形状, 圆圈

描述已自动生成

**卡通人物

中度可信度描述已自动生成**

**深度学习与自然语言处理第三次作业**

LDA主题模型进行文本分类

|  |  |
| --- | --- |
| 院（系）名称 | 自动化科学与电气工程学院 |
| 专业名称 | 电子信息 |
| 学生姓名 | 彭程 |
| 学号 | ZY2103807 |
| 指导老师 | 秦曾昌 |

2022年 5月

# 实验背景

## 问题描述

从给定的语料库中均匀抽取200个段落（每个段落大于500个词）， 每个段落的标签就是对应段落所属的小说。利用LDA模型对于文本建模，并把每个段落表示为主题分布后进行分类。验证与分析分类结果。

## 实验原理

**1.2.1** **LDA主题模型简介**

在文本挖掘领域中大量的数据都是非结构化的，难以从信息中直接获取相关和期望的信息。主题模型（Topic Model）能够识别在文档里的主题，并且挖掘语料里隐藏信息，在主题聚合、特征选择等场景有广泛的用途。

LDA（Linear Discriminant Analysis），是一种文档主题生成模型。它可以将文档中每篇文档的主题按照概率分布的形式给出。也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA的核心思想是寻找到最佳的投影方法，将高维的样本投影到特征空间(feature space)，使得不同类别间的数据“距离”最大，而同一类别内的数据“距离”最小。LDA采用了“词袋”的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是没有考虑词与词之间的顺序，这简化了问题的复杂性。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

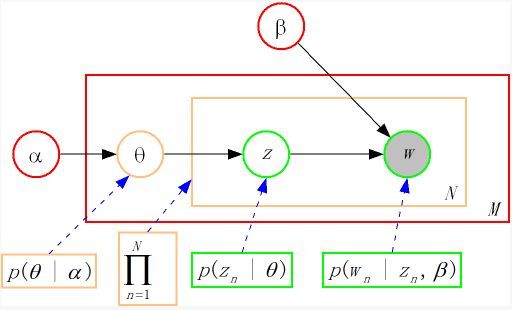
****

图1 LDA模型

**1.2.2 SVM分类器**

支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的的学习算法就是求解凸二次规划的最优化算法。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示， 即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。

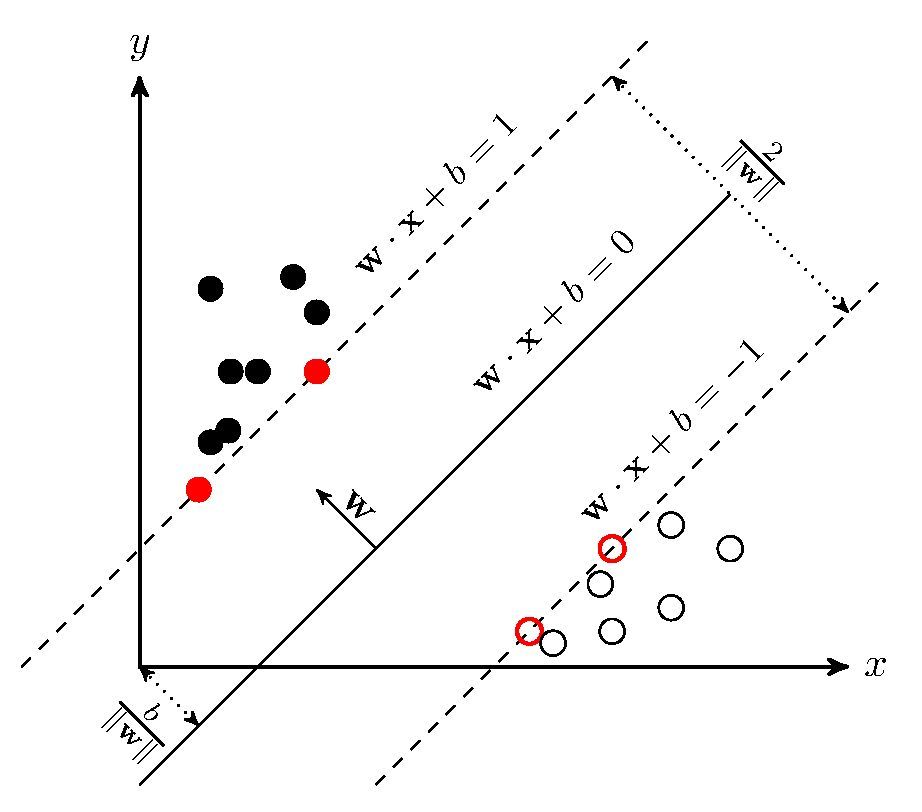


图2 SVM分离超平面示意图

# 实验方法

本实验采用的如下思路进行文本分类。

1. 生成文本段落数据。按照题目要求从给定的16本小说中均匀抽取200个段落，每个段落包含500个字，每段文本的标签为对应小说的小说名；
2. 生成的200个段落中，作为训练样本，剩余作为测试样本；
3. 利用训练样本训练LDA模型，其中主题数为*d*，每个训练样本对应的主题分布为一个的向量
4. 利用上述训练样本特征向量以及对应标签训练一个线性SVM分类器；并预测训练样本的标签，得到训练准确率；
5. 利用训练好的LDA模型得到测试样本的主题分布。计算测试样本文本分类准确率。

# 实验结果与分析

## 生成文本段落

按如下参数生成分类文段落：

|  |  |  |  |
| --- | --- | --- | --- |
| 参数 | 文本数 | 每段长度 | 段落数 |
| 数值 | 16 | 1000 | 200 |

表 1. 文本参数表

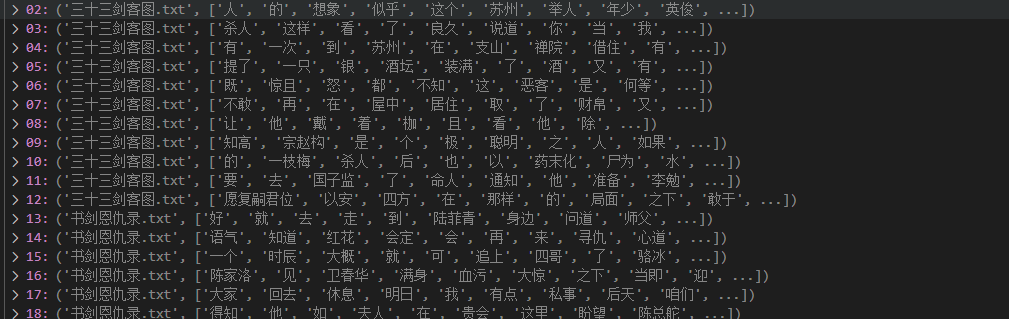


图1. 生成结果

## 实验结果

在实验中，考虑了不同划分的段落书、每个段落的段落字数以及不同主题数。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验序号 | 主题数 | 段落数 | 段落字数 | 训练集  准确率 | 测试集  准确率 |
| 1 | 16 | 200 | 500 | 0.256 | 0.075 |
| 2 | 16 | 200 | 1000 | 0.2062 | 0.100 |
| 3 | 50 | 200 | 500 | 0.2938 | 0.095 |
| 4 | 50 | 500 | 500 | 0.2750 | 0.090 |
| 5 | 100 | 200 | 500 | 0.3187 | 0.075 |
| 6 | 100 | 200 | 1000 | 0.2875 | 0.025 |

表 2 运行结果

## 结果分析

通过以上实验可以看出，以上三个变量都能决定训练集和与测试集的预测准确率。而在不同条件下，其影响不一定相同。通过1和2可以看出，增大段落字数能提高测试集准确率，而5和6则得出了相反的实验结果。

**附录：**

**实验代码见工程文件夹。**

DataProcess 对文本数据进行处理；

LDA进行模型训练和测试；

Main 为主函数。