

数据挖掘-GBDT特征选择

概述

特征选择是为算法服务的，选择不同的特征会直接影响到模型的效果。GBDT特征选择，就是使用GBDT算法，来自动选择相关性高的特征。

- 概述
- 输入/输出
- 参数设置
- 示例

输入/输出

输入	一个输入端口，用于接收前置节点传下来的数据集。
输出	一个输出端口，用于输出增加了离散后的字段的数据集。

参数设置

参数名称	说明	备注
选择特征列	选择需要的特征列，必须是数值列	必填（特征列中不能含有null）
选择标签列	选择做为标签列的字段	必填
需选择的特征数量	从待选择的特征列中输出特征列的数量	默认值为1, 范围是[1, 已选择特征的数量]的整数
选择方法	分类（二分类）	
	回归	
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	参数范围为：>=0的整数，默认值为30	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。
子采样比例	取值范围：(0, 1]的数；默认为0. 8。	对样本进行不放回的采样比例，取值为(0, 1]，取值小于1，则不使用全部样本去构建GBDT的决策树，小于1的比例，可减少方差，防止模型过拟合，默认是1，即不使用子采样。
子节点最少样本数	取值范围：>=1的整数，默认值为空	这个值用来限制叶子节点最少的样本数，如果某叶子节点数目小于样本数，则会和兄弟节点一起被剪枝。

示例

使用“居民用电数据”，预测是否漏电。特征选择3个特征和一个标签列，需选择的特征数量为2，选择方法为分类（二分类），其他参数默认。结果输出“featureSelector”列，表示为从3个特征列的值进行特征选择相关性较高的2个特征。如下图：

# 电量趋势下降指标	# 线损指标	# 告警类指标	# 是否窃漏电	As: featureSelector
4	1	1	1	[4.0,1.0,1.0]
4	0	4	1	[4.0,0.0,4.0]
2	1	1	1	[2.0,1.0,1.0]
9	0	0	0	[9.0,0.0,0.0]
3	1	0	0	[3.0,1.0,0.0]
2	0	0	0	[2.0,0.0,0.0]
5	0	2	1	[5.0,0.0,2.0]
3	1	3	1	[3.0,1.0,3.0]
3	0	0	0	[3.0,0.0,0.0]
4	1	0	0	[4.0,1.0,0.0]
10	1	2	1	[10.0,1.0,2.0]

注意：表头中表示特征列，*表示标签列

表头真名 表头别名

点击鼠标右键查看分析结果：
从3个特征列中选取2个相关性最高的特征进行展示。如下图：

查看分析结果

×

特征名称	重要程度
告警类指标	0.513137860182956
电量趋势下降指标	0.3472523569282953

