Deep Learning 读书会第四次讨论记录（由@极视角小助手整理）

下面为 2016 年 12 月 11 日 Deep Learning 读书会第五章机器学习基础前四节，即 5.0-5.4。本次讨论问题由@人工智障 v1.04 整理。笔记由极视角小助手整理。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

引言.....1
话题一. 有哪些实例可以佐证"No Free Lunch"理论吗？ （人工智障 v1.04）.....1
话题二. 怎样来衡量一个模型的表达能力(representational capacity)？ （人工智障 v1.04）.....5
话题三. iid 假设具体是怎么保证训练误差和测试误差接近的？ （人工智障 v1.04）..6
话题三. 问题 3 反过来说成立吗？即如果把从两个未知分布产生的同样多的样本用同一个模型同样的算法训练，得到的训练误差和测试误差很接近，能够说明这两个分布也很接近吗？ （人工智障 v1.04）.....8
写在最后.....12

引言

首先讲了机器学习的常见定义(同 Tom.Mitchell 书所讲)和常见任务分类，然后讲训练的流程，从而引出 underfit,overfit 和 VC 维等概念，最后讲防止过拟合用到的技术，包括正则化和交叉验证等。（由@人工智障 v1.04 *总结）*

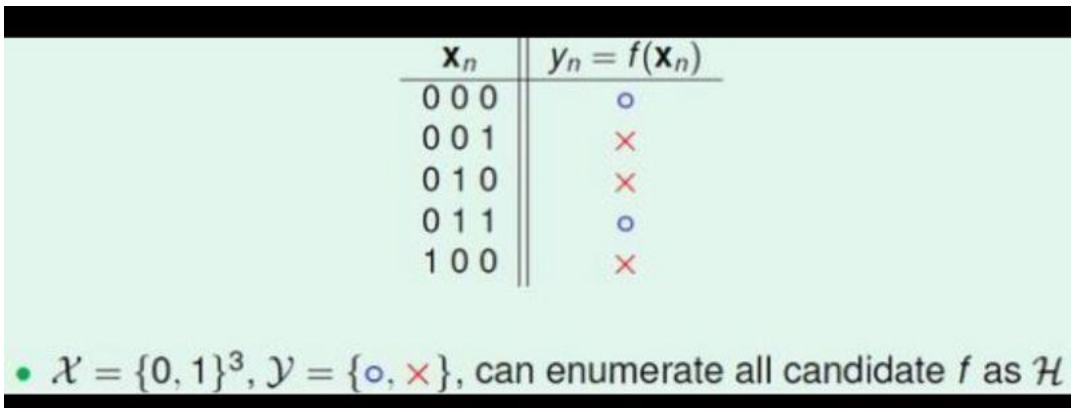
话题一. 有哪些实例可以佐证"No Free Lunch"理论吗？ （人工智障 v1.04）

人工智障 v1.04

这个理论说的是在一些问题上表现好的算法，在其他问题上则不然
就是说所有算法的优劣程度是一致的，考虑到潜在的问题

安兴乐

台大的机器学习基石里 认为是算法不能够有效区分
这个习题



果冻儿
没有完美的理论，更没有完美的算法

纸鸢
这个就仅仅是个结论吗？涉及到一些相关的公式推导吗？其实对这个问题比较晕

人工智障 v1.04
就是不知道大家有没有什么具体的例子

squirrel16
假设有这样一个问题，我们要学习一个包含 100 个变量的布尔函数，手头拿到了 100 万个样本 ,可是完整的样本应当是 2^100 个。这样我们就有 2^100-100 万个（大量的）样本的类别是不知道的。在没有更进一步信息可用的情况下，除了瞎猜也没有其他办法了

清
周志华书里有一个例子 ,在南京从鼓楼到新街口可以骑自行车 ,这是一个好方法。但是从南京到北京骑自行车就不行，可以坐飞机

squirrel16
所以每个学习器都要包含一些“数据之外的知识或者假设” ,才能够讲数据泛化。这个概念的形式化就是“No free lunch 定理” ~

张小彬（Bruce）
所以就是具体问题具体对待？

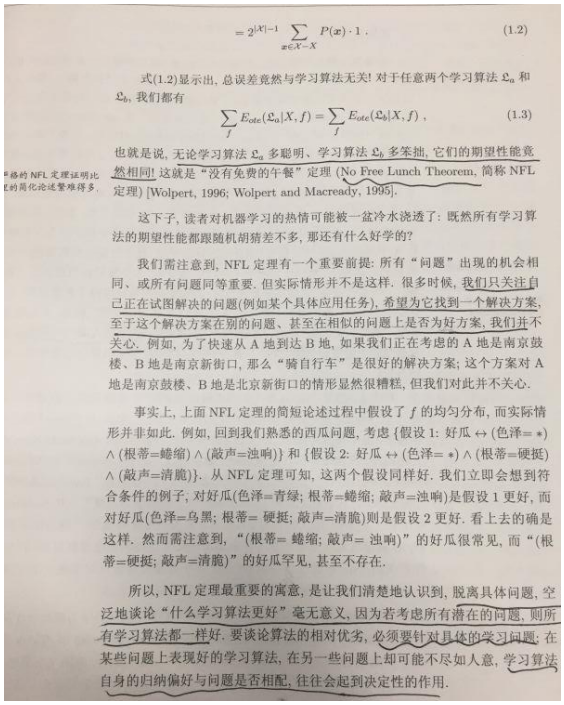
人工智障 v1.04
就是离开业务谈算法毫无意义？

清
因为 NFL 基础是所有问题一样重要，当然找不到一个大统一的算法

婷婷
就是任何一种机器学习算法都无法从不全的样本中学习到一般的结论

张小彬（Bruce）
free lunch 这个词字面意思也好难理解，免费的午餐。？

果冻儿



$$= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x} \in \mathcal{X}-X} P(\boldsymbol{x}) \cdot 1 \,. \tag{1.2}$$

式(1.2)显示出, 总误差竟然与学习算法无关! 对于任意两个学习算法 $\mathfrak{L}_a$ 和 $\mathfrak{L}_b$, 我们都有

$$\sum_f E_{ote}(\mathfrak{L}_a|X,f) = \sum_f E_{ote}(\mathfrak{L}_b|X,f) \,, \tag{1.3}$$

也就是说, 无论学习算法 $\mathfrak{L}_a$ 多聪明、学习算法 $\mathfrak{L}_b$ 多笨拙, 它们的期望性能竟然相同! 这就是“没有免费的午餐”定理 (No Free Lunch Theorem, 简称 NFL 定理) [Wolpert, 1996; Wolpert and Macready, 1995].

这下子, 读者对机器学习的热情可能被一盆冷水浇透了: 既然所有学习算法的期望性能都跟随机胡猜差不多, 那还有什么好学的?

我们需注意到, NFL 定理有一个重要前提: 所有“问题”出现的机会相同、或所有问题同等重要. 但实际情形并不是这样. 很多时候, 我们只关注自己正在试图解决的问题(例如某个具体应用任务), 希望为它找到一个解决方案, 至于这个解决方案在别的问题、甚至在相似的问题上是否为好方案, 我们并不关心. 例如, 为了快速从 A 地到达 B 地, 如果我们正在考虑的 A 地是南京鼓楼、B 地是南京新街口, 那么“骑自行车”是很好的解决方案; 这个方案对 A 地是南京鼓楼、B 地是北京新街口的情形显然很糟糕, 但我们对此并不关心.

事实上, 上面 NFL 定理的简短论述过程中假设了 f 的均匀分布, 而实际情形并非如此. 例如, 回到我们熟悉的西瓜问题, 考虑 {假设 1: 好瓜 $\leftrightarrow$ (色泽= $*$) $\wedge$ (根蒂=蜷缩) $\wedge$ (敲声=浊响)} 和 {假设 2: 好瓜 $\leftrightarrow$ (色泽= $*$) $\wedge$ (根蒂=硬挺) $\wedge$ (敲声=清脆)}. 从 NFL 定理可知, 这两个假设同样好. 我们立即会想到符合条件的例子, 对好瓜(色泽=青绿; 根蒂=蜷缩; 敲声=浊响)是假设 1 更好, 而对好瓜(色泽=乌黑; 根蒂= 硬挺; 敲声=清脆)则是假设 2 更好. 看上去的确是这样. 然而需注意到, “(根蒂= 蜷缩; 敲声= 浊响)”的好瓜很常见, 而“(根蒂=硬挺; 敲声=清脆)”的好瓜罕见, 甚至不存在.

所以, NFL 定理最重要的寓意, 是让我们清楚地认识到, 脱离具体问题, 空泛地谈论“什么学习算法更好”毫无意义, 因为若考虑所有潜在的问题, 则所有学习算法都一样好. 要谈论算法的相对优劣, 必须要针对具体的学习问题; 在某些问题上表现好的学习算法, 在另一些问题上却可能不尽如人意, 学习算法自身的归纳偏好与问题是否相配, 往往会起到决定性的作用.

人工智障 v1.04

就是不能一劳永逸

风牛也马
@果冻儿嗯，书上讲的很明白

果冻儿
实例一：西瓜问题不错
实例二：中国实际国情出发

极视角小助手
求推荐书名给大家。。。

清
就叫 机器学习

风牛也马
周志华的书吧

人工智障 v1.04
周老板的机器学习

风牛也马

清

感觉这本书后半部分介绍的东西略范，得读 reference

话题二. 怎样来衡量一个模型的表达能力(representational capacity)？（人工智障 v1.04）

清

vc dimension

人工智障 v1.04

自由度

不过深层模型是怎么衡量的？

就是我怎么知道模型不够大，然后怎么知道要加多少神经元

果冻儿

暂时只知道用的比较多的是准确率和召回率

清

很显然参数越多 范围越大 模型越深 capacity 越大

人工智障 v1.04

感觉现在都是瞎蒙，看心情

调参前先开光

清
PR 只是用来评估结果的

安兴乐
召回率是模型分类好坏 不是表达能力

果冻儿
嗯嗯，我好像搞混了
模型表达能力

人工智障 v1.04
加多大很多时候还是经验上的东西
总感觉冗余度会很大

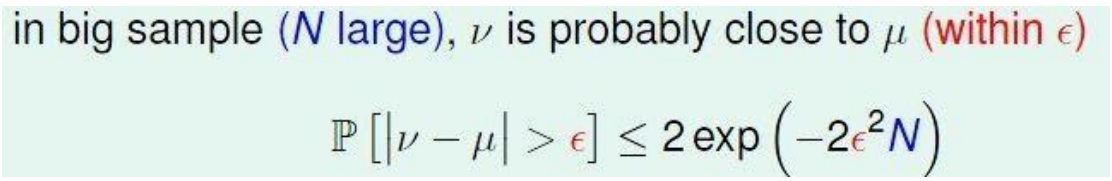
缨宁
当模型表达能力足够的时候，加上合适的训练方法和数据，总会得到不错的较小的训练误差，由此反推，当训练误差都无论如何都降不下来，排除其他因素的情况下，可以考虑一下加大模型的描述空间。

话题三. iid 假设具体是怎么保证训练误差和测试误差接近的？

（人工智障 v1.04）

安兴乐
台大的机器学习课程里面有个例子：大罐子里面有橘色和绿色的弹珠。取一定样本 N 估计橘色弹珠比例为 v,罐子中橘色弹珠比例为 u。我们通过 Hoeffding’s Inequality 知道当 N 足够大时，u 约等于 v.

公式表示为：



清

对 这个得用统计的知识推一下

安兴乐

或者误差在可以接受的程度内

人工智障 v1.04

那这个 N 有一个范围吗

安兴乐

所以尽可能让 N 大

N 和可接受误差有关

缨宁

联系一下大数定律

话题三. 问题 3 反过来说成立吗？即如果把从两个未知分布产生的同样多的样本用同一个模型同样的算法训练，得到的训练误差和测试误差很接近，能够说明这两个分布也很接近吗？（人工智障 v1.04）

安兴乐
这不是 GAN 吗？

人工智障 v1.04
对抗生成网络？

安兴乐
gan 在这么做 nips 正讨论的如火如荼

yc
是不是和 consistency 有关啊

Dandelion
数据是最大的偏执

极视角小助手
这个怎么解释呢？

人工智障 v1.04
@安兴乐 gan 具体是怎么做的
具体还是个生成模型？

安兴乐

一个生成器负责生成，一个辨别器来分类

Dandelion
算法是为了生成特定模型，如果数据是偏执的，那么最终生成的模型也会有偏执

安兴乐
生成器负责模拟真实分布

人工智障 v1.04
然后辨别器会修正生成器的参数吗

安兴乐
模拟一个正态分布
http://nooverfit.com/wp/%e6%89%8b%e6%8a%8a%e6%89%8b%e6%95%99%e4%bd%a0%e5%86%99%e4%b8%80%e4%b8%aa%e7%94%9f%e6%88%90%e5%af%b9%e6%8a%97%e7%bd%91%e7%bb%9c-%e7%94%9f%e6%88%90%e5%af%b9%e6%8a%97%e7%bd%91%e7%bb%9c%e4%bb%a3/#comment-102

人工智障 v1.04
很强
所以那个问题的答案是可以这样反推的是吧

果冻儿
……，可以说两个分布很接近吗

人工智障 v1.04
对啊，可以吗

wy
我觉得是可以的。

人工智障 v1.04
应该说是以假乱真

yc

我不理解啊

人工智障 v1.04
撸一下 gan 的论文应该就懂了，我也是一脸懵逼

yc
测试误差和训练误差很接近不是只是说在一个平面（同一个模型）下，极小值附近的分布差不多
那不代表 data 是同样分布的吧
比如一个平面（模型）有两个凹点，对称分布，然后初始化在中间，但是两组 dataset 分别引导 gradient decent 向两个方向走，也是差不多的误差，但是数据集的分布是不是不太一样

果冻儿
感觉这个问题好有趣啊，就是不知道怎么回答.....

清
我感觉还是得看模型吧
如果模型特别的复杂，对两组数据集经过不同的变化之后，他们的输出分布是一样的，那我觉得可以认为这两个的输入分布式一样的
当然，果模型比较简单，虽然输出分布一样，但是不能就说输入分布是一样的

人工智障 v1.04
就是根据模型复杂度分布相同是有个置信度的这样？

卷心菜
那模型复杂度是否需要一个判定标准呢？

人工智障 v1.04
对啊，界限在哪里，也许很难划清

缨宁
假设两个分布，由三个分量组成，前两个分量采样分布可能比较相近，但第三个分量相差悬殊，然后模型经过计算后发现前两个分量起决定性作用，对两个分布

计算出来的误差相近，但其实这两个联合分布差异还是可以很大的。也就是虽然分布不同，但经过模型特征变换后决定性的特征一致，也可以做到如讨论所说的效果。这有点让我联想到迁移学习。看似不同的数据但对于某个问题有特定的本质联系。

人工智障 v1.04
也就是由于模型未知，相异的分布也可能产生同样的特征，所以到特征层面已经很难反映原来的分布，了？
缨宁
对，就像特征分解一样，不同的向量也可以产生一样的正交基，支撑起同一片 n 维空间

卷心菜
你说的这个三分量和图像检索中的三元组有什么异同吗？

人工智障 v1.04
这样的话糊弄模型是不是变得很容易，我记得 cs231n 有个实验就是在一张猫的图片上猛加噪声最后分类还是猫，是因为主成分还是没多少变化但人已经看不出来了

缨宁
人类说糊弄模型只是因为模型表现得还不够像人类，人类觉得自己看不出来了模型还看的出来就是以噪声糊弄了他，其实也有可能其中的不变性确实捕捉的比人类好了

希希
模糊就是人类自己的思维还没被自己完全理解 所以认为是模糊的
你们说的那个是大数率。样本足够大的时候 分布概率和样本差不多
至于要样本多大，才相差的小。这个可以有二阶矩估计
也可以由四阶矩估计
daiwk
还和模型的复杂程度有关吧

比如一个处理中文的 nn 的 embedding 层，你是对分词去做 embedding 或者

分字去搞，为了保证模型的表达能力比较 ok，分词和分字需要的训练样本是差很多的，因为字典大小差很多，分字汉语可能常用的就几万个字，而分词可能有几十万几百万，embedding 的参数空间是词典大小乘以第一个隐层的节点数，需要的数据量来训练这么多参数才不会过拟合就会差很多了

————————————————————**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#
免费线上技术分享，视觉前沿资讯关注请关注极市平台公众号。