

# 数据挖掘-LDA



“LDA”节点在V9.7版本中新增

- 概述
- 参数设置
- 示例

## 概述

LDA是一种主题模型。它是一个三层贝叶斯概率模型，包含词、主题和文档三层结构。

它是一种非监督机器学习技术，可以识别出大规模文档集或语料库中的主题。

常用于做文本挖掘聚类分析。

## 参数设置

| 参数名称      | 说明                                                                                                                                                                                                                                                                                                                                                                    |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 自动调整最优主题数 | <div>1、评估标准：LDA主题模型的评估方法。<ul style="list-style-type: none"><li>困惑度：用来度量一个概率分布或概率模型预测样本的好坏程度。困惑度越低，说明聚类的效果越好。</li></ul></div> <div>2、主题数范围：主题个数的范围，1~100的整数。</div> <div><div>自动调整最优主题数</div><div><div>评估标准</div><div>困惑度</div></div><div><div>主题数范围</div><div>2</div><div>-</div><div>5</div><div>(请输入1到100的整数)</div></div><div><div>确定</div><div>取消</div></div></div> |
| 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。                                                                                                                                                                                                                                                                                                                                            |
| 主题数目      | 主题数，或者聚类中心数。默认值为2。                                                                                                                                                                                                                                                                                                                                                    |
| 迭代次数      | 模型的迭代次数，达到该迭代次数即退出。默认值为10                                                                                                                                                                                                                                                                                                                                             |
| 文档主题分布    | 文章分布的超参数(Dirichlet分布的参数)，必需>=0，默认值为1。<br>值越大，推断出的分布越平滑                                                                                                                                                                                                                                                                                                                |
| 主题词分布     | 主题分布的超参数(Dirichlet分布的参数)，必需>=0，默认值为1。<br>值越大，推断出的分布越平滑                                                                                                                                                                                                                                                                                                                |

## 示例

使用文本数据，分析主题词分布情况以及各词的概率权重。

