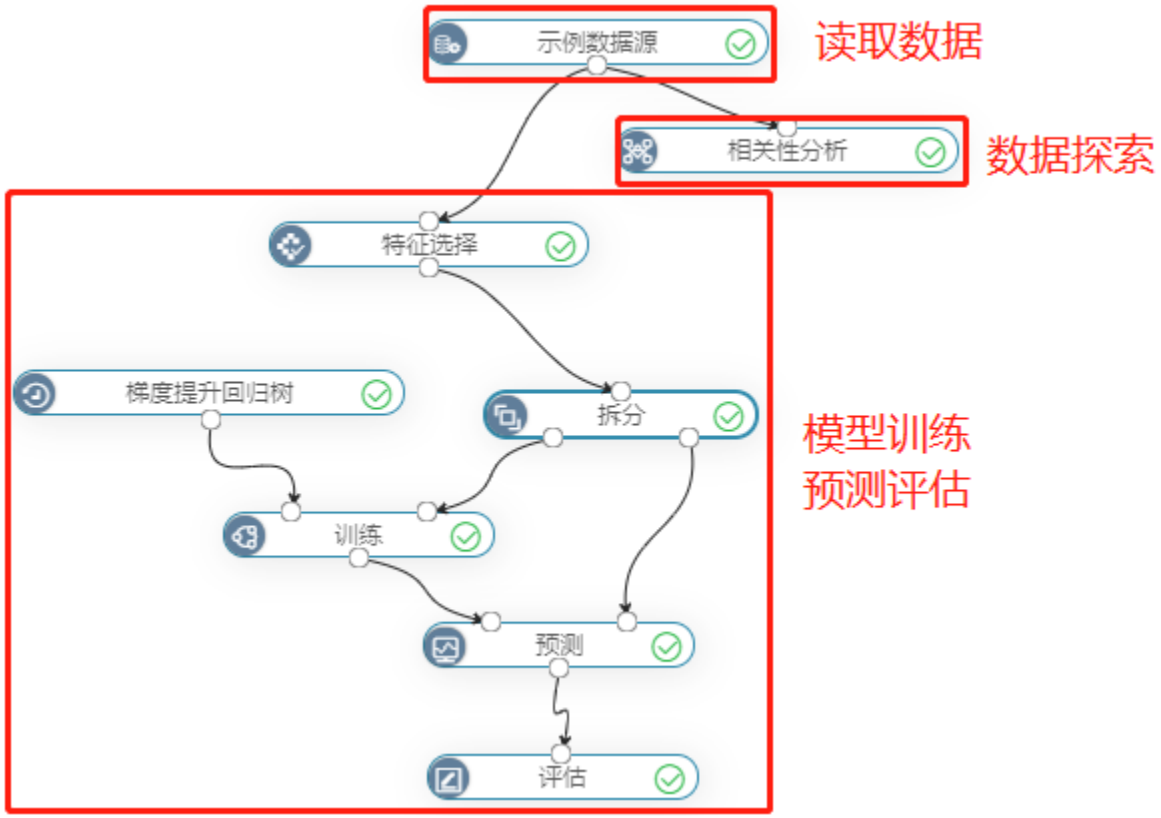
数据挖掘-梯度提升回归树

概述

梯度提升回归树是梯度提升树算法，原理是训练多棵回归树，每棵树建立是基于前一课树的残差，基函数为CART树，损失函数为平方损失函数的回归算法。

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。

参数设置

梯度提升回归树参数如下：

参数名称	值		说明
归一化	方法选择	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	
	参数	单位标准差归一化	勾选后，归一化后数据的标准差等于1
		平均数据中心化	勾选后，归一化后数据的均值等于0
自动调参设置	系统默认的各项参数值范围。		必须结合“启用自动调参”功能使用。系统将对设置指定或范围内的参数值循环调参，匹配出最优的组合。详情请参考 自动调参设置 。
启用自动调参	勾选该项，则系统自动调参数，不需要用户手工设置参数。		

分裂特征的数量	取值范围：>=2的整数； 默认值：32。	对连续类型特征进行离散时的分箱数； 该值越大，模型会计算更多连续型特征分裂点且会找到更好的分裂点，但同时也会增加模型的计算量；
树的深度	取值范围：[1,30]的整数；默认值为4。	当模型达到该深度时停止分裂； 树的深度越大，模型训练的准确度更高，但同时也会增加模型的计算量且会导致过拟合；
最大迭代数	取值范围：大于等于10且小于500的整数	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。
学习率	参数范围：[0.00000001, 1]，默认为0.01	每次迭代的参数学习步长倍率
子采样比例	取值范围：[0.1, 1]	训练每棵树时使用的训练数据的比例
衡量准则	gini entropy	裂分标准，Entropy表示熵值，Gini表示基尼指数；
子节点最少样本数	取值范围：大于0且小于等于1000的整数	每次分裂后每个子节点必须拥有的样本数； 该值越大，决策树允许分裂的次数就越少。可以防止模型过拟合；
最小分裂信息增益	取值范围：[0, 10000]	每次分裂必须达到的信息增益； 该值越大，决策树允许分裂的次数就越少。可以防止模型过拟合；

自动调参设置

系统将对设置指定或范围内的参数值循环调参，匹配出最优的组合。

自动调参的方式分为两种：

- 指定值调参：指定一个固定的值进行自动调参。
- 范围调参：在指定的范围内进行自动调参。

设置项说明如下：

设置项	说明
拆分比例	将选择的数据拆分为两部分，一部分部分用于模型的评估，另一部分数据用于训练模型。
评估标准	用于选择数据的评估指标，包括：mae、mse、rmse。

参数		自动调参的参数项。 计算信息增益的方式：目前只支持variance。 注意：参数默认的范围提供了一个推荐值，并不是算法限制的最大值和最小值。
指定值调参	指定值	指定一个固定的值进行自动调参。需要先勾选“是否使用指定值”才能配置。
	是否使用指定值	控制是否使用使用指定值进行调参。
范围调参	范围	设置自动调参参数的范围。 若运行速度比较慢，可将参数范围调小一点。
	步数	进行范围调参时，在设置的范围内生成多少个参数值。 示例： 1）范围为[3,5]，步数为3时，生成的参数值：3，4，5 2）范围为[40,100]，步数为4时，生成的参数值：40，60，80，100

示例

设置自动调参设置如图：

在训练节点查看分析结果如图：


查看分析结果
模型参数
超参数名称
超参数值
分裂特征的数量
24
树的深度
4
最大迭代数
50
学习率
0.1
子采样比例
1.0
衡量准则
variance
子节点最少样本数
1
最小分裂信息增益
0.0