机器的崛起*

Larry Wasserman, trans by Sheldon

机器的崛起

在统计界50周年的主席会议上，我感受到机器学习的崛起，以及它对统计将意味着什么。它意味着很多新的研究领域，新的应用领域和很多新的朋友（同事）。我们的学生将和机器学习的学生竞争工作。我乐观地觉得空想主义的统计系会热情地拥抱这个新生的领域；然而那些无视或是避开机器学习的人终将发现他们被遗弃在一片陈旧过时的瓦砾堆中。

引子

统计是一门讲如何从数据中“学习”的科学，机器学习也是一门讲如何从数据中学习的科学。它们目标相同，但是却有着历史，风格惯例，着重点以及文化上的差异。

我们并不讳言统计在科学甚至社会学上的重要性以及应用上的种种成功。我也和自豪于身为此中一员。所以这篇短文仅仅聚焦于机器学习将给我们带来的机会和挑战。

在这25年的职业生涯中，我一直亲眼目睹着机器学习从一坨相当原始（虽然非常聪明的）分类算法，发展到如今不论在理论还是应用上都相当复杂的“科学”。

只要扫视一眼The Journal of Machine Learning Research(mlr.csail.mit.edu)或NIPS(books.nips.cc)，我们就能看到相当多的熟面孔：

> 条件似然，序列设计，再生核Hilbert空间，聚类，生物信息，　mini-max理论，稀疏回归，大协方差估计，模型选择，密度估计，　图模型，小波，非参数回归

即使是在一本旗舰级的统计杂志上，上面的话题也大多会出现。

现在我们明白了，那些机器学习的研究者们，那些曾经忽视着主流统计方法和主流统计理论的家伙，现在不仅仅意识着这些曾经的统计课题，更加是积极主动地参与在这些前沿的研究之中。

*本文原文在http://www.stat.cmu.edu/\protect\unhbox\voidb@x\penalty\@M\{}larry/Wasserman.pdf由Wasserman教授所写，翻译纯属个人兴趣，若有错误，还望谅解

然而反方面，机器学习中的一些热点话题却基本被统计界忽视了。作为一个统计学家，我们不想被排斥在外，我们要跟上ML的潮流，改变过时地只是传授方法和改革研究生课程。

会议文化

ML以超过统计的步伐发展着。一开始，ML的研究者开发的专家系统里完全排斥使用概率。然而很快他们就开始使用一些比较高级的统计概念，比如empirical process和度量集中度的概念。这样的转换也仅发生于短短的数年之间。这样的高速一部分原因便是就来源于这个群体的会议文化。ML研究的主战场是会议文集而非杂志。

研究生们可以发一连串的文章，带着强悍的简历毕业。这种高节奏的一个重要原因还是，会议的文化。

准备一篇统计学的文章的流程一般是这样的：先有个想法（方法），慢慢地琢磨它，发展它，证明相关的一些结论，最终你将它写了出来，然后投稿。接着审稿人开始审你的文章，差不多一两年，你终于发了一篇paper。

在ML界，会议文章才是硬通货。每个主流会议（NIPS，AISTAT，ICML，COLT）的期限都很紧。这使得你不得不停下精雕细琢你的想法，马上开始写文章。教员和学生一样被期限压着，这带来了某种荣辱以共，各方都得到了好处。比如没有人会在意你在NIPS的截稿日前取消一节课。而且，　截稿日后，所有人又同时面临另一个期限：审对方的文章；所以人都会非常高效地做事。如果你有主意，却没有投稿，那么你的厄运可能就来了，因为其他人可能就把你的主意给占了。

这种压力其实也不错；整个领域都保持着高节奏。你也许觉得这将导致各种低质量的文章和主意泛滥，那么我只好建议你读一读nips.cc的文章。那上面不少文章写得相当地深刻，虽然确实会有一些糟糕的论文。就像我们的杂志一样，ML的论文一样要被审稿，不过速度快得多，　如果有个细节需要补充，作者可以再发一篇杂志版本的文章作为补充。

如果没有这样截稿期催生的动力，这个领域不会发展地这么快。统计界为了自己也好，要应对ML的竞争也好，都得把这种文化当个事。

当然，这种会议文化也有问题。东西都是匆匆忙忙做出来的，过程也没有细细推敲，很多细节都一闪带过。不过瑕不掩瑜，得到的还是要多些的。

被忽视的研究领域

有很多ML里的中心议题，在统计里却几乎被忽略了。这令人尴尬，我们统计本可以给地更多。举例来说，半监督推断，计算拓扑，在线学习，序列化博弈，哈希，能动学习（active learning），深度学习，differential privacy，随机投影和再生核Hilbert空间。具有讽刺意味的是，　一些概念—例如序列化博弈和再生核Hilbert空间，是源于统计的。

案例研究

我很幸运。我身处一个有机器学习系的学院，而且，更加重要的是，ML的人们欢迎统计学家参与进来。所以我幸运地和ML的同事们及他们的学生一起工作，参加他们的讨论班，并且在ML的系里开始课程。

有太多的研究是很参与ML的研究相关的。比如，统计拓扑，图模型，半监督推断，共性预测和differential privacy.

由于这篇短文仅仅反映了我个人的观点，我就来简要介绍下我有幸做过的两个课题。我想说明的是，对于机器学习来说，统计的思维是多么地有用。

案例I：半监督推断

设有一组数据 $(X_1, Y_1), ..., (X_n, Y_n)$ ，我们想通过X预测Y。如果Y是离散的，那么这是个分类问题。如果Y是实值的，这是个回归。进一步，我们还假设测试数据$X_{n+1}, ..., X_N$没有相应的Y值。现在我们有带标签的$\mathcal{L} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$和无标签的$\mathcal{U} = \{X_{n+1}, ..., X_N\}$。有没有办法利用无标签数据提供的信息来提高预测的精度？这被称为半监督推断。

考虑图1，协变量是$x = (x_1, x_2) \in \mathbb{R}^2$。结果是个二元分类，如圆圈方块所示。分类边界不是很清晰。图2增加了无标签数据。我们可以比较明显的看到两类。如果加上$P(Y = 1|X = x)$光滑的条件，那么就可以使用无标签数据来细化分类的边界。

很多文章想出了种种有启发性的方法来利用无标签数据。简单解释如下。试想我们下载了一百万张带有猫狗图片的网页。选取100张对它们进行手工标注。半监督方法可以让我们使用其他999,900张图片来构造一个更好地分类器。

但半监督推断可行么？或者说，什么条件下它可行？文[1]里，我们给出了以下结论（这里不严格地叙述如下）：

设$X_i \in \mathbb{R}^d$。设S_n为只使用有标签数据的监督学习的估计子的集合，用SS_N记半监督估计子（使用全部的有标签和无标签数据）的集合。令m为无标签数据的个数，假设$m \geq n^{2/(2+\xi)}$，其中$0 < \xi < d - 3$。令$f(x) = \mathbb{E}(Y|X = x)$。那么存在一大类的非参数分布$\mathcal{P}_n$使得以下结论为真：

1. 存在半监督估计子$\hat{f}$使得

$$\sup_{P \in \mathcal{P}_n} R_P(\hat{f}) \leq \left(\frac{C}{n}\right)^{\frac{2}{2+\xi}}$$

，这里$R_P(\hat{f}) = \mathbb{E}(\hat{f}(X) - f(X))^2$称为分布$P$下估计子$\hat{f}$的风险。

2. 对监督估计子S_n有

$$\inf_{\hat{f} \in S_n} \sup_{P \in \mathcal{P}_n} R_P(\hat{f}) \geq \left(\frac{C}{n}\right)^{\frac{2}{d-1}}$$

3. 由二者得到以下，

$$\frac{\inf_{\hat{f}\in SS_N}\sup_{P\in\mathcal{P}_n}R_P(\hat{f})}{\inf_{\hat{f}\in S_N}\sup_{P\in\mathcal{P}_n}R_P(\hat{f})}\leq\left(\frac{C}{n}\right)^{\frac{2(d-3-\xi)}{(2+\xi)(d-1)}}\to 0$$

，即半监督的估计被监督的估计控制。

分布类$\mathcal{P}_n$的选取使得X的边缘分布在某个低维子空间上非常集中，从而相应的回归函数也光滑。我们还没有证明半监督推断对监督推断的改进一定上上面这种样子的，但是我们猜想这个结论是对的。我们的框架里引入了一个参数α来刻画半监督假设的强度。我们事实上证明了，总可以找到正确的α来匹配数据。

案例II：统计拓扑

计算拓扑许和机器学习的研究人员发明了很多函数和数据的形状的方法。这里我简单回顾下我们再流形估计上的工作[2, 3, 4]。

设M是嵌入$\mathbb{R}^D$中的d维流形。设$X_1,\ldots,X_n$是以P为分布的样本，而P以M为支撑。假设观测值如下：

$$Y_i=X_i+\epsilon_i,\ i=1,\ldots,n$$

这里$\epsilon_1,\ldots,\epsilon_n\sim\Phi$为噪音。

机器学习的人想出了很多办法来估计流形M，但是有个统计学上的问题长期没有解决：估计得精度怎么样？一种方法是在某种损失函数下计算minimax风险。令$\hat{M}$为M的估计。我们可以自然地定义损失为Hausdorff的：

$$H(M,\hat{M})=\inf\left\{\epsilon:M\subset\hat{M}\oplus\epsilon,\hat{M}\subset M\oplus\epsilon\right\}$$

如果记$\mathcal{P}$为分布类，待估流形M作为参数，是P的支撑集，那么minimax的风险度量定义如下：

$$R_n=\inf_{\hat{M}}\sup_{P\in\mathcal{P}}\mathbb{E}_P[H(\hat{M},M)]$$

显然，这个定义依赖于我们对M和Φ的假设。

References

[1] M. Azizyan, A. Singh, and L. Wasserman, Density-sensitive semisupervised inference, The Annals of Statistics (2013).

[2] Christopher R. Genovese, Macro Perone-Pacifico, Isabella Verdinelli, and Larry. Wasserman, Manifold estimation and singular deconvolution under hausdorff loss, The Annals of Statistics (2012), no. 40, 941 - 963.

[3] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman, Minimax manifold estimation, Journal of Machine Learning Research (2012), 1263 – 1291.

[4] _____, Nonparametric ridge estimation, arXiv:1212.5156 (2012).