**1、前言**

LDA是文本集合的生成概率模型。假设每个文本由话题的一个多项分布表示，每个话题由单词的一个多项分布表示，特别假设文本的话题分布的先验分布是狄利克雷分布，话题的单词分布的先验分布也是狄利克雷分布。LDA能更好解决过拟合现象。

LDA模型是含有隐变量的概率图模型。

**2、狄利克雷分布**

**①多项分布**

多项分布是一种多元离散随机变量的概率分布。

**②狄利克雷分布**

狄利克雷分布是一种多元连续随机变量的概率分布，是贝塔分布的扩展。在贝叶斯学习中，狄利克雷分布常作为多项分布的先验分布使用。

**③二项分布和贝塔分布**

二项分布是多项分布的特殊情况，贝塔分布是狄利克雷分布的特殊情况。

**共轭先验**

狄利克雷分布属于指数分布族

狄利克雷分布是多项分布的共轭先验。

贝叶斯学习中尝试用共轭分布。如果后验分布与先验分布属于同类，则先验分布与后验分布称为共轭分布。

**3、潜在迪利克雷分配模型**

LDA是文本集合的生成概率模型。模型假设话题由单词的多项分布表示，文本由话题的多项分布表示，单词分布和话题分布的先验分布都是狄利克雷分布。

LDA模型表示文本集合的自动生成过程，首先，基于单词分布的先验分布生成多个单词分布，及决定多个话题内容，之后，基于话题分布的先验分布生成多个话题分布，即决定多个文本内容；基于话题分布生成话题序列，基于单词分布生成单词。文本的单词序列是观测变量，文本的话题序列是隐变量，文本的话题分布和话题的单词分布也是隐变量。

利用LDA进行话题分析，就是对给定的文本集合，学习到每个文本的话题分布，以及每个话题的单词分布。

**LDA和PLSA**

相同点是两者都假设话题是单词的多项分布，文本时话题的多项分布。

不同点是LDA使用狄利克雷分布作为先验分布，而PLSA不适用先验分布（或者假设先验分布是均匀分布）两者对文本生成过程有不同假设；学习过程LDA基于贝叶斯学习，PLSA基于极大似然估计。LDA优点是，使用先验概率分布，可以防止学习过程中产生的过拟合。

**4、算法**

**4.1LDA的吉布斯抽样算法**

优点是实现简单，缺点是迭代次数较多。

即要对联合概率分布p进行估计，其中第一项为观测变量，其他都为隐变量。

为了估计多元随机变量x的联合分布p(x)，吉布斯抽样法选择x的一个分量，固定其他分量，按照其条件概率分布进行随机抽样，依次循环对每一个分量执行这个操作，得到联合分布p(x)的一个随机样本，重复这个过程，在燃烧期之后，得到联合概率分布p(x)的样本集合。

LDA模型的学习通常采用**收缩的吉布斯抽样方法。**

**4.2LDA的变分EM算法**

MCMC通过随机抽样的方法近似的计算模型的后验概率，变分推理通过解析的方法计算模型的后验概率的近似值。