Introduction

书籍推荐：

- 李航《统计学习方法》
  - 侧重统计机器学习
  - 老版本12章，共讲了10个算法
    - 感知机
    - K近邻
    - 朴素贝叶斯
    - 决策树
    - 逻辑回归和最大熵原理
    - 支持向量机
    - Boosting
    - EM算法(属于概率图模型)
    - 隐马尔代夫模型HMM(属于概率图模型)
    - 条件随机场(属于概率图模型)
  - 最新版似乎22章
- 周志华“西瓜书”
  - 各种算法都涉及，但讲的不深
- PRML（2006年的一本模式识别与机器学习）
  - 注重概率图模型
  - 回分神核稀、图混近采连、顺组
- MLAPP（以概率的视角看模型）
  - 百科全书
  - 注重概率图模型
- ESL（Elements of Statistic Learning，统计学习的基本元素/本质）
  - 侧重统计机器学习
- Deep Learning圣经（花书）

视频推荐

- 台大-林轩田：基石篇（如VC theory、正则化、线性模型等，值得一看），技法篇（SVM、决策树、随机森林等早期模型）
- 张志华老师：机器学习导论（频率派）、统计机器学习（贝叶斯的角度，如共轭等，偏数学），张志华课的数学推导很多
- 吴恩达老师CS229(斯坦福的课，不是cousera公开课，有很多数学推导)
- 徐亦达：主要讲概率模型（EM、MCMC、各种滤波算法、狄利克雷算法等，github上有Notes很全）

- 台大–李宏毅：机器学习、机器学习高阶（涉及到了自然语言处理等方面）

正文：

- 对概率的诠释有两大学派，频率派和贝叶斯派，前者发展出统计机器学习，后者发展出概率图模型。
- 符号说明：

$$X_{N\times p}=(x_1,x_2,\cdots,x_N)^T_{N\times p}=\begin{bmatrix}x_{11} & \cdots & x_{1p}\\ x_{21} & \cdots & x_{2p}\\ \cdots & \cdots & \cdots\\ x_{N1} & \cdots & x_{Np}\end{bmatrix}$$

  这个记号表示有 N 个样本，每个样本都是 p 维向量。其中每个观测都是由 $p(x|\theta)$ 生成的。

频率派的观点

- $p(x|\theta)$中的 θ 是一个未知的常量。假设 $x_i \overset{i.i.d}{\sim} p(x_i|\theta)$ ，对于 N 个观测来说观测集的概率为 $p(X|\theta)\underset{iid}{=}\prod_{i=1}^{N}p(x_i|\theta))$ 。为了便于计算，加上 log 使乘法变成加法。
- 为了求 θ 的大小，采用最大对数似然估计MLE的方法：

$$\theta_{MLE}=\underset{\theta}{argmax}\log p(X|\theta)\underset{iid}{=}\underset{\theta}{argmax}\sum_{i=1}^{N}\log p(x_i|\theta)$$

- 频率派的基本流程是：提出概率模型，然后设计损失函数，最后优化。其最终演变的研究对象是优化方法

贝叶斯派的观点

- 贝叶斯派认为 $p(x|\theta)$ 中的 θ 不是一个常量，即这个 θ 满足一个预设的先验分布 $\theta\sim p(\theta)$ 。
- 先验概率： $p(\theta)$
- 后验概率： $p(\theta|X)$
- 似然概率： $p(X|\theta)$
- 通过贝叶斯定理将先验概率和后验概率联系起来：

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)} = \frac{p(X|\theta) \cdot p(\theta)}{\int_\theta p(X|\theta) \cdot p(\theta) d\theta} \propto p(X|\theta) \cdot p(\theta)$$

- 为了求 θ 的值，我们需要调整参数以满足最大后验概率估计MAP：

$$\theta_{MAP} = \underset{\theta}{argmax}\, p(\theta|X) = \underset{\theta}{argmax}\, p(X|\theta) \cdot p(\theta)$$

- 其中第二个等号是由于分母和 θ 没有关系，是一种简化表达。
- 最本质的贝叶斯估计是要求分母的积分，因此相关发展都是在研究如何求这个积分，解析解求不出来，可以用蒙特卡洛方法求数值积分解，如MCMC，从而发展出概率图模型
- 求解这个 θ 值后计算 $\frac{p(X|\theta) \cdot p(\theta)}{\int_\theta p(X|\theta) \cdot p(\theta) d\theta}$ ，就得到了参数的后验概率。其中 $p(X|\theta)$ 叫似然，是我们的模型分布。

- 得到了参数的后验分布后，我们可以将这个分布用于贝叶斯预测，本质是使用 θ 作为一个桥梁，找到新数据和原数据之间的关系：

$$\begin{aligned} p(x_{new}|X) &= \int_\theta p(x_{new}, \theta|X) d\theta \\ &= \int_\theta p(x_{new}|\theta) \cdot p(\theta|X) d\theta \end{aligned}$$

  - 其中积分中的 $p(x_{new}|\theta)$ 是模型，$p(\theta|X)$ 是后验分布。

小结

频率派和贝叶斯派分别给出了一系列的机器学习算法。频率派的观点导出了一系列的统计机器学习算法而贝叶斯派导出了概率图理论。在应用频率派的 MLE 方法时最优化理论占有重要地位。而贝叶斯派的算法无论是后验概率的建模还是应用这个后验进行推断时积分占有重要地位。因此采样积分方法如 MCMC 有很多应用。