

数据挖掘-线性回归

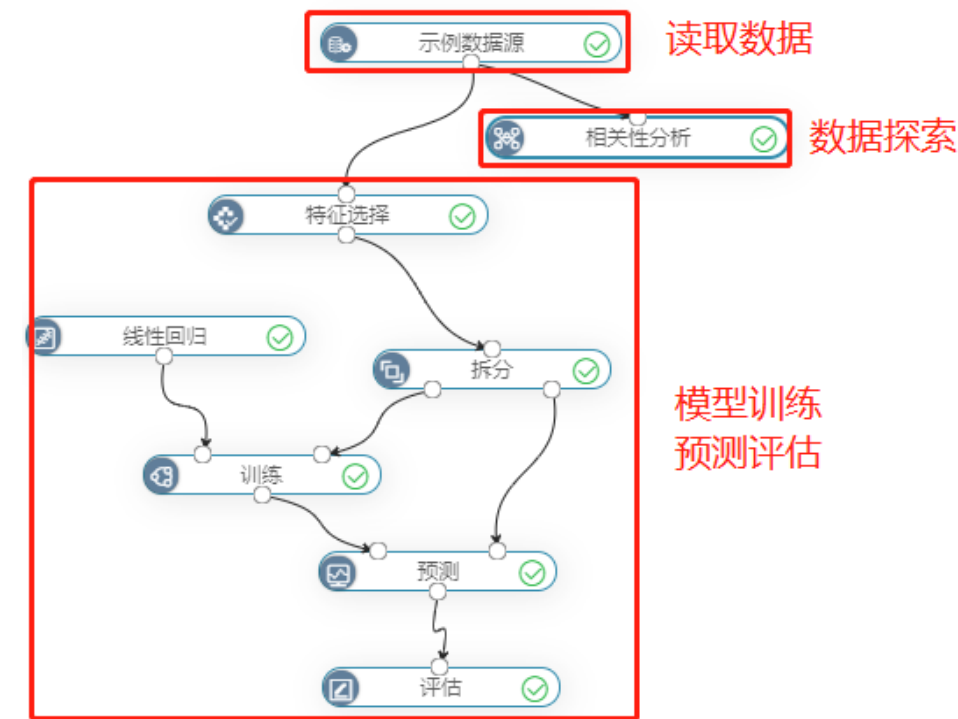
概述

一种常用的回归方法，它是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计方法，通过凸优化的方法进行求解，以达到预测评估的效果。

- 概述
- 示例

示例

使用“波士顿房价预测”案例数据，预测波士顿房价。



其中，相关性分析是为了分析特征变量与目标变量的相关性系数，方便特征选择进入模型训练。

线性回归的参数如下：

参数名称	值	说明
归一化	正则化	详情请参考 归一化 介绍说明。
	标准化	
	最小最大值归一化	
	最大绝对值归一化	
最大迭代数	参数范围为：>=0的整数，默认值为10	算法的最大迭代次数，达到最大迭代次数即退出。 最大迭代次数的值越大，模型训练更充分，但会耗费更多时间。
混合参数	参数范围为：[0,1]的数，默认值为0	控制惩罚类型，平方误差损失函数中的 ρ ，参数范围为：[0,1]的数。其中：0表示L2惩罚，1表示L1惩罚，0~1表示L1和L2惩罚的结合。 对模型系数惩罚（或称正则化）可减少模型过拟合。
正则参数	参数范围为：>=0的数，默认值为：0。	正则项系数，损失函数中的 λ 。 正则化可以解决模型训练中的过拟合现象； 正则项系数越大，模型越不会过拟合。

epsilon	参数范围为：>1的数。默认值为1.35	huber损失函数中的 δ ；调节损失函数，用于控制算法模型的健壮性； epsilon 越大，损失函数对异常点惩罚就越大，也就是对异常点越敏感；
收敛阈值	参数范围为：>=0的数，默认值为：0.000001。	收敛误差值。 收敛误差值，当损失函数取值优化到小于收敛阈值时停止迭代。
损失函数	squaredError	可待优化的损失函数，用于衡量模型的输出值和真实值之间的差距。
	huber	squaredError表示平方误差，huber表示平滑平均绝对误差。