Deep Learning 读书会第六次讨论记录（由@极视角小助手整理）

下面为 2016 年 12 月 18 日 Deep Learning 读书会第五章机器学习基础后面章节 5.5-5.11。本次讨论问题由@livic 整理，AG-GROUP 元芳组织讨论。笔记由极视角筱雅整理。如有想加入读书会讨论的，请联系小助手（微信：Extreme-Vision）.

讨论话题

话题一. 为什么说 MLE 是一个渐进最优的估计器？（@livic）..... 1

话题二. Bayesian Statistics，请以 MLE 和 MAP 为例简述频率派和贝叶斯派的差异。（@livic）..... 2

话题三. 对 bayesian 理论，极大后验估计的理解。不仅仅停留在公式表面。（@livic）..... 3

话题四. Svm 的核函数的选择有什么原则吗？线性 svm 和 LR 的区别除了 max margin 还有其他区别吗？svm 和 hard negative mining 的联系和区别呢？（@livic）..... 5

话题五.神经网络和 svm 最大的不同是否在于前者的核函数是通过层级结构学习出来的，而后者是手工设计的呢？除此之外还有什么不同吗？（@livic）..... 9

话题六.结合 KNN 和 Kmeans 理解有监督学习和无监督学习的差异。（@livic）.....10

话题七.结合标准 bp 算法与累积 bp 算法深入理解 GD 与 SGD 的差异。(@livic).....11

话题八.对于 weight decay 项的选取，目前我们大多数是根据实验是否过拟合去调整，有没有其他一些经验，结合具体的例子讲讲。(@livic) 13

写在最后.....16

话题一. 为什么说 MLE 是一个渐进最优的估计器？（@livic）

极视角小助手

今晚第一个话题，为什么说 MLE 是一个渐进最优的估计器？大家有什么想法吗？

Daiwk

啥叫渐近最优呢？

枫

$$E[\hat{\theta}] = \theta$$
$$E[\hat{\theta}^2] = \theta^2$$

第二条基本是无法满足的，这是因为对于一个无偏估计而言，满足第二条即说明方差为 0.那么自然，我们希望方差越小越好。所以方差只能逐渐逼近。

话题二. Bayesian Statistics，请以 MLE 和 MAP 为例简述频率派和贝叶斯派的差异。（@livic）

AG-GROUP 元芳

那么我们第二个问题其实是上一个问题的延续。Bayesian Statistics，请以 MLE 和 MAP 为例简述频率派和贝叶斯派的差异。

PATAPONPONPON

我感觉 MLE 是对 theta 的点估计，MAP 是对 theta 的全估计。

Joffery

两个不是就差了一个先验吗？

枫

啥叫全估计？

PATAPONPONPON

区别应该就是就贝叶斯学派 theta 服从一个先验分布。全估计是我看资料看到的一个跟点估计对立的词。

Joffery

是不是就是区间估计呢？

极视角小助手
这是 5.6 的内容，要不翻书看看？
要不我们接着下一个问题吧，这两个问题感觉比较抽象，但都是 5.5 和 5.6 的。

AG-GROUP 元芳
大家对 bayesian 理论，极大后验估计的理解，不仅仅停留在公式表面。还是贝叶斯的问题。是在问什么是极大后验么？

枫
最大似然估计和 bayes/map 有很大的不同，原因在于后两种估计方法利用了先验知识，如果利用恰当，可以得到更好的结果。其实这也是两大派别（frequentists and Bayesians）的一个区别。

话题三. 对 bayesian 理论，极大后验估计的理解。不仅仅停留在公式表面。（@livic）

极视角小助手
第三个问题主要是大家对 bayesian 理论，极大后验估计的理解。因为可能大家对这个理解停留在公式里面，不够深入，直观。

AG-GROUP 元芳
我的理解是还是最大后验概率和最大似然的区别吧。然而我也不懂这个问题在问啥，感觉和上一个问题一样。或者是怎么理解 MAP？

极视角小助手
Bayesian 理论呢？

Joffery
或者是它们在现在的机器学习里有什么应用？

张小彬（Bruce）
MLE 的参数估计结果是一个极值点，MAP 的参数估计结果是一个分布。

Joffery
对。

AG-GROUP 元芳
MAP 在教科书上，说的是和极大似然估计不同的是，MAP 寻求的是能使后验概率最大的变量值。应该这个问题是这个意思。

张小彬（Bruce）
比如投硬币问题（伯努利模型），MLE 的参数估计结果可能是，正面朝上的概率是 0.6，但是 MAP 的结果却是一个分部，峰值可能在 0.6 附近分布。

Yc
那怎么理解贝叶斯里面的似然？先验是我们人为对模型的估计？

张小彬（Bruce）
后验=先验*似然

魑魅魍魉
正比吧。

枫
先验是一个分布吧。

安兴乐
似然可以理解为大致分布吧。

Yc
P（x|theta）怎么理解呢？

张小彬（Bruce）
比如上来就假设硬币朝上的概率是 0.5，这个是先验知识；投了一百次硬币，61 次是正面，这个是似然，似然估计出来的结果是 0.61.

Yc
那后验呢？

枫
P（x|theta）是取 theta 的条件下生成 x 的概率。

安兴乐
这不是 theta 情况下 x 的概率分布吗？

张小彬（Bruce）
P（x|theta）就是似然；p（theta）是先验。

Yc
理解了。

张小彬（Bruce）
似然就是在某个模型（参数为 theta）下，数据发生的概率，所以某个样本 x 的似然概率就是 p（x|theta）。

话题四. Svm 的核函数的选择有什么原则吗？线性 svm 和 LR 的区别除了 max margin 还有其他区别吗？svm 和 hard negative mining 的联系和区别呢？（@livic）

AG-GROUP 元芳
0 -- linear: u'*v
1 -- polynomial: (gamma*u'*v + coef0)^degree
2 -- radial basis function: exp(-gamma*|u-v|^2)
3 -- sigmoid: tanh(gamma*u'*v + coef0)
4 -- precomputed kernel (kernel values in training_set_file)
这个是 libSVM 文档里给出的常用核公式选择。主要有线性内核，多项式内核，径向基内核（RBF），sigmoid 核。

Joffery
最常用的是 0 1 2 吧，后面两个没怎么见过。

AG-GROUP 元芳
然后具体怎么选用我觉得是看数据。

安兴乐
看参数是否线性可分。

AG-GROUP 元芳
1. Linear 核：主要用于线性可分的情形。参数少，速度快，对于一般数据，分类效果已经很理想了。
2. RBF 核：主要用于线性不可分的情形。参数多，分类结果非常依赖于参数。有很多人是通过训练数据的交叉验证来寻找合适的参数，不过这个过程比较耗时。我个人的体会是：使用 libsvm，默认参数，RBF 核比 Linear 核效果稍差。通过进行大量参数的尝试，一般能找到比 linear 核更好的效果。
其实我觉得最好的方法就是试。试一下就知道怎么出来了。然后接下来，SVM 核 LR 的区别？

张小彬（Bruce）
Loss function 不同？

枫
一个经验风险最小化，一个结构风险最小化？

张小彬（Bruce）
不是吧。那个只是加个正则项而已。

Yc
一个是分类器，一个是 regression？

PATAPONPONPON
LR 也是分类器啊。

AG-GROUP 元芳
其实我觉得没那么复杂，就是在样本空间中切一刀，LR 希望所有数据都离这一刀（分离面）远远的，而 SVM 只是给了个距离（理解为铺了条路），之在范围内的样本希望他们离得尽量远。范围内的样本叫做支持样本，所以 SVM 叫做支持向量机。我是这么理解的。

Badrobot
核还真挺神奇。

安兴乐
所以除了 max margin，可以把他们理解为本质上一样？

Yc
好像有道理。

AG-GROUP 元芳
这个只是线性核和 LR 的区别吧，SVM 可以解决非线性问题。

枫
结构风险=经验风险+置信风险，max margin 就是置信风险，不知道理解对不对。

安兴乐
我们也可以在 lr 上加个“激活”啊。

枫
其实 LR 加个核也能解决非线性问题。

张小彬（Bruce）
我觉得 svm 的损失只跟少数的样本有关（支持向量？），但是 LR 的损失却跟全部的样本有关。

AG-GROUP 元芳

就是铺路区域覆盖的样本吧。

张小彬（Bruce）
还有这个名字啊？

魑魅魍魉
LR 会拟合离群点，svm 不会。

张小彬（Bruce）
哈哈，好吧，我以为是术语呢。

枫
Logistic regression 也不会拟合离群点。

Yc
我喜欢粗暴的解释啊。
安兴乐
那就把它过拟合了。

AG-GROUP 元芳
这个看有没有奇异值的处理吧？

张小彬（Bruce）
LR 会受 outliner data 影响吧，但是 SVM 理论上就不理了。

AG-GROUP 元芳
如果没有判定为奇异值自然会拟合。

Joffery
离群点有没有可能是支持向量呢？

AG-GROUP 元芳
Svm 因为只处理支持样本，而一般在区间内的，就不太可能是支持样本。毕竟奇异点和大量样本比是少数。

枫
LR 模型保存的是分割超平面，SVM 保存支持向量，SVM 分类是要分类的点与支持向量相乘，LR 是放到模型里算距离，好像没表述清楚。

AG-GROUP 元芳
然后后面的负样本挖掘，我的理解是一个数据的压缩感知过程。当负样本数量远大于正样本时，执行一下对负样本进行筛选。

枫
样本不均衡问题？

AG-GROUP 元芳
这个概念是 RCNN 里最早提出来的。

枫
LR 的离群点问题，可以看看。

AG-GROUP 元芳
就是因为在提取人脸特征时，负样本的数量会比较大，所以提出了一个方法。后面不知道我的理解对不对。

话题五.神经网络和 svm 最大的不同是否在于前者的核函数是通过层级结构学习出来的，而后者是手工设计的呢？除此之外还有什么不同吗？（@livic）

枫
Nn 主要是特征好吧，跟核函数有关系吗？

AG-GROUP 元芳
其实为啥我觉得现在跑的 DNN 核基本上都是 reLu 呢？他这个问题的核应该说

的是激活函数，激活函数的选择。

Joffery
两个应该是有联系的，我也忘了是啥了。

安兴乐
神经网络更应该理解为“多级线性分类”吧。

Joffery
非线性吧？

人工智障 v1.04
把数据扭来扭去扭到最后可以一刀切。

安兴乐
加了多级。

AG-GROUP 元芳
应该是这个意思。

话题六.结合 KNN 和 Kmeans 理解有监督学习和无监督学习的差异。(@livic)

Joffery
是不是因为无监督总是在找一种合适的概率分布？比如 kmeans 就是 GMM，knn 是一种均匀分布？

枫
Knn 的距离度量方式不一样，分布就不一样吧。

Joffery

对，它是非参数估计。

AG-GROUP 元芳
其实我觉得从结果上来看的话，Kmeans 的数据不需要标定，而 KNN 要，就是他俩最大的区别，不知道对不对。

Yc
我看是。

Joffery
那不就是有监督和无监督的区别吗？

枫
问题好奇怪。

Yc
不过，knn 之后也不用标定了吧？不是放一个 test 进去就自动归类了吗？

AG-GROUP 元芳
可是 test 时候用的数据不就是标定的吗。

话题七.结合标准 bp 算法与累积 bp 算法深入理解 GD 与 SGD 的差异。(@livic)

Joffery
是单样本更新和批处理的区别？

Badrobot
累积 bp 不太懂。

AG-GROUP 元芳
不太懂提问者的意思，大家可以说说 GD 和 SGD 的区别？

Yc
Gd 是整个 dataset 的，sgd 是分 batch 的。

Badrobot
Gd 一次运算整个数据集，运行慢。

Yc
因为 gd 的计算量大。

Badrobot
还有个 mini batch gd。

AG-GROUP 元芳
应该就是这个意思吧，累积 BP，一点一点学习。

Yc
Mini batch 和 batch 有什么区别？

Joffery
一样的吧。

安兴乐
就是一个。
Badrobot
Mini batch 可以是 10 个数据一组，gd 和 sgd 的折中情况吧。

Joffery
一般是 32 个？weight decay 是正则项的意思吗？跟 drop out 有关系吗？

话题八.对于 weight decay 项的选取，目前我们大多数是根据实验是否过拟合去调整，有没有其他一些经验，结合具体的例子讲讲。(@livic)

AG-GROUP 元芳
权值衰减吧。

安兴乐
我都是看着 train_loss 曲线手动改。

Joffery
加正则项就可以让它衰减呀。

枫
神经网络里面的正则化怎么弄到？一直不懂。

AG-GROUP 元芳
怎么弄到？

Joffery
我们的好几次实验都是加正则项能抑制过拟合，但是训练的准确率和测试的准确率都挺低，似乎不是很有效。之前过拟合的时候是训练的准确率很高测试很低。加了之后，训练的就降下来了。

人工智障 v1.04
加的太大？

Joffery
很小的和很大的都试过，效果都不是很理想。

AG-GROUP 元芳
防止过拟合最早的就是 Cross Validation，还有 Drop out 吧。

Yc
我也发现。

AG-GROUP 元芳
正则也能防止？

Yc
调参太难。Batch norm 吗？

安兴乐
是加快训练。

Joffery
正则项在损失函数里，最小化损失函数不就把 w 降下来了？

曲晓峰
在字典学习和度量学习里面，正则项添加时都是有明确需要抑制的对象和目标的。

枫
有参考资料吗？

曲晓峰
在神经网络里面，设计损失函数时，加正则也需要根据目标来设计，直接遍历似乎不是好方法。

Joffery
意思是需要选择一部分 w 吗？

曲晓峰
一般对样本正则是担心某些样本权重太高，对偏差的正则，是抑制误差。对模型参数正则抑制过拟合。

Yc

偏差指的是什么？

王晨曦
能说下三种正则的具体使用场景吗？或者举例说明下？

Joffery
在 adaboost 里提高错误样本的权重算是样本正则吗？

曲晓峰
第一个也可以说是样本的标准化，第二个例如信号重建的一范数二范数。第三个一般是额外叠加的约束。

极视角小助手
感谢大家今天的积极讨论，今晚的问题稍微偏抽象啦。如果大家自己还有其他问题，欢迎提出一起来交流~

Livic
呃，非常抱歉，我是提出今天讨论问题的人之一，由于一些原因很抱歉没有参加今天的实时讨论。刚才我回顾了一下大家讨论的过程，感觉大家对部分问题可能没有太 get 到要讨论的点，所以这里我对由我给出的四个讨论题进行一些解释，希望大家能够了解到我提出这个讨论题的缘由，以稍微减少大家的疑惑。

1. 为什么说 MLE 是一个渐近最优的估计器？这个问题是出自 5.5 节 MLE，主要是想讨论下 5.5.2 节中介绍的 MLE 的重要性质。 首先，机器学习是基于统计学的、由 Repersentation + Evaluation + Optimization 三个部分组成的一个研究方向。（Pedro Domingos, A Few Useful Things to Know about Machine Learning, 2012.） 5.5 节中前面部分花了很多篇幅讨论了 MLE 是 Evaluation 这个部分中最最广泛使用的一个评价准则。5.5.2 节就告诉了我们 MLE 受欢迎的具体原因，是因为 MLE 是一个渐近最优的估计器。所以我提出这个讨论题的主要目的是想让大家讨论下 5.5.2 节中提及的 MLE 渐近最优的两个条件以及一致性和高效性。

2. 请以 MLE 和 MAP 为例简述频率派和贝叶斯派的差异。 这个问题是出自 5.6 节 Bayesian Statistics，主要是想讨论下 MLE 作为频率派的常用的 Evaluation

手段，MAP 作为贝叶斯派惯用的 Evaluation 手段具体是怎么操作的，然后两者的差异是什么。 这个问题确实是抽象了一点（或者说偏统计了一点），但是对于机器学习来讲这两个派别太重要了，感觉机器学习基本上可以根据 Evaluation 的不同而直接划分成这两个学派，所以还是提了这个问题让大家讨论下。

6. 请结合 KNN 和 Kmeans 理解有监督学习和无监督学习的差异。 这个问题是出自 5.7 节监督学习算法和 5.8 节无监督学习算法，大家可以注意到 5.7 和 5.8 分别提及了 SVM、Desicion Tree、KNN 和 PCA、K-means 等算法，由于不可能面面俱到，所以我这里拿了最简单但是最容易混淆的 KNN 和 Kmeans 来讨论。感觉基本上搞清楚了 KNN 和 Kmeans 的具体细节、优缺点和两者的相似处和差异之处，就能够对有监督和无监督有个较为清晰的认识了。

7. 请结合标准 bp 算法与累积 bp 算法深入理解 GD 与 SGD 的差异。 这个问题是出自 5.9 节 SGD。个人感觉 5.9 节讲了很多，但是关于 SGD 最最关键的其实只有两句话和一个单词，那就是 "The insight of stochastic gradient descent is that the gradient is an expectation. The expectation may be approximately estimated using a small set of samples." 和"minibatch"，所以我是想针对这个点进行讨论的。然后由于个人感觉 GD 与 SGD 和标准 bp 算法与累积 bp 算法的联系太紧密，所以就直接给抛了出来，如果对这个地方有疑惑，感觉可以去看下周志华老师《机器学习》第 5 章中的标准 bp 算法与累积 bp 算法的对比。。

————————————————————**End**————————————————————

写在最后

非常感谢此次进行讨论交流的朋友们以及群内支持的朋友们 ,希望我们读书会能让大家学到更多，并且讨论后可以对原书有更独到的理解。

#广告时间#