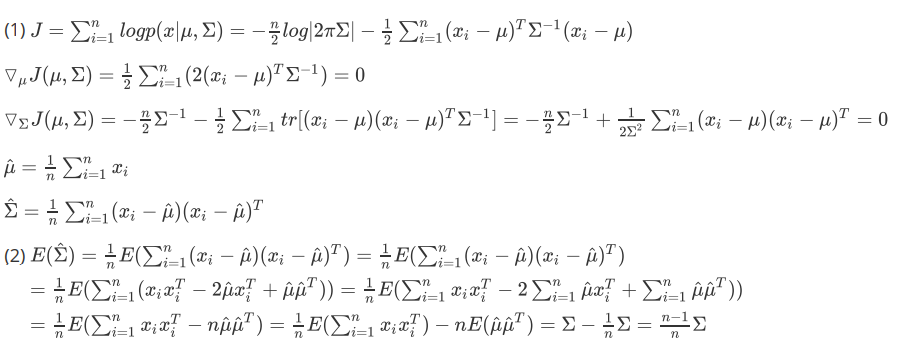
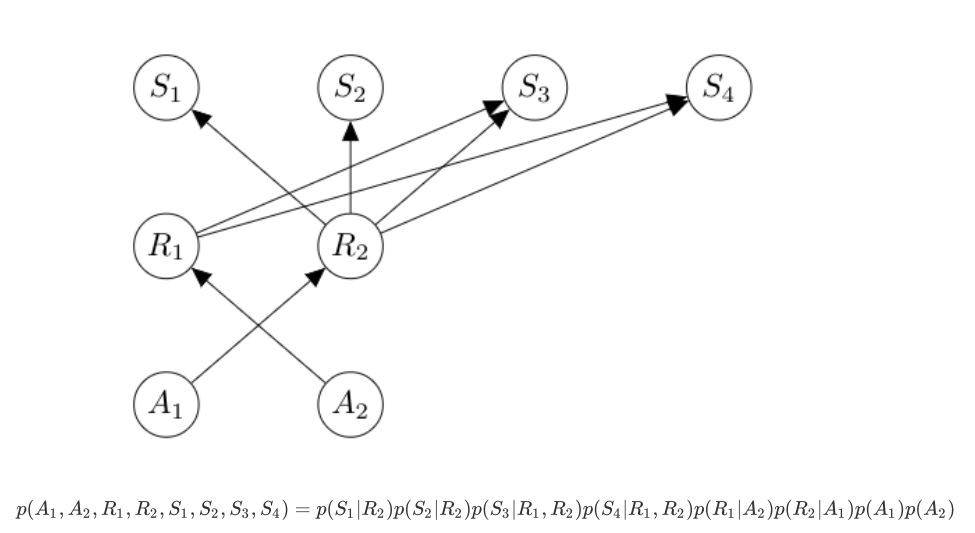
人工智能导论第三次作业

2018080111 覃逸成

第一题



第二题

1.

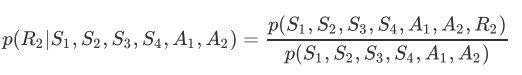
2.{R1,R2}

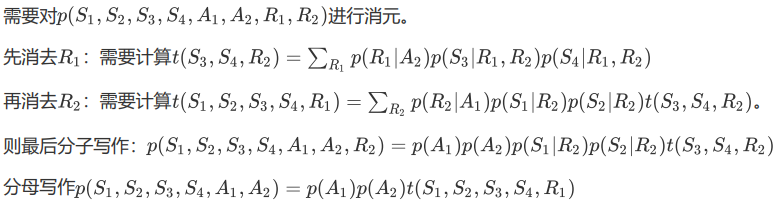
3.需要18个独⽴参数 需要255个独⽴参数

4.患新冠肺炎R2的概率会变化

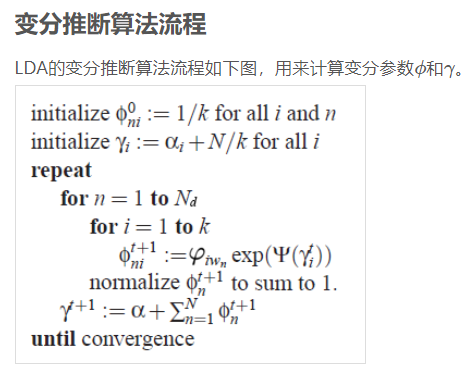
吸烟经历会影响患有普通肺炎R1、新冠肺炎R2的概率

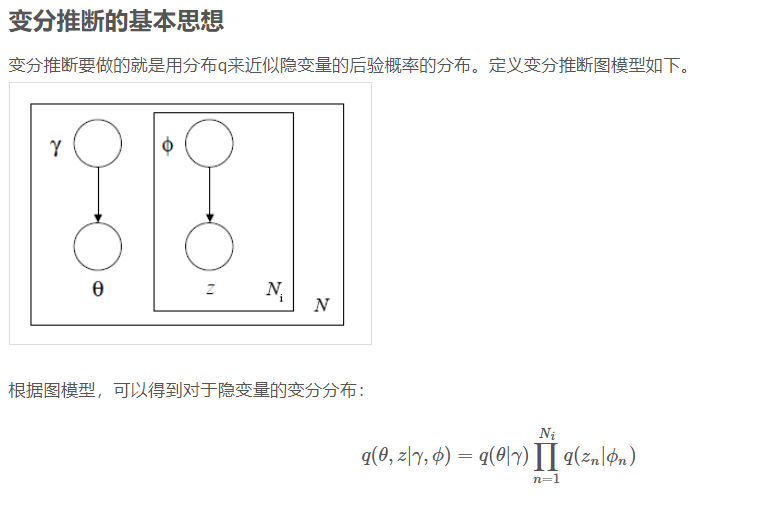
5.

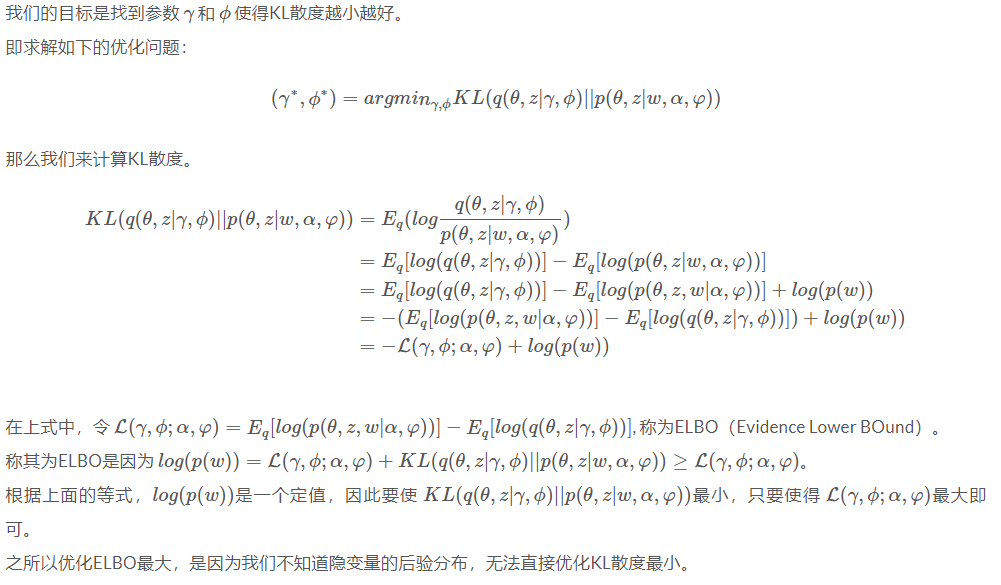


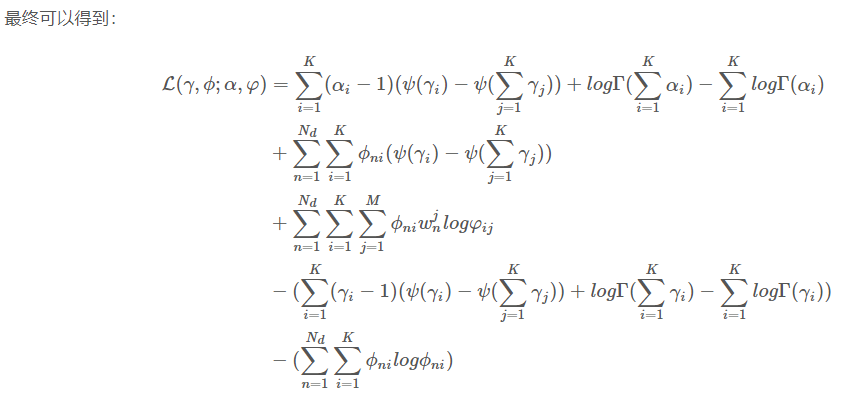


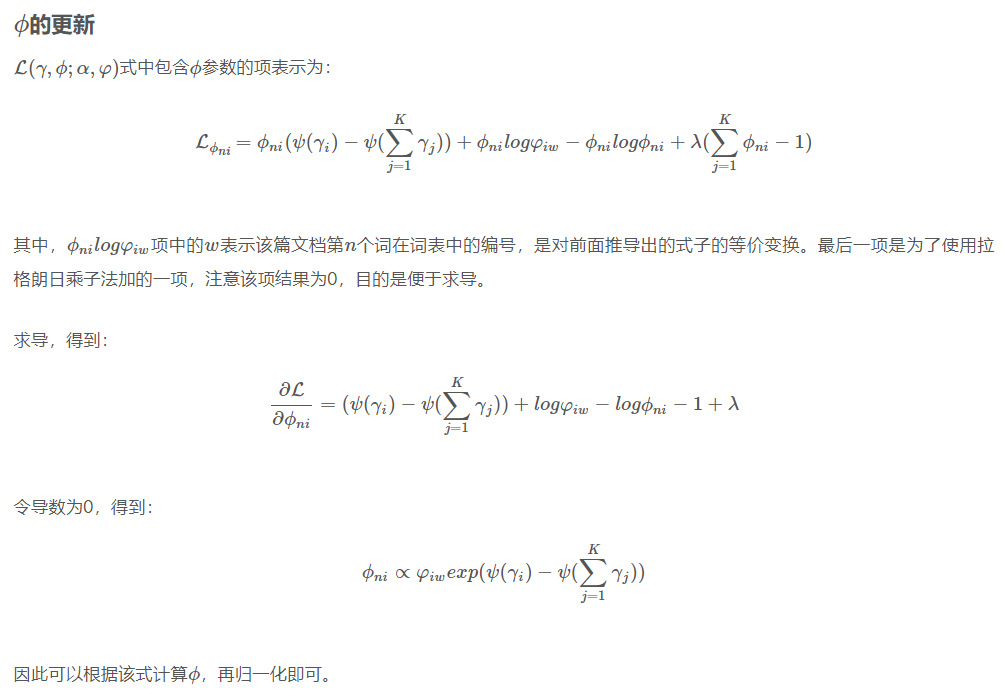
实验报告

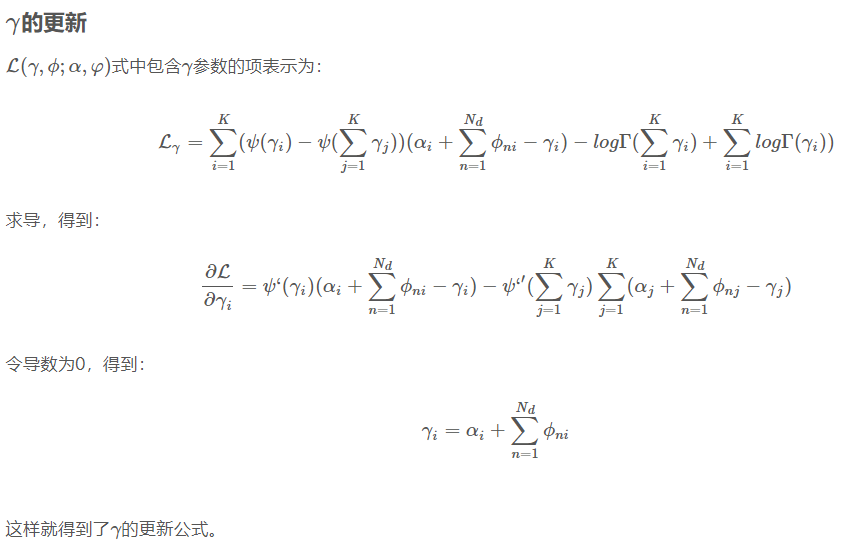








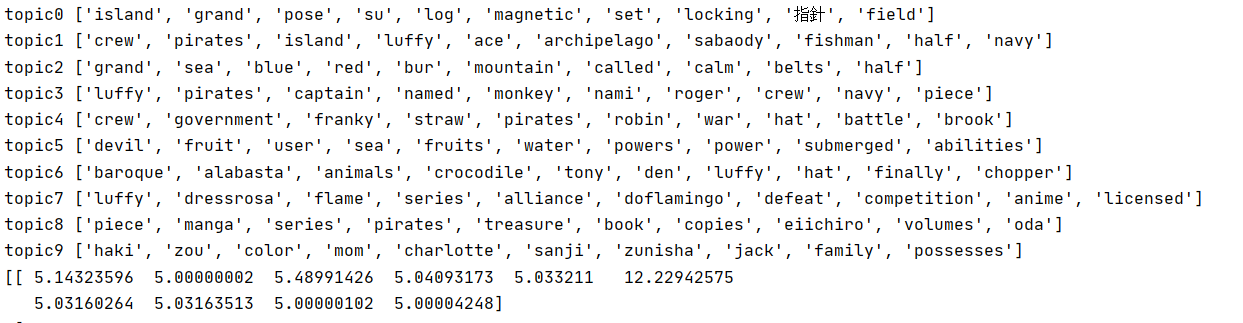




实验数据分析

分别对几个不同的K取值进行了比较，并选取出最好的取值。

K=10时，我们得到各个主题的top words分布和测试集两篇文档的如下:



从topic words来看，这里各个topic之间top words的重合程度较低，且每个topic均具有一定的语义信息。而文档1的显示，其具有属于主题3的先验倾向，而主题3中的词汇与文档1中的词汇 确实有较高的重合度。同样，文档2具有属于主题8的先验倾向，而主题8的高频词也确实和文档2重合度较高。

K=20时，我们得到各个主题的top words分布和测试集两篇文档的 如下:



当K增大到20时，各个topic之间高频词之间的重复度显著增加，很多主题具有较为相似的高频词。但是较好的一方面是，根据文档1，文档2的推断其先验主题，发现文档1,2均有偏向主题15和 14的倾向，观察主题15和14，其高频词组成和K=10时基本一致。说明K=20时依然保留了K=10时的主题信息。

K=30时，我们得到各个主题的top words分布(截取部分)和测试集两篇文档的 如下:



可以发现，与K=20相比，其主题的分布更加细致，有一些较新的主题的出现，例如主题14。但是与 K=20相比，主题高频词的冗余进一步增加。观察可以发现，文档1和文档2具有倾向于主题5和主题2的 先验，这与K=20和K=10时的结果基本一致。

K=50时，我们得到各个主题的top words分布(截取部分)和测试集两篇文档的 如下:



K=50时，有较多主题的高频词分布非常接近，同时观察 ，发现此时文档1和文档2分别具有倾向于主题44和主题21的先验，观察这两个主题，发现与之前相比，其与文档词汇的符合程度有一定的降低。

综上，最好的K值为20。