机器学习业务实践之路

课程6：文本分析-新闻自动分类系统

阿里云 李博（傲海）

目录

文本分类算法思路

随着互联网的发展，网络上已经有大量文本数据的积累。传统的互联网媒体的分类以及打标等操作主要通过人肉处理的方式进行，效率低、成本高。目前正在通过文本算法大量的替代人肉的处理方式。

文本分析常见需求：

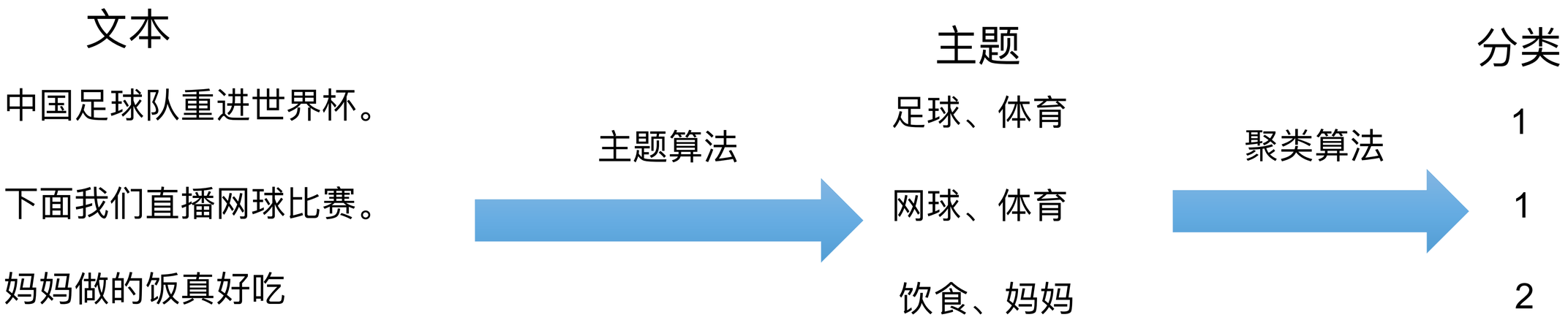
- 文本分类
- 文本打标签（长文本、短文本）
- 文本情感分析

文本分类常见方法：

- 基于语义向量距离：将文本按照语义映射成高维向量特征，通过向量距离进行分类
- 基于文本关键词、主题：首先提取文本的关键词、主题等信息，然后通过这些词语的对照关系进行分类

基于主题的文本分类

思路：具有相似主题的一类文本，属于相同的类别。



主题模型-LDA算法

$$p(词语|文档) = \sum_{主题} p(词语|主题) * p(主题|文档)$$

通过上面的这个概率密度公式，我们就可以抽象出文章产生的场景，下面来看下如何通过贝叶斯公式来生成文章的：

- 当我们准备开始写作的时候，我们先考虑这篇文章的主题是什么，有哪些，这个主题可以通过主题的概率分布来获得。
- 拿到了主题之后，当我们起笔开始写做的时候，就会从这个主题中的单词分布中选择一个词，这个词一定是要符合这个主题的概率分布。
- 循环遍历整篇文章，例如我们要写一篇500个词的作文，就把上面的1和2步骤遍历500遍，整篇文章就生成了。

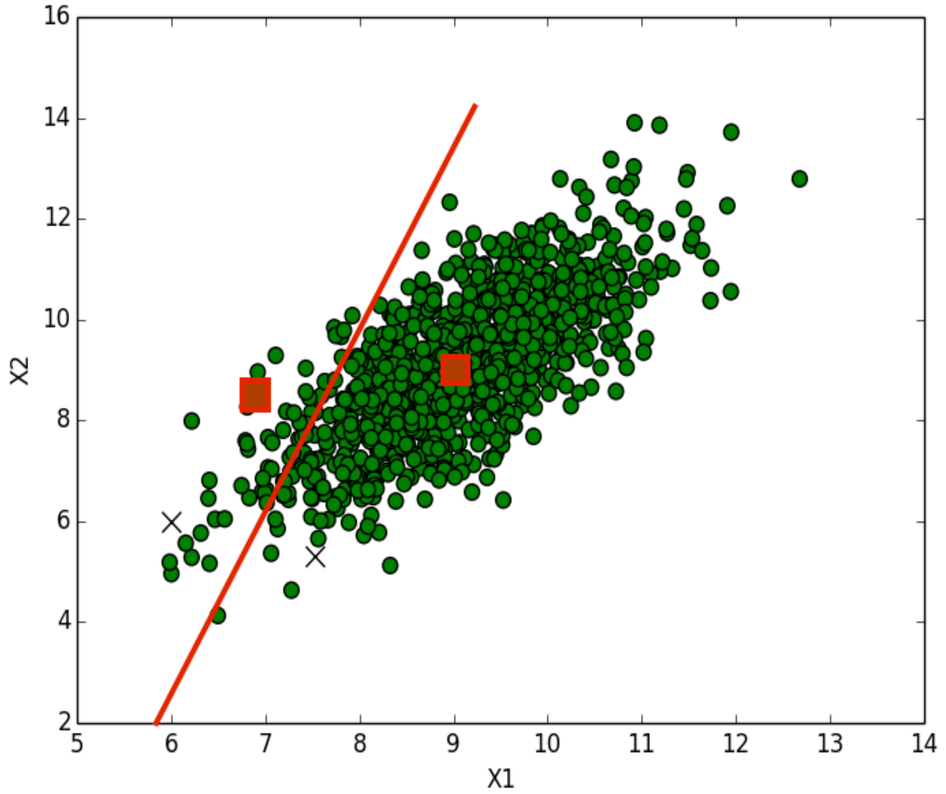
KMeans算法

第一步：设置分类K值

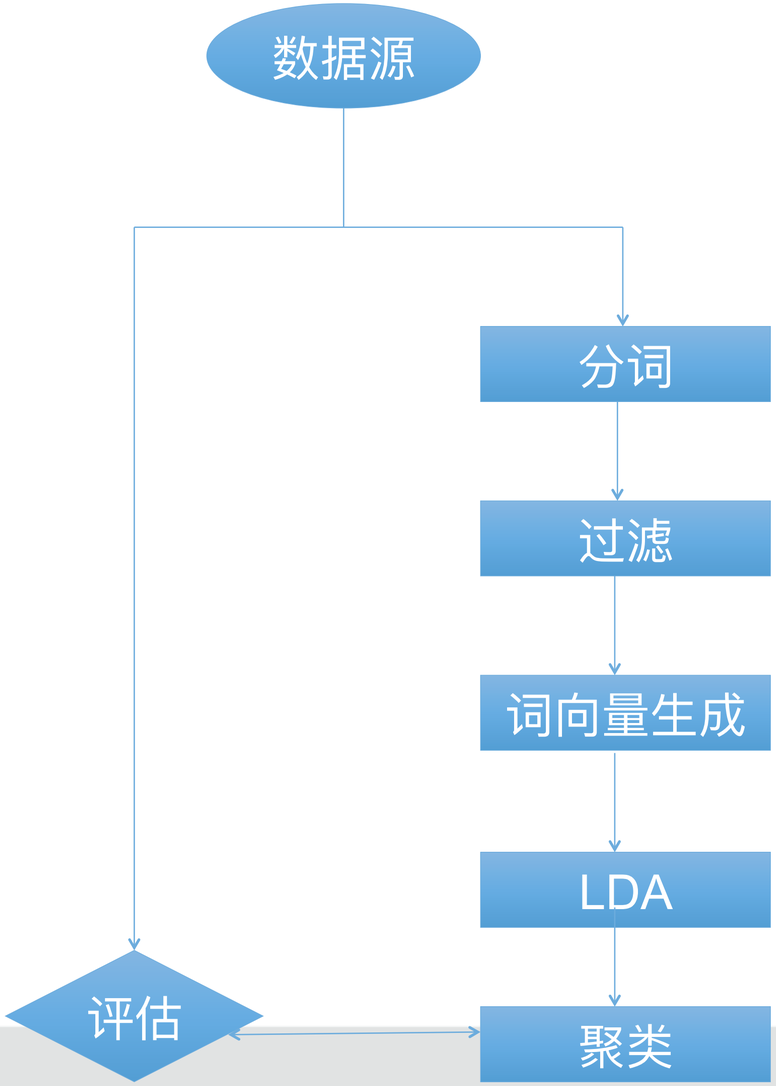
第二步：设置初始质心簇的位置

第三步：不断迭代寻找新分类簇的簇心点

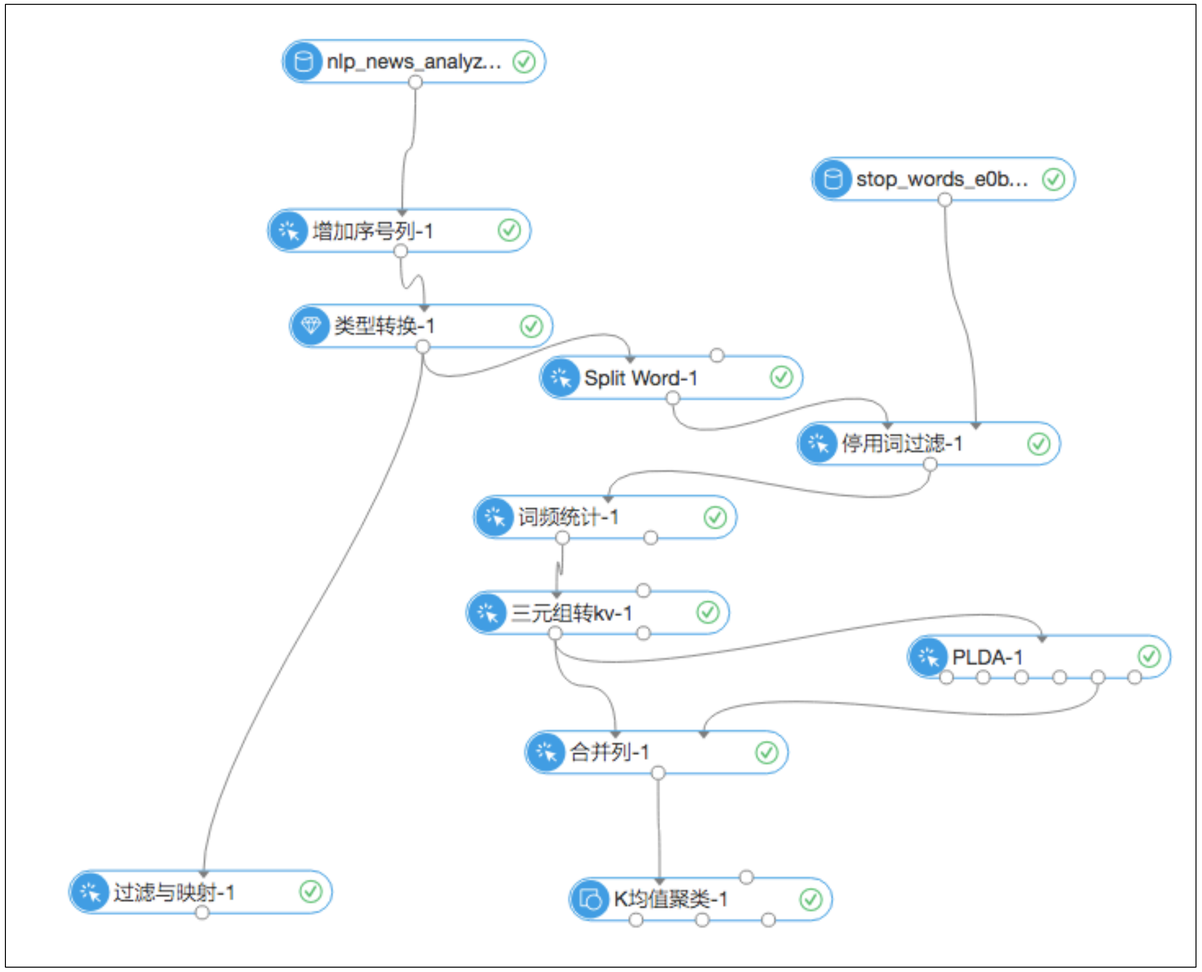
第四步：收敛完成聚类



PAI平台实现文本分类



PAI平台架构图



基于主题算法的新闻分类

详细介绍文档：

https://
yq.aliyun.com/
articles/59205

相关资料

推荐学习材料：

- 《机器学习实践》
- 《统计学习方法》
- 吴恩达的机器学习相关课程

推荐实验环境：机器学习PAI https://data.aliyun.com/product/learn

我的个人微信公众号（与我交流）：凡人机器学习

为了无法计算的价值 | 阿里云