

数据挖掘-PCA(主成分分析)

概述

PCA（主成分分析）是统计学上一种常用的方法，主要用来对高维数据进行降维，通过对多个维度进行线性组合，获得较少且能够描述数据特征的主成分指标，减少由于数据维度过多带来的庞大计算量，降低算法的复杂度，使用最少数量的主成分来解释大量的方差。因为它可减少变量数目以此避免多重共线性，适用于预测变量较多大于观测值数目的情况。

- [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，用于输出接收到的数据集。

参数设置

参数名称	说明	备注
选择特征列	用于设置待选择的特征列	必填（特征列中不能含有null）
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	必填范围是[1, 已选择特征的数量]的整数，默认值为1

示例

使用“鸢尾花数据”，选择4个特征列和设置需选择的特征数量为2，则输出的结果为2个主成分指标。结果如下图：

⌚ 当前显示 100 条 / 总共有 150 条数据 提示: 点击单元格可查看超出的内容

Sepal_Width	# Petal_Length	# Petal_Width	Species	# PCA_0	# PCA_1
3.0	4.4	1.4	versicolor	-6.402539505781175	-5.655457050662012
2.8	4.8	1.4	versicolor	-6.834389569331097	-5.571393452844031
3.0	5.0	1.7	versicolor	-7.060167287675079	-5.594448024137944
2.9	4.5	1.5	versicolor	-6.315655782426549	-5.163602276140076
2.6	3.5	1.0	versicolor	-5.196781354584699	-4.9586903870731005
2.4	3.8	1.1	versicolor	-5.434238640540644	-4.621780445028225
2.4	3.7	1.0	versicolor	-5.312742660230506	-4.646665813299557
2.7	3.9	1.2	versicolor	-5.638793844167023	-5.012920138578452
2.7	5.1	1.6	versicolor	-6.882391568524519	-4.905998289513812
3.0	4.5	1.5	versicolor	-6.090371575948871	-4.842665156846481
3.4	4.5	1.6	versicolor	-6.30922344510942	-5.5211348915408465

注意：表头中📌表示特征列，*表示标签列

表头真名 ☒ 表头别名