

极客大学机器学习训练营

机器学习基本概念

王然

众微科技 AI Lab 负责人

二〇二一年五月十日

- 1 怎样学数学
- 2 机器学习的各种角度和建模流程
- 3 概率论和统计学复习
- 4 极大似然估计
- 5 贝叶斯估计和变分贝叶斯
- 6 矩阵和张量求导
- 7 总结

- 1 怎样学数学
- 2 机器学习的各种角度和建模流程
- 3 概率论和统计学复习
- 4 极大似然估计
- 5 贝叶斯估计和变分贝叶斯
- 6 矩阵和张量求导
- 7 总结

- ▶ AI 的语言 → 不理解数学，不可能理解模型
- ▶ 创新的根基 → 看起来创新不多，但是实际上有很多地方可以创新，而且创新没有那么难
- ▶ 数学锻炼思维

- ▶ 把数学当做语言：不管它的意思，严格按照要求 → 我们主要讲方法
- ▶ 数学真正的学法，是以证明为目的的

核心：

- ▶ Frame and Hypotheses
- ▶ Elements and Relationships
- ▶ Patterns
- ▶ Intuition
- ▶ Retrospect and Empathetic
- ▶ Bucket(In/Out/New)
- ▶ Strategic minds

- ▶ 一遍听懂，不现实；不论老师讲的多细，重复一百遍也没有效果；
- ▶ 必须要回去对着自己推导，如果卡住就问助教；
- ▶ 自己推过之后就会发现；哇，这个怎么这么简单；
- ▶ 自己不推永远都是听天书；
- ▶ 如果前面概念不清楚，不可能听得懂后面的概念。

- ▶ 机器学习的各种角度和建模流程
- ▶ 概率论和统计学基础概念复习
- ▶ 极大似然体系和 EM 算法
- ▶ 贝叶斯体系和 Variational Bayes 算法
- ▶ 矩阵代数：基本概念复习和 Tensor 求导

- 1 怎样学数学
- 2 机器学习各种角度和建模流程
- 3 概率论和统计学复习
- 4 极大似然估计
- 5 贝叶斯估计和变分贝叶斯
- 6 矩阵和张量求导
- 7 总结

- ▶ 最终目的：效果好，即准确性高
- ▶ 为了达到最终目的，必须从不同角度考虑

- ▶ 最简单的是视角
- ▶ 目标：给定 X 预测 y
- ▶ 假设：存在真实的 $y = f_0(X)$
- ▶ 如果知道 f_0 ，那么就不需要做任工作
- ▶ 但是我们不知道，所以需要逼近

- ▶ 观测 $\{X_i, y_i; i \in \mathcal{I}\}$
- ▶ 可以假设 $f \in \mathcal{F}$
- ▶ 目标：给定一个损失函数 c , 最小化 $\sum_i c(f(X_i), y_i)$
- ▶ 这个估计可以称之为 \hat{f}

什么样的 \hat{f} 是好的

- ▶ 最理想状况 $\hat{f} = f_0$; 事实上 (可能) 不可能
- ▶ 不可能原因 (一): 没有所有的 x 和 y 的组合
- ▶ 不可能原因 (二): $f_0 \notin \mathcal{F}$
- ▶ 不可能原因 (三): 求解 \hat{f} 时候有困难
- ▶ 但是基本启示是: 要找到一个足够大的 \mathcal{F} 使它包含 f_0 , 并且要求 \mathcal{F} 应该足够小使得求解比较容易 \rightarrow 自相矛盾

- ▶ 本质上来说，世界上是随机的
- ▶ 随机的来源：
 - ▶ 缺乏信息 → 最主要问题，在表格化数据中最为明显
 - ▶ 测量误差 → 大部分信息都有误差
 - ▶ 比如说年龄 800 岁，收入 400 万亿
 - ▶ 模型误差 → 假设模型形式和现实的差别
 - ▶ 估计误差 → 得到模型过程中造成的误差
 - ▶ 优化误差 → 求解过程中的误差
 - ▶ 评估误差 → 评估本身也存在误差

- ▶ 假设目标是用身高预测体重
- ▶ 为什么不可以进行插值？

请思考

- ▶ 缺乏信息：人有胖有瘦，仅仅给定身高，不可能判断
- ▶ 导致结果：如果要求身高必须解释体重，身高就承担了非理性的要求
- ▶ 相关结果：bias 较大
- ▶ 统计学根本区别于函数逼近的原因
 - ▶ 函数逼近： $y = f_0(X)$
 - ▶ 统计学 $y = f_0(X) + \epsilon$

- ▶ Bias: 话说得很详细, 但是很不准
 - ▶ 北京明天下午两点四十分会发生里氏 2.6 级地震
- ▶ Variance: 含糊其词, 但是很准
 - ▶ 在这个世界上有一天会发生地震
- ▶ 往往存在 Bias 和 Variance 的权衡 (但这不是全部, 它本身的数学理论只是针对回归的)
- ▶ Bias 大: 过拟合
- ▶ Variance 大: 欠拟合

- ▶ 往往难以处理
- ▶ 是数据预处理一个重要部分

- ▶ 假设背景：存在一个上帝知道的真实的模型，但他不知道部分误差，所以模型一定会有损失
 - ▶ 但就该损失函数而言，这个真实的模型一定是预测最好的
- ▶ 现实情况：因为我们不知道真实的模型，所以只能采用一些模型来逼近
- ▶ 如果模型跟真实模型很近，则效果应该是最好的
 - ▶ 一般情况下不知道真实模型，只能选择一般的模型 → 估计方差大

- ▶ 即使对于同样的模型或问题，也有不同办法得到模型的参数
 - ▶ 极大似然估计和贝叶斯估计
 - ▶ 增强学习中的 Q-learning 和 Policy Gradient
- ▶ 好的方法可以减少其中误差

- ▶ 求解的过程，就是迭代的过程
- ▶ 迭代是否会收敛是一个很大的问题
- ▶ 在神经网络中尤其明显，但在传统模型中也存在

- ▶ 因为不知道真实的损失函数（除非有无限多的测试样本），所以必须评估
- ▶ 评估的越多，训练样本就越少 → 出现了交叉验证的概念
- ▶ 注意避免不公平的评估

- ▶ 只用训练集 → 不公平
- ▶ 无数次的测试训练集 → 不可以（否则猜就可以了）
- ▶ 建模数据和实际场景不同：在 2019 年建模预测 2020 年上半年旅游业情况

- ▶ **重要原则：**一定要看评估本身的误差多大，然后决定做法是否有提升
- ▶ **重要提示：**
- ▶ 越是误差小的领域，需要概率角度越多
- ▶ 误差大的领域，概率角度可能不能帮上太多忙，更应该找可以优化的地方

- ▶ 从概率理论上来说，预训练不应该有任何帮助：预训练和当前任务无关(?)，而且模型表达力没有变
- ▶ 预训练是深度学习最重要发明之一
 - ▶ 例子：从一个字预测出词语和预测情感没关系
 - ▶ 现实：预测词语表示了对语义的理解，所以对预测情感有帮助
 - ▶ 从优化的角度来说：有利于优化

- ▶ 很多问题要 case-by-case 分析
- ▶ 重点：从不同角度出发（数学思维）
- ▶ 从不同角度看同一个问题：其他角度的进展也可以帮助解决这个问题

- 1 怎样学数学
- 2 机器学习的各种角度和建模流程
- 3 概率论和统计学复习
- 4 极大似然估计
- 5 贝叶斯估计和变分贝叶斯
- 6 矩阵和张量求导
- 7 总结

- ▶ 概率论是描述随机的语言
- ▶ 概率论分为朴素概率论和公理性概率论
- ▶ 主要讲朴素概率论

- ▶ 一维离散意味着可以直接讨论概率
- ▶ 一维离散意味着可以假设概率取值只是整数
- ▶ 例子：男 = 1, 女 = 2, 未知 = 3
 - ▶ $P(X < 3) = \dots$
 - ▶ $p(X = 1) = \dots$
 - ▶ $P(X \leq x) = \sum_{i \leq x} p(X = i)$, 或者用更标准的写法 $P(X \leq t) = \sum_{x \leq t} p(x)$

- ▶ 连续意味着可能性至少不是有限的
- ▶ 还是可以定义 $P(X \leq x)$
- ▶ 但是定义 $p(x)$ 的时候就有问题了

思考：为什么？

- ▶ 在给定一个连续变量时，只能定义 $P(X \leq m) = \int_{-\infty}^m p(x) dx$
- ▶ 虽然离散和连续的定义有所不同，但是积分本身就是一种非常复杂的加法
- ▶ $F_X(t) := P(X \leq t)$ 就是所谓的概率 Cumulative Distribution Function
- ▶ $p(x)$ 就是所谓的 Probability Density Function，**不是概率值**

习题：CDF 和 PDF 的转换

指数分布的 PDF 为 $\lambda e^{-\lambda x}$, $x \geq 0$, 求其 CDF;