电子科技大学
University of Electronic Science and Technology of China

Chap. 16
Learning Graphical Models: Overview

Junhua Chen, PengFei Xlao

Data Mining Lab,
Big Data Research Center, UESTC

目录：

1. 学习动机
2. 模型学习概念
3. 学习目标
4. 优化学习
5. 学习任务

那年那兔那些章节：

目前为止，兔子们讨论的大多数问题出发点都是一个给定的图模型。如：

推理中是给定图的结构和参数，来做一系列查询的。

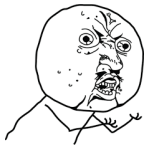
两种方法完成一个模型的查询任务（acquiring a model）。

- 一种是通过专家知识
- 一种是通过大量数据

冻鸡：

专家知识“手工”建模缺陷：

- 建模需要的经验太多
- 某些领域高级玩家少
- 计划赶不上变化



示例示例:

大数据嘛！搜集大量实例嘛！医学诊断—病例（实际上是我们的观测数据或证据）。

“模型学习”是什么gui:

这些实例，可以通过孤立（频率学派），或者混合先验知识（B学派），来为潜在的分布构建一个好的模型。

模型学习：

具体一些：

假定领域的潜在分布为P^*，由某个网络模型M^*诱导。给定从P^*中采样的M个iid样本数据集$D = \{d[1], \ldots, d[M]\}$。给定一族模型，并且任务是在这族模型中学习一个定义分布$P_{\tilde{M}}$的模型$\tilde{M}$。

希望

- 学习一个固定结构的模型参数
- 学习模型的一些或者所有结构
- 学习模型的置信估计

学习目标：

成为最接近P^*的男人：

密度估计：构建模型$\tilde{M}$，使$\tilde{P}$接近于生成分布P^*。
评价近似模型$\tilde{M}$的好坏！--相对熵距离：

$$D\left(P^*||\tilde{P}\right) = E_{\xi\sim P^*}\left[\log\frac{P^*(\xi)}{\tilde{P}(\xi)}\right]$$

命题16.1 对χ上的任意分布P，P'：

$$D(P||P') = -H_p(\chi) - E_{\xi\sim P}[\log P'(\xi)]$$

考虑$D\left(P^*||\tilde{P}\right)$，对近似模型的评估在第二部分$E_{\xi\sim P^*}[\log\tilde{P}(\xi)]$，使$D$最小即使$E_{\xi\sim P^*}$最大。$E_{\xi\sim P^*}[\log\tilde{P}(\xi)]$称为期望的对数似然expected log-likelihood。直观的，$\tilde{M}$对真实分布采样的ξ赋予的概率越大，越能反映它是该分布。

注意：由于$-H_{P^*}(\chi)$的不可知，期望对数似然只能用于模型之间的比较。

最大与最小:

更普遍的，我们感兴趣的是数据的似然，给定模型M，数据D，似然为$P(D:M)$，或者它的对数形式$\ell(D:M) = \log P(D:M)$。

对数损失：损失函数$loss(\xi:M)$度量了模型M在实例ξ上的损失，当实例从某个分布P^*中采样时，目标是找到使期望损失最小的模型：

$$E_{\xi \sim P^*}[loss(\xi:M)]$$

当然，P^*是不知道的，所以一般用平均经验风险：

$$E_D[loss(\xi:M)] = \frac{1}{|D|}\sum loss(\xi:M)$$

学习目标：

如何预测:

✓ 给定一些列变量X，和一个预测变量集Y

一个具体的预测为：

$$h_{\tilde{p}}(x) = \text{argmax}_y \tilde{P}(y|x)$$

对此条件概率查询，我们采用以下损失函数/目标：

$$E_{(x,y)\sim P^*}[\log \tilde{P}(y|x)]$$

称为条件似然conditional likelihood。

优化学习：

学习任务→优化问题:

✓ 一个假设空间hypothesis space，即，一个候选模型的集合，
✓ 一个目标函数objective function和一个定量刻画模型好坏的准则。

学习任务可以描述为：
在模型类中找到得分最高的模型。

- 学习目标中提到，模型的优化目标是损失函数$E_{\xi\sim P^*}[loss(\xi:M)]$，然而我们不知道$P^*$，所以一般用经验风险替代$E_D[loss(\xi:M)]=\frac{1}{|D|}\sum loss(\xi:M)$。

过拟合，偏差，方差:

- 考虑用数据D来定义经验分布$\hat{P}_D$：

$$\hat{P}_D(A) = \frac{1}{M}\sum_m 1\{\xi[m]\epsilon A\}$$

事件A的概率为A在训练数据中的得分。当$M \to \infty$，$\hat{P}_D(A)$将依概率收敛于$P^*(A)$。

然而，利用经验分布带来了无法避免的过拟合问题。

举个栗子：
假如有100个二值变量，2^{100}种可能；
现在观察到1000个实例；
那么，经验分布会给其中1000个变量赋0.001的概率，其他为0。

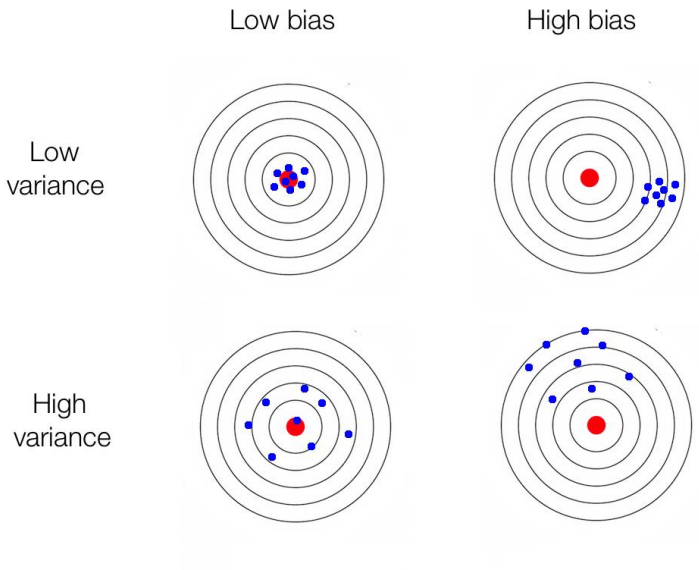
过拟合，偏差，方差:

幼齿，使用经验损失作为真实损失由很大的问题。倾向于过拟合。

方差-偏差都熟悉了，不罗嗦

在概率这里，可以总结为:

- 假设空间小→表达能力差，高偏差，低方差；
- 假设空间大→需要更多的实例，低偏差，高方差。



PAC界:

那么，在一个给定学习过程L（假设空间+目标函数），要多大的数据才能学习/找到一个误差较低的模型呢？--PAC界：

令$\epsilon > 0$是近似参数，$\delta > 0$是置信参数。那么对于足够大的M，我们有

$$P_M^*\left(\left\{D: D\left(P^* || P_{M_{L(D)}}\right) \leq \epsilon\right\}\right) \geq 1-\delta$$

即，对于足够大的M有：对于大多数从P^*中采样的规模为M的数据集D，学习过程用于D，将学习到P^*的一个比较准确的近似。达到这个结果的M的数目称为样本复杂性sample complexity。这类结果称为PAC界。

PS：PAC界只有当P^*在假设空间中才能获得。

优化学习：

判别式与生成式:

- 判别式discriminativel training：
目标使$\tilde{P}(\mathrm{Y}|\mathrm{X})$接近于$P(Y|X)$。

- 生成式generative training：
目标使$\tilde{M}$接近于总体联合分布$P^*(X,Y)$。

- **很有用的直觉：**
  - 一般，生成模型有较大的偏差—他们对分布有更多的假设（X独立性假设），判别模型仅关于Y和它们对X的依赖性做了独立性假设。
  - 或者，生成模型定义了$\tilde{P}(\mathrm{Y},\mathrm{X})$，并且需导出$\tilde{P}(\mathrm{X})$，$\tilde{P}(\mathrm{Y}|\mathrm{X})$，为了获得$P^*$的好的拟合，必须调整同时或者三者的最优。而判别模型只需$P^*(Y|X)$拟合更好。

学习任务：

抽象的学习过程:

- 输入：

某些先验知识或者关于$\tilde{M}$的约束。
从P^*中独立同分布采样的数据实例的集合D。

- 输出：

可能包括结构，参数或者两者都包含的模型$\tilde{M}$。

学习任务：

学习任务的三个变化方向：

- 模型输出：
  贝叶斯，马尔科夫。
- 模型限制：
  给定图结构，学习参数—数值优化问题；
  结构可能未知，额外的结构选择优化问题；
  部分变量已知，？？？
- 数据可观测性：
  数据完备；
  数据不完备：
  数据含隐变量；

总结：

- 模型学习的概念；
- 学习目标，为何最大/小就可以近似P^*；
- 学习优化，不可避免的过拟合问题，如何折中；
- PAC界，判别式&生成式的直觉；
- 不同类学习任务。

电子科技大学

University of Electronic Science and Technology of China

Thanks!

Data Mining Lab,
Big Data Research Center, UESTC