Randomized Linear Algebra; Tractable Inference

http://freemind.pluskid.org/machine-learning/randomized-linear-algebra-tractable-inference

posted on Free Mind on September 27, 2012
generated with pandoc on December 3, 2015
category: Machine Learning

tags: Talk, Probabilistic Graphical Model, Randomized Algorithm

这是今天听的两个 talk，觉得还是比较有意思的，就跟大家分享一下。

第一个是关于 Randomized Algorithm in Linear Algebra，演讲者是 Petros Drineas，slides 可以在他的主页下载到：Randomized Algorithms in Linear Algebra and the Column Subset Selection Problem。

因为机器学习中很多问题都要用到 linear algebra，所以这个 topic 也是很重要的，他在这里涉及到的 randomized algorithm 主要分为两种，一种是做 sampling，可以是对矩阵的行、列或者是元素进行采样；另一种是 random projection。使用 randomized algorithm 的主要 motivation 是两个方面：第一是数据量太大，无法处理；第二是问题本身是 combinatorial 问题，比如说是 NP 问题，可能需要用随机算法来进行近似。

特别地，他在这里介绍了 SVD 分解和 CX 分解，为了降低计算复杂度，希望对矩阵的列进行 sampling，他在这里讲到了利用列的 leverage score 来进行采样的方式。不过有一个问题就是 leverage score 是需要先 SVD 之后才能计算的..... `=_=bb` 对于一个 $m \times n$ 的矩阵 A 来说（假设 $m > n$），SVD 的复杂度大概是 $O(mn^2)$，不过幸运的是 leverage score 也是可以通过近似算法来逼近的，计算复杂度可以降低到 $O(mn \log m)$。

另外一个方向的研究在于在给定的精度要求下如何能采样更少的列。具体的结果和参考文献见他的 slides。

第二部分主要是讲 random projection，当然是从著名的 Johnson & Lindenstrauss Lemma 开始。然后介绍了一些不同的随机投影算法，主要集中在如何 clever 地构造随机投影矩阵使得计算复杂度尽量降低，其中有一个叫做 Fast JL Transform 的算法需要乘上 P、H、D 三个矩阵.....于是下面的 PhD 们都笑了。最后当然还是回到如何通过这里的方法来加速 SVD 分解。

第二个 Talk 是由 Pedro Domingos[1] 做的关于 Learning Tractable but Expressive Models 的 talk，主要是关于 Graphical Model 和 Bayesian Inference 的。

我之前一直对 Probabilistic Graphical Model 敬而远之，因为好像听说它啥都能 model，不管是什么问题，模型建一下然后就可以做 inference 了，总觉得好像是类似于遗传算法之类的 “万能” 算法，反而不太有什么好玩的了。不过最近多多少少接触了一些，觉得至少从问题的有趣程度上来说还是有挺多好玩的地方的，而且在实际应用中也确实有很多成功的例子。不过，总的来说，虽然总是可以建立概率图模型来建模各种不同的问题，但是经常出现的情况是模型的 inference 并不那么好搞，计

[1] 他也在 Coursera 上开了一门 Machine Learning 课，目前还没有开始，感兴趣的同学可以关注一下。

算是 intractable 的，没有办法，只要利用近似计算，什么 MCMC 还有 Variantional Bayes 之类的。

Figure 1: .....

Speaker 在这里说这样的近似计算实际上对最终效果影响很大，然后他整个 talk 的主要思想是与其使用 intractable 的模型来进行近似 inference，应该是使用 tractable 模型来进行 exact inference 是更好的。当然 tractable model 不是那么容易得到的，由于需要限制模型的复杂性，tractable model 的表达能力通常受到了很大的限制，无法很好地建模实际应用中的问题。然后在这个 talk 中他介绍了几个 tractable 并且很 expressive 的模型，由于不知道他把 slides 放在哪里的，也没有记笔记，所以不太记得所有的方法了，只记得有一个叫做 Arithmetic Circuit 的模型，然后还有一个叫做 Sum-Product Network 的模型是 Arithmetic Circuit 的推广。

后面再推广到 Statistical Relational Learning 方面的主题我不是很感兴趣就没仔细听了。总之主要思想是 tractable model + exact inference 比 intractable model + approximate inference 好，当然这个似乎是公说公有理，婆说婆有理的吧？因为前者（在许多情况下）可以说是对模型的近似，而后者是计算的近似，究竟哪个好呢？Speaker 在演讲中提到他们的方法在实际问题中结果很好，在许多问题上都 beat 了 state-of-the-art，并且能做许多比以前复杂得多的 inference，比如说在 Matrix Completion 方面居然能给定人的上半脸给你补全出下半脸……如果没有使用其他额外信息的话，我怎么都觉得很神奇，不过过分的是 Speaker 的 slides 里一个图都没有，我还指望他展示一个实际的实验效果图呢。不过他说他们在今年的 NIPS 上有一篇这方面的文章，对这方面感兴趣的同学可以期待一下，或者也可以参考他主页上其他已经发表了的文章。

最后水几句，CSAIL 的 Talk 真是很多很多，每天邮件列表里都会像发

大水一般，其中很大一部分就是近几天的 Talk 内容，各个领域都是有的，光过滤掉不相关的领域之后剩下的 Talk 仍然量很大，真是一件很不错的事情。更何况 Talk 上总是有 Free Food，至少也是有 Free Fruit 的！^_^

CSAIL 邮件列表里的另外的水就是一些 “愤青” 们讨论的话题，其中 Richard M. Stallman 就是一只超级大愤青，以前我一直觉得 Linus Torvalds 是个大愤青，现在觉得 RMS 完全有过之而无不及呀，经常在邮件列表里乱喷，比如 Vmware 要来做个 talk，他在那里呼吁大家去现场组织抗议；然后有人发了一个（知名的）拼车的系统鼓励大家拼车，他在那里大喷什么系统搜集的信息泄漏隐私，big brother is watching you 之类的。有时候还真会引起一大堆讨论来。一开始还觉得好玩，时间长里就觉得比较痛苦了，因为现在用智能手机，邮件是有提醒的，一天到晚都有邮件的话..... `=.=bbb`，真希望手机上的邮件客户端也有 Gmail 那样的 mute 功能咯。

邮件列表里第三件神奇的事情就是所谓的 Spontaneous Social Event，事件的主要内容基本上是吃 pizza，不过神奇之处不在这里，而在于时间。一开始我没注意，后来有几次睡觉忘关机，半夜收到邮件，仔细看里下时间，一次是半夜两点五十，邮件说一小时后大家来吃 pizza 呀！我还怀疑是不是邮件服务器在处理邮件发送时间的时候有些 bug，后面又一次邮件里明确说了 12:47am[2] 到 1am 大家来吃 pizza 呀！然后我就只好怀疑 CSAIL 这帮家伙们到底是生活在哪个时区或者哪个维度的了。啊，对了，说到 food，CSAIL 还专门有一个邮件列表叫做 Vulture，顾名思义就是像秃鹫一样吃残羹剩饭咯，听起来有点可怜的样子，不过其实是挺好的避免浪费嘛，基本是因为办活动像有些晚上的 talk 或者中午的组会什么的都会提供食物的，如果是有许多剩余的食物的话，就会被重定向到 Vulture 邮件列表。啊，不过我好像忘记 subscribe 到那里了呢！^_^!!

[2] 想起了 POM 的 2:37pm。