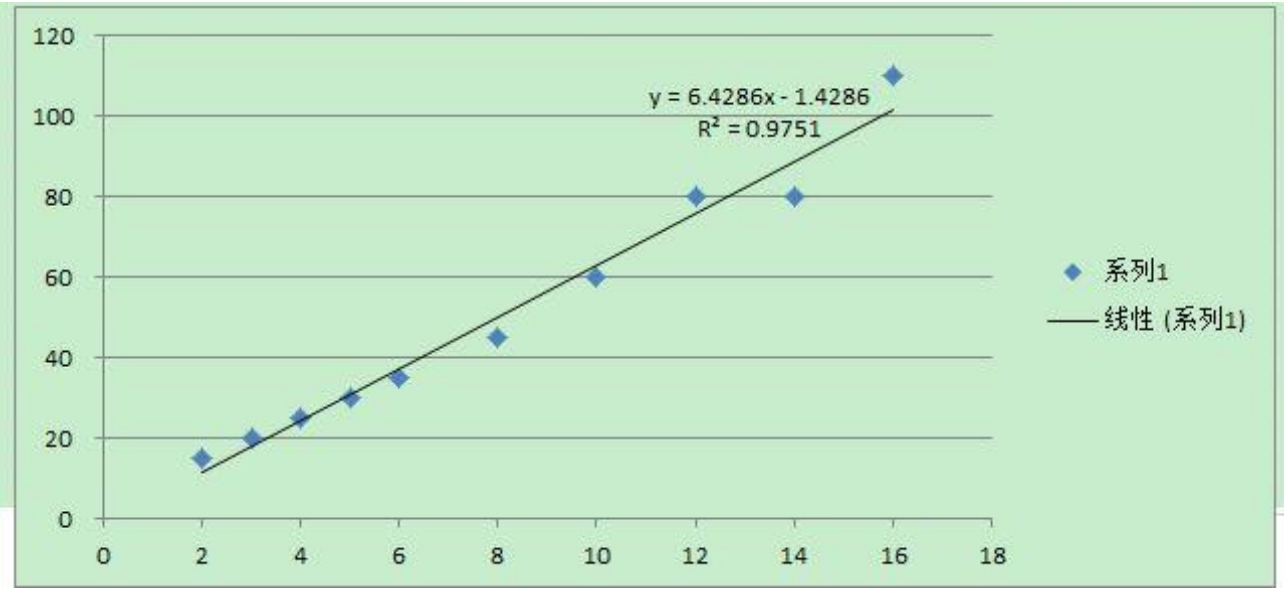
线性预测： $f(x_i) = \omega x_i + b$ 从训练集得到的ω和b的值，使得$f(x_i) \approx y_i$

如何衡量?

$$(\omega^*, b^*) = \arg\min_{(\omega,b)} \sum_{i=1}^{m} [f(x_i) - y_i]^2 = \arg\min_{(\omega,b)} \sum_{i=1}^{m} [\omega x_i + b - y_i]^2$$

最小二乘法

Least square method



$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right) ,$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right) ,$$

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2} ,$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i) ,$$

多属性线性回归

给定数据集 D，有m个样本，d个属性，可以写成矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{bmatrix}=\begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{bmatrix}$$

属性：$\vec{x}_i=(x_{i1},\dots,x_{id})^T$

权重：$\vec{\hat{\omega}}=(\omega_1,\dots,\omega_d,b)^T$

$$\hat{\boldsymbol{\omega}}^*=\arg\min_{\hat{\boldsymbol{\omega}}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\omega}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\omega}})$$

矩阵求导之后 得：$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega}=\boldsymbol{X}^T\boldsymbol{y}$

$$\frac{\partial E_{\hat{\omega}}}{\partial \hat{\omega}}=2X^T(X\hat{\omega}-\mathrm{y})$$

当$\boldsymbol{X}^T\boldsymbol{X}$可逆时 $\hat{\boldsymbol{\omega}}^*=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

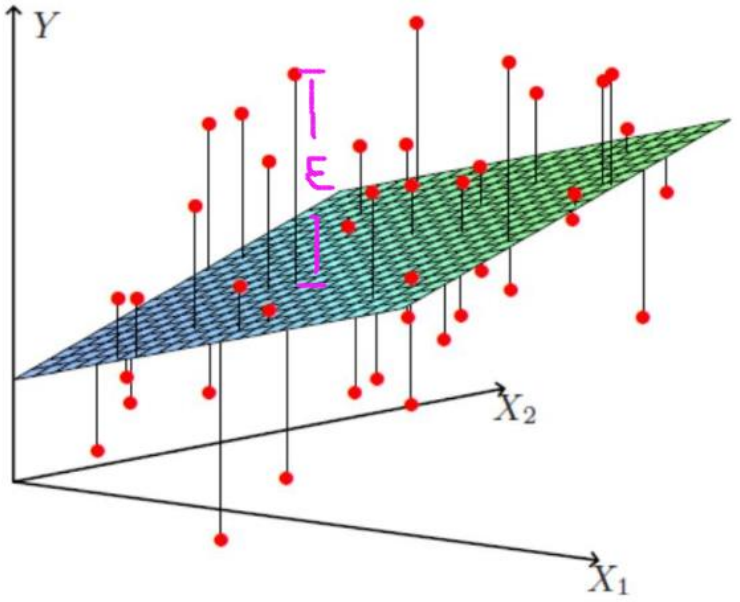
当$\boldsymbol{X}^T\boldsymbol{X}$不可逆时（模型参数太多，训练样本太少） 可以引入正则项$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}$ （或求$(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}$）

对 $J(\boldsymbol{\omega})=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\omega}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\omega}})+\lambda\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}$ 进行整体优化，λ称为超参数

附函数对向量求导公式：

$$\frac{d\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{d\boldsymbol{x}}=(\boldsymbol{A}+\boldsymbol{A}^T)\boldsymbol{x}$$

$$\boldsymbol{A}=\boldsymbol{A}^T\Rightarrow\frac{d\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{d\boldsymbol{x}}=2\boldsymbol{A}\boldsymbol{x}$$



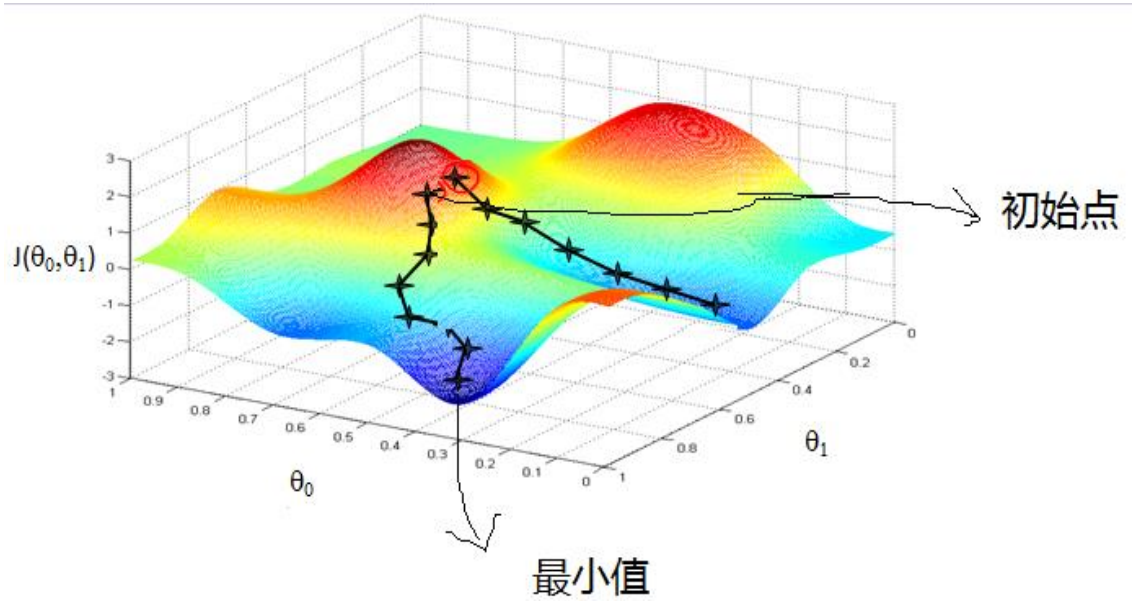
两个属性的线性回归

梯度下降法

损失函数： $J(\omega) = \min_{\omega} \|X\omega - y\|^2$

关于ω的梯度： $\nabla J(\omega) = 2X^T(X\omega - y)$

梯度下降法： $\omega_k = \omega_{k-1} - \rho_k \nabla J(\omega)$



当数据量过大或者行列式接近于0，X^TX求逆困难时，可以选择梯度下降法

线性回归优缺点

优点：线性回归的理解和解释都非常直观，还能通过正则化来避免过拟合。可以预测、求出函数，很容易通过随机梯度下降来更新数据模型。还可以进行残差检验，检验精度。

缺点：线性回归在处理非线性关系时非常糟糕，在识别复杂的模式上也不够灵活，而添加正确的相互作用项或多项式又极为棘手且耗时。

PART 01

线性回归算法的基本原理

线性回归算法的概念及数学解释

线性回归算法是一种较为基础的机器学习算法，基于特征（自变量、解释变量、因子、协变量）和响应变量（因变量、被解释变量）之间存在的线性关系。线性回归算法的数学模型为：

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots + \beta_n x_n + \epsilon$$

矩阵形式为：

$$y = \alpha + X\beta + \varepsilon$$

其中：$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ 为响应变量；$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$ 为截距项；$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$ 为待估计系数；

$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$ 为特征；$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ 为误差项。

并且假定特征之间无多重共线性；误差项ε_i（i=1,2,...,n）之间相互独立，且均服从同一正态分布 $N(0,\sigma^2)$，σ^2是未知参数，误差项满足与特征之间的严格外生性假定，以及自身的同方差、无自相关假定。

响应变量的变化可以由$\alpha + X\beta$组成的线性部分和随机误差项ε_i两部分解释。对于线性模型，一般采用最小二乘估计法来估计相关的参数，基本原理是使残差平方和最小，即 $\min \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - \hat{\alpha} - \hat{\beta})$，残差就是响应变量的实际值与拟合值之间的差值。

线性回归算法的优点

一是线性回归算法基于特征和响应变量之间的线性关系，理解起来比较简单，实现起来也比较容易，在处理样本示例容量较小的数据集时比较有效。

二是线性回归算法是许多强大的非线性模型的基础，一方面根据微积分相关原理，在足够小的范围内，非线性关系可以用线性函数来近似，另一方面，很多非线性模型可以通过线性变换的方式转化为线性模型，比如针对二次函数、三次函数、对数函数等非线性关系，我们可以将特征的二次项、三次项、交互项、对数值等命名为新的特征，与原特征值共同构成线性关系。

线性回归算法的优点

三是线性回归模型十分容易理解，结果具有很好的可解释性，在线性回归算法中，特征的系数即为该特征对于响应变量的边际效应，也就是说特征每一单位的增长能够引起响应变量多少单位的增加。系数又可以分两方面来看，一方面通过计算系数的显著性水平观察其统计意义显著性，当其显著性水平小于显著性P值（通常为0.05）时，特征对于响应变量的影响是统计显著的；另一方面通过计算系数大小观察其经济意义显著性，系数越大，说明特征对于响应变量的影响程度越大。

四是线性模型中蕴含着机器学习中的很多重要思想，比如其算法原理中使得均方误差（MSE）最小化，本质上实现的是偏差与方差的权衡等等。

线性回归算法的优点

五是线性模型具有一定的稳定性。从技术角度，我们在评价模型的优劣好坏时，通常从两个维度去评判，一是模型预测的准确性，二是模型预测的稳健性，两者相辅相成、缺一不可。关于模型预测的准确性，如果模型能够尽可能的拟合了历史数据信息，拟合优度很高，损失的信息量很小，而且对于未来的预测都很接近真实的实际发生值，那么模型一般是被认为是质量较高的。而关于模型的稳健性，我们期望的是模型在对训练样本以外的样本进行预测时，模型的预测精度不应该有较大幅度的下降。一般来说，神经网络、决策树的预测准确性要优于判别分析和Logistic回归分析等线性分析，但是其稳健性弱于线性分析。

线性回归算法的缺点

线性回归算法的缺点主要体现在：对于非线性数据或者数据特征间具有相关性多项式回归难以建模，难以很好地表达高度复杂的数据。比如针对商业银行信贷客户违约量化评估与预测问题，如果我们能够较为合理的判定信用风险和各个特征变量是一种线性关系，那么我们完全可以选择线性回归算法。但是如果我们无法较为合理的判定信用风险和各个特征变量之间的关系，那么使用神经网络、决策树建模技术可能就是更好的选择，这些相对更加复杂的建模技术对模型结构和假设施加最小需求，应用到响应变量和特征变量之间关系不明确的情形中。

PART 02

数据准备

数据准备

本节我们用于分析的数据是数据4.1.记录的是XX生产制造企业1994—2021年营业利润水平（profit）、固定资产投资（invest）、平均职工人数（labor）、研究开发支出（rd）数据，以营业利润水平作为响应变量，以固定资产投资、平均职工人数、研究开发支出作为特征变量，开展线性回归算法。

载入分析所需要的模块和函数

本例中需要载入pandas、numpy、matplotlib、seaborn、statsmodels、sklearn等模块。其中pandas、numpy用于数据读取、数据处理、数据计算；matplotlib.pyplot、seaborn、probplot用于绘制图形，实现分析过程及结果的可视化；stats模块用于统计分析；statsmodels中的statsmodels.formula.api、以及sklearn中的LinearRegression用于构建线性回归模型；train_test_split用于把样本随机划分为训练样本和测试样本；mean_squared_error，r2_score模块分别用于计算均方误差（MSE）和可决系数，评价模型优劣。

示例

参阅教材内容

数据读取及观察

大家首先需要将本书提供的数据文件放入安装python的默认路径位置，并从相应位置进行读取。

示例

参阅教材内容

PART 03

描述性分析

描述性分析

在进行数据分析时，当研究者得到的数据量很小时，可以通过直接观察原始数据来获得所有的信息。但是，当得到的数据量很大时，就必须借助各种描述性指标来完成对数据的描述工作。用少量的描述性指标来概括大量的原始数据，对数据展开描述的统计分析方法被称为描述性统计分析。

示例

参阅教材内容

PART 04

图形绘制

图形绘制

在构建线性回归模型之前，我们可以通过针对变量绘制图形的方式，初步研究下变量的分布特征，常用的图形绘制包括直方图、密度图、小提琴图、箱图、正态 QQ 图、散点图和线图、热力图、回归拟合图、联合分布图等，这些图形绘制方法可以帮助用户快速了解数据点的分布，还可以发现异常值的存在。

直方图

直方图（Histogram）又称柱状图，是一种统计报告图，由一系列高度不等的纵向条纹或线段表示数据分布的情况。一般用横轴表示数据类型，纵轴表示分布情况。通过绘制直方图可以较为直观地传递有关数据的变化信息，使数据使用者能够较好地观察数据波动的状态，使数据决策者能够依据分析结果确定在什么地方需要集中力量改进工作。

示例

参阅教材内容

密度图

密度图（Density Plot）用于显示数据在连续时间段内的分布状况，是直方图的进化，使用平滑曲线来绘制数值水平，从而得出更平滑的分布。密度图的峰值显示数值在该时间段内最为高度集中的位置。相对于直方图，密度图不受所使用分组数量（直方图中的条形）的影响，所以能更好地界定分布形状。

示例

参阅教材内容

箱图

箱图（Box-Plot）又称为盒须图、盒式图或箱线图，是一种用于显示一组数据分散情况的统计图。箱图很形象地分为中心、延伸以及分部状态的全部范围，提供了一种只用5个点总结数据集的方式，这5个点包括最小值、第一个四分位数Q1、中位数点、第三个四分位数Q3、最大值。数据分析者通过绘制箱图不仅可以直观明了地识别数据中的异常值，还可以判断数据的偏态、尾重以及比较几批数据的形状。

示例

参阅教材内容

小提琴图

小提琴图其实是箱式图与密度图的结合，箱式图展示了分位数的位置，小提琴图则展示了任意位置的密度，小提琴图可以展示密度较高的位置。通过使用密度曲线描述一组或多组的数值数据分布。

示例

参阅教材内容

正态 QQ 图

正态 QQ 图是由标准正态分布的分位数为横坐标，样本值为纵坐标的散点图，通过把测试样本数据的分位数与已知分布相比较，从而来检验数据是否服从正态分布。如果QQ图中的散点近似地在图中的线附近，就说明是正态分布，而且该直线的斜率为标准差，截距为均值。

示例

参阅教材内容

散点图和线图

作为对数据进行预处理的重要工具之一，散点图（Scatter Diagram）深受专家、学者们的喜爱。散点图的简要定义就是点在直角坐标系平面上的分布图。研究者对数据制作散点图的主要出发点是通过绘制该图来观察某变量随另一变量变化的大致趋势，据此可以探索数据之间的关联关系，甚至选择合适的函数对数据点进行拟合。

示例

参阅教材内容

热力图

热力图用于表现某种事物密集度的图形化显示，是展示差异非常直观的方法。热力图的右侧是颜色带，也叫图例说明，上面代表了数值到颜色的映射，数值由小到大对应色彩由浅到深。数据值在图形中以颜色的深浅来表示数量的多少，可以快速找到最大值的与最小值所在位置，在机器学习的分类中经常用来作混淆矩阵的比较。

示例

参阅教材内容

回归拟合图

散点图只能大致显示响应变量和特征之间的关系，为了深入研究其拟合关系，可以通过绘制回归拟合图的方式进行观察，回归拟合图应用最小二乘法原理，让误差的平方和最小，但回归拟合图反映的只是大概，并不够精确，只能为后续真正做数据拟合提供参考信息。

示例

参阅教材内容

联合分布图

联合分布图是一个多面板图形，比如散点图、二维直方图、核密度估计等，用来显示两个变量之间的双变量关系及每个变量在单独坐标轴上的单变量分布。

示例

参阅教材内容

PART 05

正态性检验

正态性检验

正态分布，又称高斯分布（Gaussian distribution）。若随机变量 X 服从一个数学期望为μ、方差为σ^2的正态分布，记为 $N(\mu, \sigma^2)$，其中期望值μ决定了其位置，其标准差σ决定了分布的幅度。当$\mu = 0, \sigma = 1$ 时的正态分布是标准正态分布。有相当多的统计程序对数据要求比较严格，它们只有在变量服从或者近似服从正态分布的时候才是有效的，所以在对整理收集的数据进行预处理的时候需要对它们进行正态检验。Python 中常用的正态性检验包括 Shapiro-Wilk test 检验和 kstest 检验。

Shapiro-Wilk test检验

示例

参阅教材内容

kstest检验

示例

参阅教材内容

PART 06

相关性分析

相关性分析

相关性分析通过计算皮尔逊简单相关系数、斯皮尔曼等级相关系数、肯德尔系数展开。其中皮尔逊简单相关系数是一种线性关联度量，适用于变量为定量连续变量且服从正态分布、相关关系为线性时的情形。如果变量不是正态分布的，或具有已排序的类别，相互之间的相关关系不是线性的，则更适合采用斯皮尔曼、肯德尔等级相关系数。

相关系数r有如下性质：

① $-1 \leqslant r \leqslant 1$，$r$绝对值越大，表明两个变量之间的相关程度越强。

② $0 < r \leqslant 1$，表明两个变量之间存在正相关。若$r=1$，则表明变量间存在着完全正相关的关系。

③ $-1 \leqslant r < 0$，表明两个变量之间存在负相关。$r=-1$表明变量间存在着完全负相关的关系。

④ $r=0$，表明两个变量之间无线性相关。

应该注意的是，相关系数所反映的并不是一种必然的、确定的关系，也不能说明变量之间的因果关系，而仅仅是关联关系。

相关性分析

说 明

皮尔逊：线性关联度量，适用于变量为定量连续变量且服从正态分布、相关关系为线性时的情形。若随机变量 X 、 Y 的联合分布是二维正态分布， x_i 和 y_i 分别为 n 次独立观测值，皮尔逊相关系数公式为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

“斯皮尔曼”“肯德尔”为等级相关系数，当数据资料不服从双变量正态分布或总体分布未知，或原始数据用等级表示时，宜选择斯皮尔曼、肯德尔等级相关系数。

示例

参阅教材内容

PART 07

使用 smf 进行线性回归

使用 smf 进行线性回归

在python中，进行线性回归的模块主要包括statsmodels模块和sklearn模块。其中statsmodels模块的优势在于不仅可以进行预测，还可以进行统计推断，包括计算标准误差、p值、置信区间等；而sklearn模块则无法进行统计推断，也就是不提供标准误差、p值、置信区间等指标结果，但在机器学习方面效能相对更佳。

使用 smf 进行线性回归

示例

参阅教材内容

使用 smf 进行线性回归

多重共线性检验

示例

参阅教材内容

使用 smf 进行线性回归

解决多重共线性问题

示例

参阅教材内容

使用 smf 进行线性回归

绘制拟合回归平面

示例

参阅教材内容

PART 08

使用 sklearn 进行线性回归

使用 sklearn 进行线性回归

大家不难发现，在前面的分析中，我们是使用样本示例全集进行的分析，并没有像第三章介绍的那样，将样本划分为训练样本和测试样本的方式，进而也无法考察算法模型的“泛化”能力，所以从严格意义上讲，前面的内容更多地属于使用python开展统计分析的范畴，而非真正的机器学习。而前面提及的可决系数、修正的可决系数、MSE（均方误差）、AIC及BIC信息准则等指标结果，也都是针对样本示例全集，反映的也都是样本内的预测效果，而不是模型真实泛化能力的度量。下面我们使用sklearn 进行线性回归，讲解真正意义上的机器学习操作。

使用 sklearn 进行线性回归

使用验证集法进行模型拟合

示例

参阅教材内容

使用 sklearn 进行线性回归

更换随机数种子，使用验证集法进行模型拟合

示例

参阅教材内容

使用 sklearn 进行线性回归

使用10折交叉验证法进行模型拟合

示例

参阅教材内容

使用 sklearn 进行线性回归

使用10折重复10次交叉验证法进行模型拟合

示例

参阅教材内容

使用 sklearn 进行线性回归

使用留一交叉验证法进行模型拟合

示例

参阅教材内容

PART 09

习　题

习题

1、使用数据4.3进行分析，数据是某商业银行相关经营数据（已经过处理，不涉及泄露商业秘密）。限于篇幅仅展示部分数据。其中code为客户编号，Profit contribution为利润贡献度，作为响应变量；Net interest income为净利息收入、Intermediate income为中间业务收入、Deposit and finance daily为日均存款加理财之和，均作为特征变量。请使用数据具体进行以下操作：

（1）对Profit contribution, Net interest income, Intermediate income, Deposit and finance daily进行描述性分析；

（2）对Profit contribution, Net interest income, Intermediate income, Deposit and finance daily使用Shapiro-Wilk test检验和kstest检验两种方法进行正态性检验；

习题

（3）对Profit contribution，Net interest income，Intermediate income，Deposit and finance daily四个变量进行皮尔逊相关分析并绘制热图、进行斯皮尔曼、肯德尔相关分析。

（4）以Profit contribution为响应变量，其他三个变量为特征变量，使用smf 进行线性回归并进行分析，在此基础上开展多重共线性检验。

（5）以Profit contribution为响应变量，其他三个变量为特征变量，使用sklearn 进行线性回归，包括使用验证集法进行模型拟合、更换随机数种子使用验证集法进行模型拟合、使用10折交叉验证法进行模型拟合、使用10折重复10次交叉验证法进行模型拟合、使用留一交叉验证法进行模型拟合等。

感谢聆听

THANKS