Xu Wenhao Follow
Facebook Messenger & Chatbot, Machine Learning & Big Data
Mar 19, 2011 · 4 min read

建议的程序员学习LDA算法的步骤

这一阵为了工作上的关系，花了点时间学习了一下LDA算法，说实话，对于我这个学CS而非学数学的人来说，除了集体智慧编程这本书之外基本没怎么看过机器学习的人来说，一开始还真是摸不太到门道，前前后后快要四个月了，算是基本了解了这个算法的实现，记录一下，也供后来人快速入门做个参考。

一开始直接就下了Blei的原始的那篇论文来看，但是看了个开头就被Dirichlet分布和几个数学公式打倒，然后因为专心在写项目中的具体的代码，也就先放下了。但是因为发现完全忘记了本科学的概率和统计的内容，只好回头去看大学时候概率论的教材，发现早不知道借给谁了，于是上网买了本，花了几天时间大致回顾了一遍概率论的知识，什么贝叶斯全概率公式，正态分布，二项分布之类的。
后来晚上没事儿的时候，去水木的AI版转了转，了解到了Machine Learning的圣经PRML，考虑到反正也是要长期学习了，搞了电子版，同时上淘宝买了个打印胶装的版本。春节里每天晚上看一点儿，扫了一下前两章，再次回顾了一下基本数学知识，然后了解了下贝叶斯学派那种采用共轭先验来建模的方式。于是再次尝试回头去看Blei的那篇论文，发现还是看不太懂，于是又放下了。然后某天Tony让我准备准备给复旦的同学们share一下我们项目中LDA的使用，为了不露怯，又去翻论文，正好看到Science上这篇Topic Models Vs. Unstructured Data的科普性质的文章，翻了一遍之后，再去PRML里看了一遍Graphic Models那一张，觉得对于LDA想解决的问题和方法了解了更清楚了。之后从search engine里搜到这篇文章，然后根据推荐读了一部分的Gibbs Sampling for the Uninitiated。之后忘了怎么又搜到了Mark Steyvers和Tom Griffiths合著的Probabilistic Topic Models，在某个周末往返北京的飞机上读完了，觉得基本上模型训练过程也明白了。再之后就是读了一下这个最简版的LDA Gibbs Sampling的实现，再回过头读了一下PLDA的源码，基本上算是对LDA有了个相对清楚的了解。

这样前前后后，也过去了三个月，其实不少时间都是浪费掉的，比如Blei的论文在没有任何相关知识的情况下一开始读了好几次，都没读完而且得到到信息也很有限，如果重新总结一下，我觉得对于我们这些门外汉程序员来说，想了解LDA大概需要这些知识:

- 有基本的概率论的知识，这个拿个大学的课本大概翻一下就好了

  PRML的前两章和Graphic Model那部分需要浏览一下　了解一下所谓的

- 对照着Probabilistic Topic Models直接看LdaGibbsSampling.java的源码

基本上这样一圈下来，基本概念和算法实现都应该搞定了，当然，数学证明其实没那么容易就搞定，但是对于工程师来说，先把这些搞定就能干活了，这个步骤并不适合各位读博士发论文的同学们，但是这样先看看也比较容易对于这些数学问题的兴趣，不然，成天对这符号和数学公式，没有整块业余时间的我是觉得还是容易退缩放弃的。

发现作为工程师来说，还是看代码比较有感觉，看实际应用的实例比较有感觉，看来不能把大部分时间花在PRML上，还是要多对照着代码看。