1. **模型推导**



η是一个二维矩阵,行是标记,列是主题,第i是表示标记i在不同主题上的分布，是文档d在各个主题上的分布，是文档d中单词的主题赋值，和超参和无监督主题模型中的相同。

参数估计和预测的重点在于计算参数η，，的后验，

在Svn上, 最新版本在host\_tag\trunk\htmb3下。

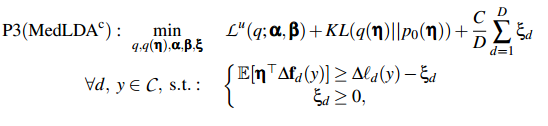
### 入口

shdir/ host\_label.sh

### 原理

一.模型定义

MedLDA 分类模型是解决下面的二次规划问题来获得模型参数。



(在后面的求解中，，因为 服从正态分布，均值为**** ，所以 ****)

其中代表 。



其中是一个特征向量，元素从到和相同，其它为0。



最优化方法



直接最优化上面的模型是不可解的，令是的近似，上界表示为：



其中，



最小化上界：

分两步进行，每一步都使目标函数减小。

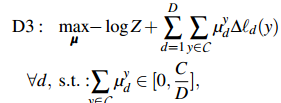
1. 固定，求解和

约束中没有，所以更新过程和无监督LDA及sLDA相同。



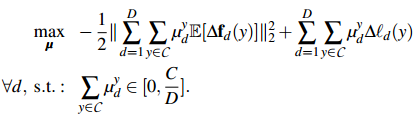
更新，最优解具有下面的形式：





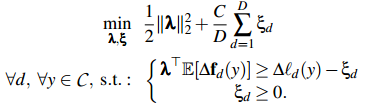
选择normal先验







上式是对偶形式，原始形式是：





1. 固定和，求解





与 是同一个意思 ，这里的)



二．程序实现具体的流程如下图所示：

因为stanford 分词模块，stanford词性标注模块，MedLDA预测分类模块的模型数据较大，每分类少数几个host就记载一次模型，会浪费很多的时间。所以将三个模块做成了三个服务，通过TCP socket 调用，

echo temp/conv\_train\_data.txt |php stanford\_sen.php

Stanford 分词模块，调用 localhost:15001的服务

php pos\_tag.php

Stanford 词性标注模块，调用 localhost:15000的服务

./medlda inf 28 temp120\_\*

Medlda host分类模块，调用localhost:10001的服务



三 模型代码实现过程：

* 训练样本的生成:

格式: line .=. tag host

例如：

交友 www.751314.com

交友 [www.luobohua.com](http://www.luobohua.com)

团购 fushun.lashou.com

团购 tg.fuling.com

团购 tuan.bangzhufu.com

… …

训练样本生成是在 host\_tag\_medlda\_bootstrap 文件夹中进行：

cat ana\_data.txt |php cateMain.php

利用cat bootdir/moutput.txt |php dictGen.php >> dictdir/dict\_new.txt 生成词典

再利用echo dictdir/dict.txt temp/moutput.txt |php testSamGen.php >> temp/nntrainSample.txt生成样本

生成的中间文件 temp/nntrainSample.txt

再调用./tag\_host\_train\_sam\_gen.sh 生成训练样本

svm训练样本结果：

temp/ntrainSample.txt

temp/hosts.txt

转换为medlda训练样本：

cat temp/ntrainSample.txt |php svm\_to\_med.php >> temp/mtrainSample.txt

* medlda 模型的生成:

此工作在medlda文件夹下进行

cd data

cp ../../ host\_tag\_medlda\_bootstrap/temp/mtrainSample.txt .

cp ../../ host\_tag\_medlda\_bootstrap/temp/hosts.txt .

mv mtrainSample.txt mmtrainSample.txt

程序运行前data文件夹下应该有下面三个文件

hosts.txt mmtrainSample.txt m\_tag\_index.txt

运行 ./auto\_train.sh

脚本里面参数设置：类别个数一般设置为 m\_tag\_index.txt 不同标记个数 +2,如果有26个分类则设置为28 (27 亦可)

./medlda est 120 28 10 16 64 temp random >> log.txt

./medlda inf 28 temp120\_\*里类别作同样的设置

注：在用该模型进行实际的分类时，在htmb 文件夹下的 auto\_train.sh设置为相同的值。

生成模型参数存储在类似temp120\_c10\_f519149632的文件下

* 使用模型进行实际的分类:

此工作在htmb文件夹下进行

模型的拷贝

cp -r ../medlda/ temp120\_c10\_f519149632 .

cat ana\_data.txt |php cateMain.php

或者

./crawl\_and\_cate.sh crawl\_src/test\_url.txt

从分光点数据到最终分类结果:

./flow\_cate.sh

生成统计结果：

cat dictdir/m\_tag\_index.txt scnode/1\_ChangChun\_CNC\_S1\_2013041911.txt scnode/sc\_label.txt |php flowStatis.php >> 结果文件

