LDA 主题模型

原理详解与代码实战

苏烨

统计与数学学院

2021-12-15

Section 1

写在前面

写在前面

在机器学习领域，关于 LDA 有两种含义，一是「线性判别分析（Linear Discriminant Analysis）」，是一种经典的降维学习方法；一是我们要讲的「隐含狄利克雷分布（Latent Dirichlet Allocation）」，是一种概率主题模型，主要用来文本分类，在 NLP 领域有重要应用。

LDA 主题模型，在 PLSA 模型的基础上引入了参数的先验分布的概念，能够对文本信息进行语义抽取，为各领域科研人员提供了文本主题挖掘的新途径。目前它已经被广泛应用于文本信息检索、话题检测和跟踪、线上消费者偏好特征研究、错误报告诊断等诸多领域，产生了许多研究成果。

Section 2

LDA 模型的代码实现

目前，我在网上找到的实现方式主要有以下 3 种 (都先用 jieba 将词分好)：

1. 使用 gensim(Python) 里的 doc2bow 向量化，然后用 ldamodel

(图1)

目前，我在网上找到的实现方式主要有以下 3 种 (都先用 jieba 将词分好)：

1. 使用 gensim(Python) 里的 doc2bow 向量化，然后用 ldamodel
2. 使用 sklearn(Python) 里的 CountVectorizer 向量化，然后用 LatentDirichletAllocation 建模

(图1)

目前，我在网上找到的实现方式主要有以下 3 种 (都先用 jieba 将词分好)：

1. 使用 gensim(Python) 里的 doc2bow 向量化，然后用 ldamodel
2. 使用 sklearn(Python) 里的 CountVectorizer 向量化，然后用 LatentDirichletAllocation 建模
3. R 语言 lda 包进行模型训练，使用 LDAvis 进行可视化。

(图1)

LDAvis 可视化 demo

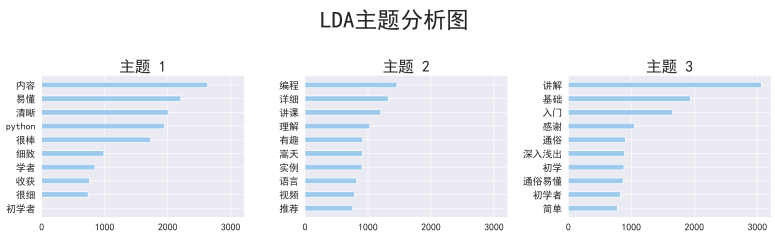


图 1: 作图样例

Section 3

LDA 的数学原理

Subsection 1

pLSA

pLSA

LDA 是一种典型的词袋模型，它的基本假设是一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

经典 LDA 主题模型的三层贝叶斯模型：

LDA 模型

图 2: Smoothed_LDA

LDA 模型

如上图所示，在 LDA 模型中一篇文档生成的方式如下：

- 从狄利克雷分布 α 中取样生成文档 i 的主题分布 θ_i
- 从主题的多项式分布 θ_i 中取样生成文档 i 第 j 个词 $w_{i,j}$ 的主题 $z_{i,j}$
- 从狄利克雷分布 β 中取样生成主题 $z_{i,j}$ 的词语分布 $\phi_{z_{i,j}}$
- 从词语的多项式分布 $\phi_{z_{i,j}}$ 中采样最终生成词语 $w_{i,j}$

LDA 主题建模的过程，概括来说就是通过给定的训练文本集学习出参数 α 和 β。而参数 α 和 β 的估计，可以由 EM 推断和 Gibbs 采样算法得到。

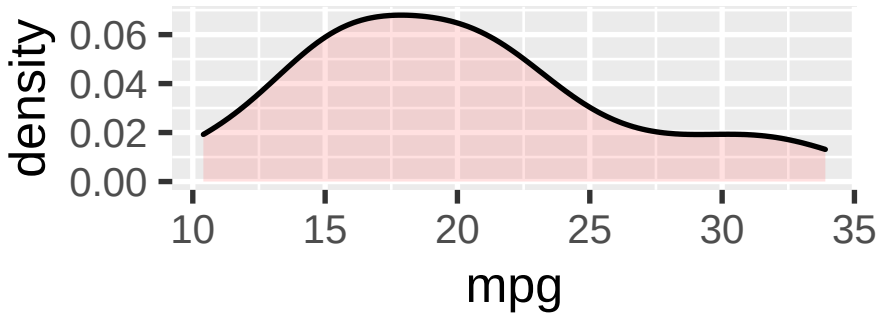
Subsection 2

对比

对比

My topic for this slide

head	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1



- Here is some Bullet Text
- And some more
  - Subtext
  - More Subtext