北风网
IBEIFENG.COM

人工智能之机器学习

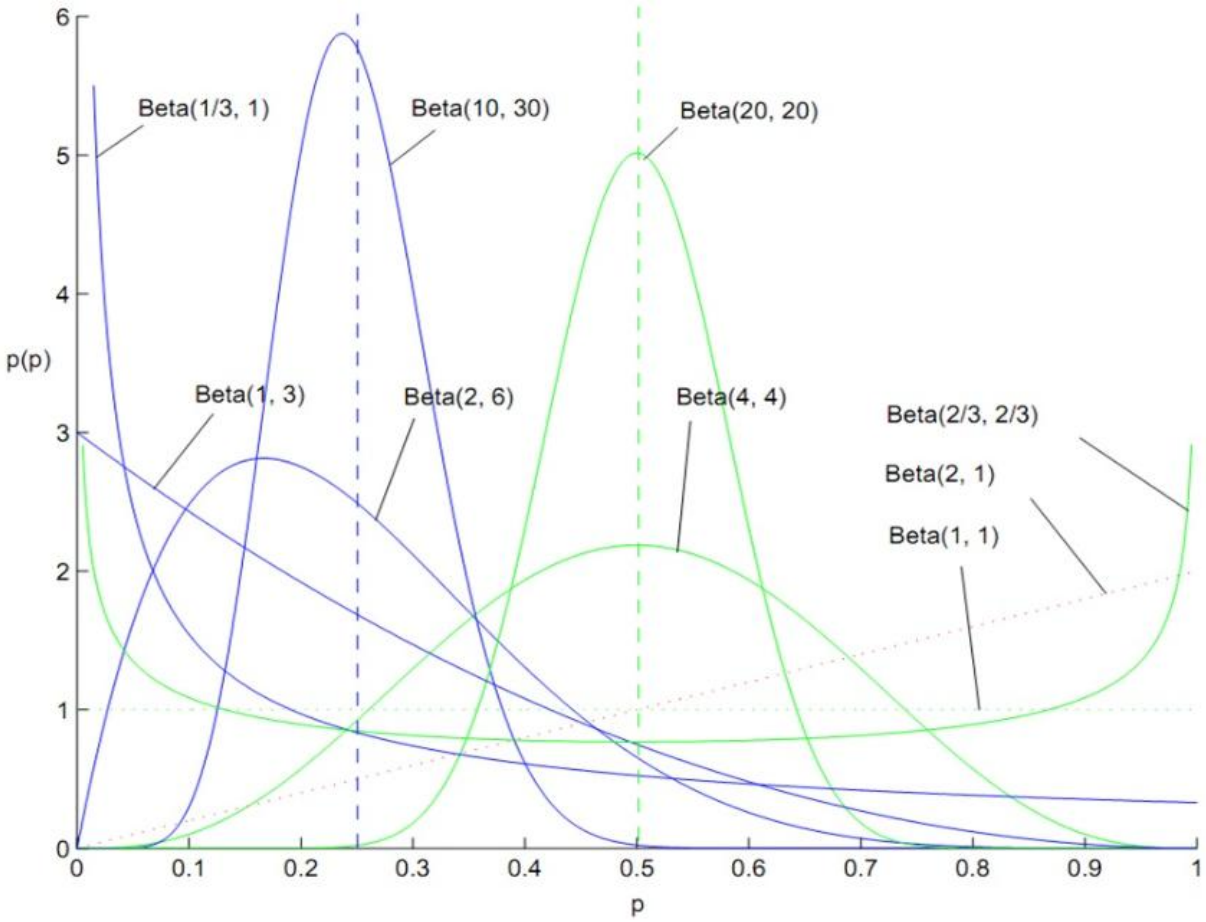
隐含狄利克雷分布
(Latent Dirichlet Allocation)

上海育创网络科技股份有限公司

主讲人：赵翌臣

LDA基础知识——beta分布

- Beta分布：
  - Beta(a,b)
  - 以抛硬币为例，抛了a+b次，参数a为出现了正面的次数，参数b为出现了反面的次数。
  - Beta分布是估计硬币正面向上概率的分布，横轴表示出现正面的概率，纵轴表示出现正面概率的可能性。



LDA基础知识——狄利克雷分布

- 狄利克雷分布：
  - 狄利克雷分布就是beta分布的推广情形。现在抛出的结果不是只有正和反两种了，可能有很多种，就像抛骰子可能有6种。

隐含狄利克雷分布（LDA）直观理解

- 作者
  - LDA是由吴恩达等人于2003年提出的一种主题模型，是无监督学习模型
  - 论文：Latent Dirichlet Allocation
- 应用
  - 文档分类、降维
- 思想
  - 模型构建
    - 训练数据：文档集，目标函数稍后谈
  - 模型使用
    - 文本分类：每个文档以不同的概率分别属于每一个主题，当有新文档待预测时，LDA可以算出这个文档属于每个主题的概率，下面举例子说明一下
    - 降维：主题是个抽象的概念，无法直接理解，但可以作为特征

隐含狄利克雷分布（LDA）直观理解

- 第一步准备训练数据：这里准备了两类文档（航空类、体育类），并带有类别y，但是训练LDA模型时不用y，y只是用来评估LDA文档分类的效果

隐含狄利克雷分布（LDA）直观理解

- 第二步：将文本转为词频，注意这里是TF，不是TF-IDF
- 第三步：训练LDA模型

```
from sklearn.decomposition import LatentDirichletAllocation
lda = LatentDirichletAllocation(n_components =2,
                                random_state=0)
```

  - 注：这里指定了2个主题，用户指定个数也是可以的，这里指定为2的目的是，因为已知是两类文档，我想对照y看一下LDA的效果。
- 效果评估
  - LDA给出的主题概率分布如图

```
[[9.99417172e-01 5.82828207e-04]
 [9.48272130e-01 5.17278701e-02]
 [8.36948741e-01 1.63051259e-01]
 ...
 [1.73361605e-03 9.98266384e-01]
 [2.72210690e-04 9.99727789e-01]
 [3.21560981e-04 9.99678439e-01]]
```

第二篇文档属于第一个主题的概率

隐含狄利克雷分布（LDA）直观理解

- 使用y评估后，准确率为0.971969696969696969
- 打印主题　print (lda.components_)

```
[[ 6.04629107   0.50931329 11.12517948 ...   2.17450839   0.50101766
   0.50485851]
 [24.41353582   1.52722691   4.06877379 ...   0.50095316   1.57205007
   2.64094396]]
```

- 6.04629107 代表第一个主题的第一个词的非规范化概率，0.50931329代表第一个主题的第二个词的非规范化概率，以此类推。

隐含狄利克雷分布（LDA）直观理解

- 打印每个主题下概率Top50的词语
- 主题1：分布,确定,使用,得到,具有,提高,工艺,故障,信号,要求,条件,状态,问题,不同,应用,实验,通过,性能,本文,误差,材料,工作,叶片,变化,航空,时间,由于,速度,温度,过程,参数,数据,发动机,试验,影响,研究,结果,可以,结构,模型,采用,控制,技术,设计,分析,测量,计算,进行,方法,系统——航空主题
- 主题2：水平,方面,目标,训练,过程,国家,作用,自己,方法,可以,基础,心理,需要,运动,重要,理论,要求,建设,他们,经济,我国,能力,知识,内容,中国,学科,管理,培养,科学,改革,提高,思想,进行,问题,文化,学习,研究,我们,工作,素质,课程,教师,活动,学校,社会,教学,学生,发展,体育,教育——体育主题

LDA介绍

- 基础概念
  - 词（word）：待处理数据的基本离散单元，如“hello”、“你好”
  - 文档（document）：是待处理的数据对象，由一组词组成，这些词在文档中不计顺序，例如一篇论文，这样的表示方式为“词袋”（bag-of-words）
  - 主题/话题（topic）：表示一个概念，具体表示为一系列相关的词，以及它们在该概念下出现的概率。

LDA介绍

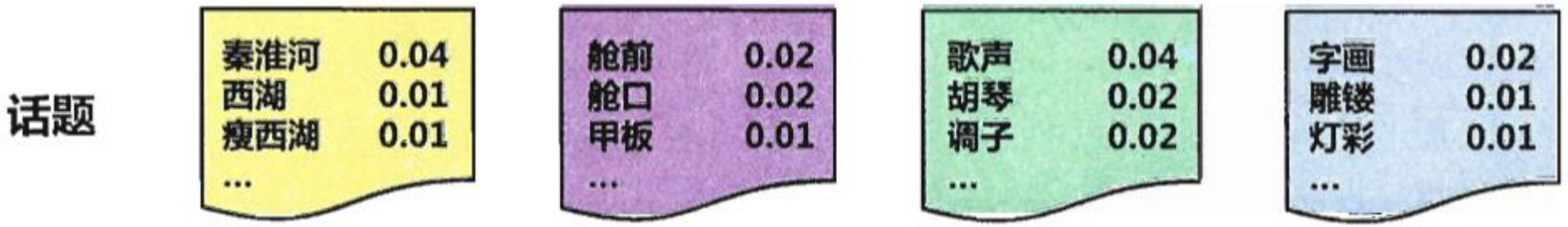
- LDA构建目标函数的思想
  - 训练数据（不携带类别信息）W={w_1，w_2，…，w_T}
  - LDA模型认为你这些数据是由我生成的，我有我的参数α和η，我可以量化出生成这样数据的概率

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{t=1}^{T} \ln p(\boldsymbol{w}_t \mid \boldsymbol{\alpha}, \boldsymbol{\eta})$$

  - 使用极大似然估计法，可以得到最优的α和η。

LDA介绍

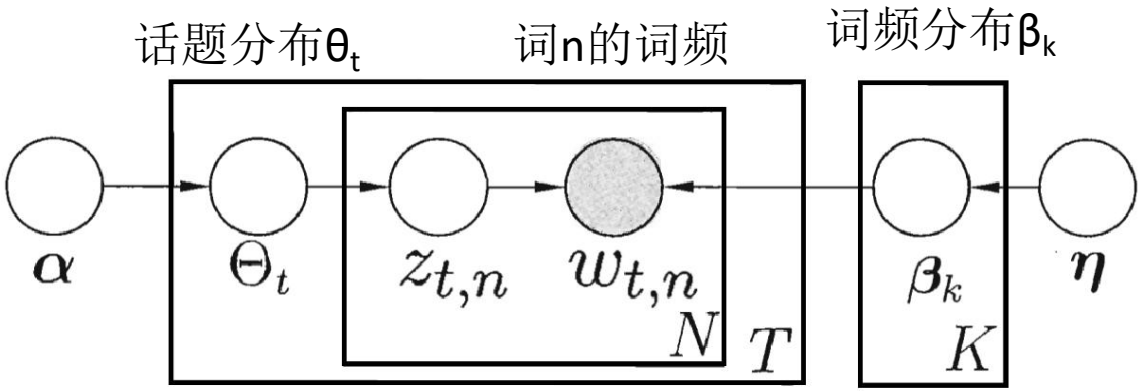
- 一个话题就像一个箱子，里面装着在这个概念下词的出现频率。



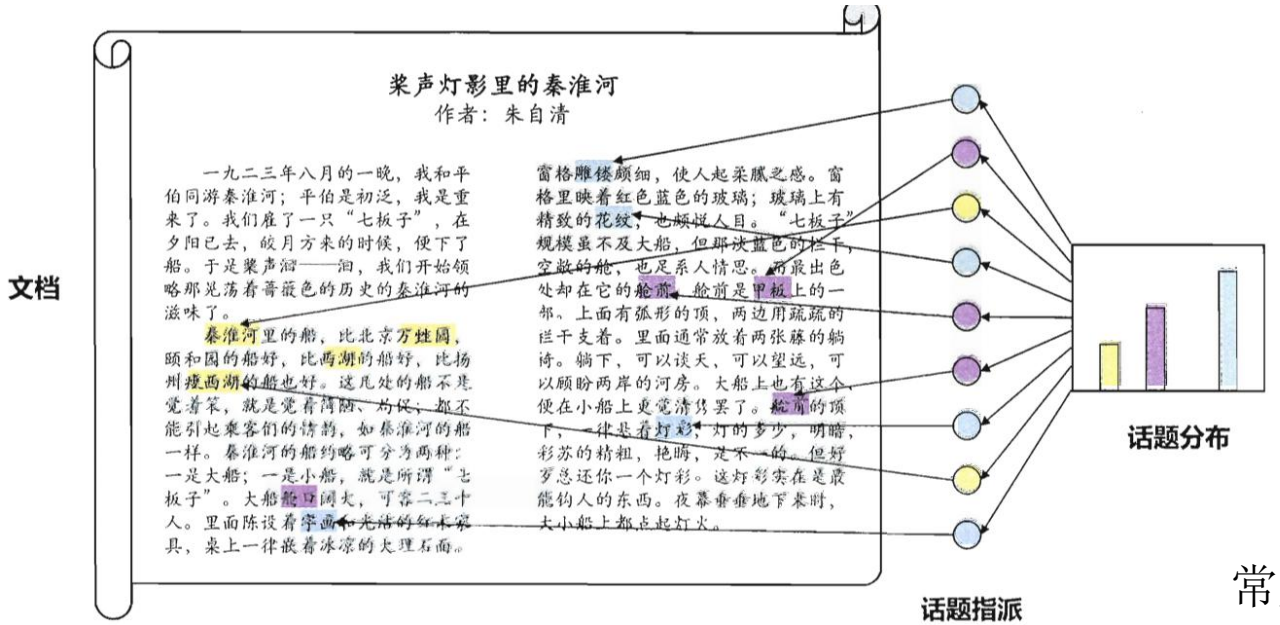
- 不妨假定数据集中一共包含K个话题和T篇文档，文档中的词来自一个包含N个词的词典。我们用T个N维向量$W=\{w_1, w_2, \ldots, w_T\}$表示数据集（即文档集合），K个N维向量$\beta_k$（k=1,2,...,K）表示话题。
- 其中$w_t \in R^N$的第n个分量$w_{t,n}$表示文档t中词n的词频
- 其中$\beta_t \in R^N$的第n个分量$\beta_{t,n}$表示话题k中词n的词频

LDA介绍

- LDA认为每篇文档有多个话题，用向量$\theta_t \in R^K$表示文档t中所包含的每个话题的比例，$\theta_{t,k}$表示文档t中包含话题k的比例，进而通过下面的步骤由话题生成文档t：
- （1）根据参数α得到一个话题分布θ_t~Dirichlet(α)；
- （2）按照如下步骤生成文档中的N个词：
  - （a）根据θ_t进行话题指派，得到文档t中的词n的话题$z_{t,n}$；
  - （b）根据指派的话题所对应的词频分布β_k随机采样生成词



话题$z_{t,n}$



$$p(\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \Theta \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{t=1}^{T} p(\Theta_t \mid \boldsymbol{\alpha}) \prod_{i=1}^{K} p(\boldsymbol{\beta}_k \mid \boldsymbol{\eta}) \left(\prod_{n=1}^{N} P(w_{t,n} \mid z_{t,n}, \boldsymbol{\beta}_k) P(z_{t,n} \mid \Theta_t) \right)$$

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{t=1}^{T} \ln p(\boldsymbol{w}_t \mid \boldsymbol{\alpha}, \boldsymbol{\eta})$$

常采用吉布斯采样或变分法进行近似推断


北风网
IBEIFENG.COM