机器学习中的数学(1):MIT 大牛的综述

“想要深入了解机器学习和深度学习背后的数学支撑？想要避免成为调包侠？想要做更前沿更基础的研究？不仅是 Import xxx as xx，这里是“机器学习中的数学”第一篇，带你从宏观上认知 AI 背后的数学❤️”

❤️ 00 导言

机器学习的数学原理（1）：宏观介绍篇

申明：本文内容来源于一位 **MIT** 大牛对机器学习中的数学的综述，原文地址已经找不到了（如果有人知道原文地址的话，不妨与大家分享），我将内容转来与大家分享，并做了一些改动和补充(主要是增加了凸优化与非凸优化的部分，并在文末贴上了一些数学课程的学习资料)，以下是文章内容。

从大学到现在，课堂上学的和自学的数学其实不算少了，可是在研究的过程中总是发现需要补充新的数学知识。Learning 和 Vision 都是很多种数学的交汇场。看着不同的理论体系的交汇，对于一个 researcher 来说，往往是非常 exciting 的 enjoyable 的事情。不过，这也代表着要充分了解这个领域并且取得有意义的进展是很艰苦的。记得在两年前的一次 blog 里面，提到过和 learning 有关的数学。今天看来，我对于数学在这个领域的作用有了新的思考。

❤️ 01 代数与统计

对于 Learning 的研究，Linear Algebra (线性代数) 和 Statistics (统计学) 是最重要和不可缺少的。这代表了 Machine Learning 中最主流的两大类方法的基础。一种是以研究函数和变换为重点的代数方法，比如 Dimension reduction，feature extraction，Kernel 等，一种是以研究统计模型和样本分布为重点的统计方法，比如 Graphical model, Information theoretical models 等。它们侧重虽有不同，但是

常常是共同使用的，对于代数方法，往往需要统计上的解释，对于统计模型，其具体计算则需要代数的帮助。

❤️

02 数学分析

以代数和统计为出发点，继续往深处走，我们会发现需要更多的数学。数学分析，其基础性作用不言而喻。Learning 研究的大部分问题是在连续的度量空间进行的，无论代数还是统计，在研究优化问题的时候，对一个映射的微分或者梯度的分析总是不可避免。而在统计学中，Marginalization 和积分更是密不可分——不过，以解析形式把积分导出来的情况则不多见。

❤️

03 偏微分方程

Partial Differential Equation （偏微分方程)，这主要用于描述动态过程，或者仿动态过程。这个学科在 Vision 中用得比 Learning 多，主要用于描述连续场的运动或者扩散过程。比如 Level set, Optical flow 都是这方面的典型例子。

❤️

03 泛函分析

Functional Analysis (泛函分析)，通俗地，可以理解为微积分从有限维空间到无限维空间的拓展——当然了，它实际上远不止于此。在这个地方，函数以及其所作用的对象之间存在的对偶关系扮演了非常重要的角色。Learning 发展至今，也在向无限维延伸——从研究有限维向量的问题到以无限维的函数为研究对象。Kernel Learning 和 Gaussian Process 是其中典型的例子——其中的核心概念都是 Kernel。很多做 Learning 的人把 Kernel 简单理解为 Kernel trick 的运用，这就把 kernel 的意义严重弱化了。在泛函里面，Kernel (Inner Product) 是建立整个博大的代数体系的根本，从 metric, transform 到 spectrum 都根源于此。

❤️

04 凸优化与非凸优化

凸优化是指一种比较特殊的优化，是指求取最小值的目标函数为凸函数的一类优化问题。其中，目标函数为凸函数且定义域为凸集的优化问题称为无约束凸优化问题。而目标函数和不等式约束函数均为凸函数，等式约束函数为仿射函数，

并且定义域为凸集的优化问题为约束优化问题。凸优化在机器学习中应用的十分广泛，最基本的梯度下降法以及其各种变形都是凸优化的中方法。

非凸优化是指不满足凸条件的优化问题，绝大部分问题都属于此类，有趣的是，近年来，机器学习领域出现了一种新方法，不对非凸问题进行松弛处理而是直接解决。引起目标是直接优化非凸公式，该方法通常被称为非凸优化方法。非凸优化方法常用的技术包括简单高效的基元，如投影梯度下降、交替最小化、期望最大化算法、随机优化及其变体。这些方法在实践中速度很快，且仍然是从业者最喜欢用的方法。

❤️

05 测度论与实分析

Measure Theory (测度理论)，这是和实分析关系非常密切的学科。但是测度理论并不限于此。从某种意义上说，Real Analysis 可以从 Lebesgue Measure（勒贝格测度）推演，不过其实还有很多别的测度体系——概率本身就是一种测度。

测度理论对于 Learning 的意义是根本的，现代统计学整个就是建立在测度理论的基础之上。在看一些统计方面的文章的时候，你可能会发现，它们会把统计的公式改用测度来表达，这样做有两个好处：所有的推导和结论不用分别给连续分布和离散分布各自写一遍了，这两种东西都可以用同一的测度形式表达：连续分布的积分基于 Lebesgue 测度，离散分布的求和基于计数测度，而且还能推广到那种既不连续又不离散的分布中去。而且，即使是连续积分，如果不是在欧氏空间进行，而是在更一般的拓扑空间（比如微分流形或者变换群），那么传统的黎曼积分就不 work 了，你可能需要它们的一些推广，比如 Haar Measure 或者 Lebesgue-Stieltjes 积分。

❤️

06 拓扑学

Topology（拓扑学)，这是学术中很基础的学科。它一般不直接提供方法，但是它的很多概念和定理是其它数学分支的基石。看很多别的数学的时候，你会经常接触这样一些概念： Open set / Closed set， set basis， Hausdauf, continuous function，metric space, Cauchy sequence, neighborhood, compactness, connectivity。很多这些也许在大学一年级就学习过一些，当时是基于极限的概念获得的。

看过拓扑学之后，对这些概念的认识会有根本性的拓展。比如，连续函数，当时是由 epison 法定义的，就是无论取多小的正数 epsilon，都存在 xxx，使得 xxx。这是需要一种 metric 去度量距离的，在 general topology 里面，对于连续函数的定义连坐标和距离都不需要——如果一个映射使得开集的原像是开集，它就是连续的——至于开集是基于集合论定义的，不是通常的开区间的意思。这只是最简单的例子。当然，我们研究 learning 也许不需要深究这些数学概念背后的公理体系，但是，打破原来定义的概念的局限在很多问题上是必须的——尤其是当你研

究的东西它不是在欧氏空间里面的时候——正交矩阵，变换群，流形，概率分布的空间，都属于此。

❤

07 微分流形

Differential Manifold (微分流形)，通俗地说它研究的是平滑的曲面。一个直接的印象是它是不是可以用来 fitting 一个 surface 什么的——当然这算是一种应用，但是这是非常初步的。本质上说，微分流形研究的是平滑的拓扑结构。

一个空间构成微分流形的基本要素是局部平滑：从拓扑学来理解，就是它的任意局部都同胚于欧氏空间，从解析的角度来看，就是相容的局部坐标系统。当然，在全局上，它不要求和欧氏空间同胚。它除了可以用于刻画集合上的平滑曲面外，更重要的意义在于，它可以用于研究很多重要的集合。一个 n-维线性空间的全部 k-维子空间(k < n)就构成了一个微分流形——著名的 Grassman Manifold。所有的标准正交阵也构成一个流形。

一个变换群作用于一个空间形成的轨迹(Orbit) 也是通常会形成流形。在流形上，各种的分析方法，比如映射，微分，积分都被移植过来了。前一两年在 Learning 里面火了好长时间的 Manifold Learning 其实只是研究了这个分支的其中一个概念的应用: embedding。其实，它还有很多可以发掘的空间。

❤

08 群论

Lie Group Theory (李群论)，一般意义的群论在 Learning 中被运用的不是很多，群论在 Learning 中用得较多的是它的一个重要方向 Lie group。定义在平滑流行上的群，并且其群运算是平滑的话，那么这就叫李群。因为 Learning 和编码不同，更多关注的是连续空间，因为 Lie group 在各种群中对于 Learning 特别重要。各种子空间，线性变换，非奇异矩阵都基于通常意义的矩阵乘法构成李群。在李群中的映射，变换，度量，划分等等都对于 Learning 中代数方法的研究有重要指导意义。

❤

09 图论

Graph Theory（图论)，图，由于它在表述各种关系的强大能力以及优雅的理论，高效的算法，越来越受到 Learning 领域的欢迎。经典图论，在 Learning 中的一个最重要应用就是 graphical models 了，它被成功运用于分析统计网络的结构和规划统计推断的流程。Graphical model 所取得的成功，图论可谓功不可没。

在 Vision 里面，maxflow (graphcut)算法在图像分割，Stereo 还有各种能量优化中也广受应用。另外一个重要的图论分支就是 Algebraic graph theory (代数图论)，主要运用于图的谱分析，著名的应用包括 Normalized Cut 和 Spectral Clustering。近年来在 semi-supervised learning 中受到特别关注。

10 数学学习资料

1.线性代数：

MIT 教材： Introduction to Linear Algebra (3rd Ed.) by Gilbert Strang；

视频：

http://ocw.mit.edu/OcwWeb/Mathematics/18-06Spring-2005/CourseHome/ind ex.htm

2.概率与统计：

Applied Multivariate Statistical Analysis (5th Ed.) by Richard A. Johnson and DeanW. Wichern；

Introduction to Graphical Models (draft version). by M. Jordan and C. Bishop

3.数学分析：

Principles of Mathematical Analysis, by Walter Rudin

4.凸优化与非凸优化

convex optimization， by Boyd;

Provable Nonconvex Methods/Algorithms

5.泛函分析：

Introductory Functional Analysis with Applications, by Erwin Kreyszig.

6.拓扑学：

Topology (2nd Ed.) by James Munkres

7.微分流形：

Introduction to Smooth Manifolds. by John M. Lee

8.群论:

Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. by Brian C. Hall

9.图论：

Graph Theory Moody&Bondy

点我跳转本文 github 链接

往期精彩推荐

[机器学习系列（1）：十分钟掌握深度学习的原理、推导与实现]
[机器学习系列（2）：初始化的一小步，网络性能的一大步]
[机器学习系列（3）：几分钟了解正则化及其 python 实现]
[机器学习系列（4）：除了上双路泰坦，你还有优化算法]

你是完美的，
是你支持我一路向前，
是你给了我安全感，
你的名字叫 Math。
-by LSayhi 的神经网络

微信公众号：AI 有点可 ai
▼更多**原创干货**和最新资讯，请**关注**我吧▼