第九章 概率模型

第 23 讲 建模的概率思想

黄定江

DaSE @ ECNU
djhuang@dase.ecnu.edu.cn

1. 23.1 生成模型和判别式模型
2. 23.2 参数化模型和参数估计
3. 23.3 非参数模型和非参数估计
4. 23.4 概率图模型
5. 23.5 统计决策理论
6. 23.6 统计推断的基本概念

23.1.1 引言

在本讲中我们将讨论一些概率方法。在这些模型里我们会对概率函数进行估计，并使用这些概率函数来指导我们的决策 (例如分类、回归)。这是一个很大的话题，并且在该领域中存在着大量的方法。本讲只会讨论一些最基本的概念和方法, 并简要介绍其他方法. 本讲的主要目是介绍一些重要的概念和方法以及进行推断和决策的概率路线。

23.1.1 分布与推断

令 $p_{(}X, Y)$ 表示 X 与 Y 的**联合**分布。由于我们假设 Y 是可以基于 X 按照某种方式进行预测的，X 和 Y 之间必定存在着某种联系。换言之，X 与 Y 不可能是独立的——这意味着我们应该期望

$$p_{X,Y}(X, Y) \neq p_X(X)p_Y(Y)$$

相反，如果我们有

$$p_{X,Y}(X, Y) = p_X(X)p_Y(Y)$$

那么知道 X 对于预测 Y 一点帮助都没有，我们也无法学习到任何有意义的模型。

23.1.1 分布与推断

- 边缘分布 $p_X(x)$ 度量的是在不考虑 Y 的影响（或者说，将 Y 的影响从联合分布中通过积分移除出去）时，数据 X 的密度函数。它被称为**边缘似然** (marginal likelihood)。
- 在不考虑 X（或者其尚未被观测到）时，边缘分布 $p_Y(y)$ 是关于 Y 的**先验分布** (prior distribution)。它反映的是我们在观测到任何输入之前就已经（如通过领域知识）了解到的那些关于 Y 的先验知识。

23.1.1 分布与推断

- 在我们观测到 X 之后，由于 X 和 Y 之间的联系，我们可以更准确地估计 Y 的值。也就是说，$p_{Y|X}(Y \mid X)$（可以简写为 $p(Y \mid X)$，如果其意义可以从上下文中安全地推理出来）是关于 Y 的、比 $p_Y(Y)$ 更好的估计。这种分布被称为**后验分布** (posterior distribution)。当给定更多的**证据**（X 的样本）时，可根据其更新我们对 Y 的**信念** (belief)。更新后的信念，即后验或条件分布 $p(Y \mid X)$ 将作为我们在给定 X 时对 Y 的最佳估计。
- 使用证据来更新信念（即更新后验分布）的过程被称为**概率推断** (probabilistic inference)。我们还需要决定在获得后验之后能做些什么，因此**决策** (decision) 过程紧随其后。分类是一种经典的决策类型。

23.1.1 生成式模型和判别式模型

- 如果直接对条件/后验分布 $p(Y \mid X)$ 进行建模，那么这是一个**判别式模型** (discriminative model)。但判别式模型不能采样（或生成）得到一个服从潜在联合分布的样本对 (x, y)。在一些应用中，生成一个样本是很重要的。因此，我们可以对联合分布 $p(X, Y)$ 进行建模，这就导致了**生成式模型** (generative model)。
- 就分类而言，通常会转而对先验分布 $p(Y)$ 和类条件分布 $p(X \mid Y)$ 进行建模。由于

  $$p(X, Y) = p(Y)p(X \mid Y)$$

  这相当于对 $p(X, Y)$ 进行建模。

23.1.1 生成式模型和判别式模型

- 当从联合分布中进行采样 (即从联合分布中产生实例) 的能力不重要时，判别式模型是适用的，并且在实践中它通常比生成式模型具有更高的分类精度。但是，如果我们的目标是对数据的生成过程进行建模而非分类的话，一个生成式模型就是必要的。
- 我们不一定非要概率式地来解释这个世界。在不考虑概率的情况下，直接找到分类边界 (也被称为判别函数 (discriminant function)) 有时甚至能比判别式模型产生更好的结果。

现在看起来我们只需要估计 $p(Y|X)$，并且这个分布自身就能给我们提供足够的信息来 (在给定了来自 X 的证据时) 做出与 Y 相关的决策。但是，仍有诸多问题尚未解决，例如：

- 我们是否需要使用贝叶斯定理来估计 $p(Y|X)$，即首先估计 $p(X|Y)$ 和 $p(Y)$ 呢？这等价于估计联合概率 $p(X, Y)$。是否还存在其他的选择呢?
- 我们该如何表示分布呢?
- 我们该如何估计分布呢?

接下来我们将讨论针对这些问题的几个重要选项。参数估计与非参数估计、概率图模型。并介绍相关的决策理论。

23.2.1 参数模型的定义

一般地, 参数模型具有如下形式:

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\},$$

其中,θ 表示在参数空间 Θ 中取值的未知参数 (或参数向量)。如果 θ 是向量，但仅关心其中的一个元素的时候，则称其他参数为冗余参数。

23.2.1 参数模型举例

例 1

(一维参数估计）令 $X_1, \cdots, X_n$ 为相互独立的 *Bernoulli*(p) 观察值，问题是如何估计参数 p。

例 2

(二维参数估计）假设 $X_1, \cdots, X_n \sim F$ 并假设 *PDF*$f \in \mathfrak{F}$，其中 $\mathfrak{F}$ 满足高斯分布。这种情况下就有两个参数 μ 和 σ，目标是根据数据去估计这两个参数，如果仅关心估计 μ 的值，则 μ 就是感兴趣的参数而 σ 就是冗余参数。

23.2.1 参数化模型估计的方法

参数化模型估计的方法一般有：

- 矩估计
- 极大似然估计
- 极大后验估计
- 贝叶斯推断

这些方法将在下一讲进行介绍。

23.3.1 非参数模型含义

一般地，参数模型具有如下形式:

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\},$$

非参数模型指一些不能用有限个参数表示的 $\mathfrak{F}$，例如 $\mathfrak{F}_{\text{所有}} = \{\text{ 所有 CDF}\}$ 就是非参数模型。

23.3.1 非参数模型举例

例 3

(*CDF* 的非参数估计）令 $X_1, \cdots, X_n$ 是来源于 *CDF* 为 F 的独立观察值，问题是在假设 $F \in \mathfrak{F}_{\text{所有}}\{$ 所有 *CDF*$\}$ 的前提下如何去估计 F.

例 4

(非参数密度估计）令 $X_1, \cdots, X_n$ 是来源于 *CDF* 为 F 的独立观察值，令 $f = F'$ 为 *PDF*。假设要估计 *PDF*f. 如果仅假设 $F \in \mathfrak{F}_{\text{所有}}$ 是不可能估计 f 的，需要假设 f 的光滑性，例如，假设 $f \in \mathfrak{F} = \mathfrak{F}_{\text{DENS}} \cap \mathfrak{F}_{\text{SOB}}$，其中，$\mathfrak{F}_{\text{DENS}}$ 表示所有密度函数的集合

$$\mathfrak{F}_{SOB} = \{f\colon \int \left(f''(x)\right)^2 dx < \infty\} \tag{1}$$

集合 $\mathfrak{F}_{SOB}$ 称为**索伯列夫空间**（*Sobolev space*），它表示一系列“波动不大”的函数的集合。

23.3.1 非参数模型举例

例 5

(函数的非参数估计) 令 $X_1, ..., X_n$ F。假定要在仅假设 μ 存在的条件下去估计 $\mu = T(F) = \int xdF(x)$，通常情况下，任何 F 的函数称为**统计泛函**，其他一些统计泛函的例子有方差 $T(F) = \int x^2 dF(X) - (\int xdF(x))^2$，中位数 $T(F) = F^{-1}(1/2)$。

23.3.1 非参数化模型估计的方法

非参数化模型估计的方法一般有：

- 直方图估计
- 核密度估计
- 非参数回归估计
- CDF 和统计泛函的估计

23.4.1 概率图模型

- 机器学习算法经常会涉及多元随机向量的概率分布
- 如果采用单个函数来描述整个随机变量的联合分布是非常低效的 (无论是计算上还是统计上)，因为这些随机变量中涉及到的直接相互作用通常只介于非常少的变量之间的。
- 利用随机变量之间的条件独立性关系，可以将随机变量的联合分布分解为一些因式的乘积，得到简洁的概率表示。
- 我们可以采用图论中的“图”的概率来表示这种分解，得到概率图模型：图中的节点表示随机变量，边表示随机变量之间的直接作用。
- 有向图和无向图均可以用于概率表示。

23.4.2 有向图

- 有向图模型 (directed graphical models,DGMs) 使用带有有向边的图，用条件概率分布来表示分解：每个随机变量 x_i 都包含着一个影响因子，这些影响因子被称为 x_i 的父节点，记为 $Pa(x_i)$：

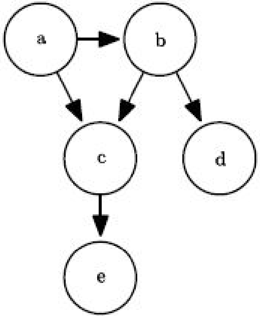
$$p(\mathbf{x}) = \prod_i p(x_i | Pa(x_i))$$

23.4.2 有向图

- 例：图中对应的概率分布可以分解为

$$p(a,b,c,d,e)=p(a)p(b|a)p(c|a,b)p(d|b)p(e|c)$$

-从图模型可以快速看出此分布的一些性质：如 a 和 c 直接相互影响，但 a 和 e 只有通过 c 间接相互影响



23.4.3 无向图

- 无向图模型 (undirected graphical models,UGM) 使用带有无向边的图，它将联合概率表示分解成一组函数的乘积
- 图中任何满足两两之间有边连接的顶点的集合被成为团 (clip)。每个团 $\mathtt{C}^{\mathtt{i}}$ 都伴随着一个因子 $\phi^i(\mathtt{C}^{\mathtt{i}})$
  -每个因子的输出都必须是非负的
  -但不像概率分布中那样要求因子的和/积为 1
- 随机向量的联合概率与所有这些因子的乘积成比例：

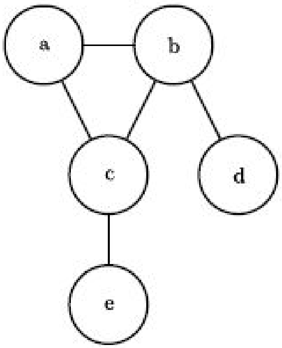
$$p(\mathbf{x}) = \frac{1}{Z}\prod_i \phi^i(\mathtt{C}^{\mathtt{i}})$$

- 其中归一化常数 Z 被定义为 ϕ 函数乘积的所有的求和或积分，使得这些乘积的求和为 1($p(\mathbf{x})$ 为一个合法的概率分布)

23.4.3 无向图

- 例：图中对应的概率分布可以分解为

$$p(a, b, c, d, e) = \frac{1}{Z}\phi^1(a, b, c)\phi^2(b, d)\phi^3(c, e)$$



23.4.3 无向图

- 从图中可以快速看出此分布的一些性质：如 a 和 c 直接相互影响，但 a 和 e 只有通过 c 间接相互影响
- 注意：这些模型表示的分解仅仅是描述概率分布的一种语言，它们不是相互排斥的概率分布族，任何概率分布都可以用这两种方式进行描述。

23.5.1 引言

- 前面已经考虑了几种点估计，如矩估计、极大似然估计、和极大后验估计。事实上，还有许多其他的估计方法。如何选择它们？答案在决策理论中找，它是比较统计过程的正规理论。
- 考虑参数空间 Θ 中的参数 θ。令 $\hat{\theta}$ 是 θ 的估计。在决策理论的语言中，点估计有时称为决策规则，决策规则可能的值称为行动。

23.5.1 常见损失函数

用损失函数 $L(\theta,\widehat{\theta})$ 来度量 θ 和 θ 的离散程度。正式地,L 把 $\Theta\times\Theta$ 映射到 $\mathbf{R}$。下面列出了一些损失函数:

$$L(\theta,\widehat{\theta})=(\theta-\widehat{\theta})^2 \quad \text{平方损失,}$$

$$L(\theta,\widehat{\theta})=|\theta-\widehat{\theta}| \quad \text{绝对损失,}$$

$$L(\theta,\widehat{\theta})=|\theta-\widehat{\theta}|^p \quad L_p \text{ 损失,}$$

$$\text{当}\theta=\widehat{\theta}\text{ 时, } L(\theta,\widehat{\theta})=0, \quad \text{当}\theta\neq\widehat{\theta}\text{ 时为}1 \quad 0-1 \text{ 损失,}$$

$$L(\theta,\widehat{\theta})=\int\log\left(\frac{f(x;\theta)}{f(x;\widehat{\theta})}\right)f(x;\theta)\mathrm{d}x \quad \text{Kullback-Leibler 损失.}$$

23.5.1 平均风险

为了衡量一个估计，可以用平均风险或损失来估计。

定义 1

估计 $\hat{\theta}$ 的风险为

$$R(\theta, \widehat{\theta}) = \mathbb{E}_{\theta}(L(\theta, \widehat{\theta})) = \int L(\theta, \widehat{\theta}(x)) f(x; \theta) \mathrm{d}x$$

当损失函数为平方误差时，风险是均方误差 MSE：

$$R(\theta, \widehat{\theta}) = \mathbb{E}_{\theta}(\theta - \widehat{\theta})^2 = \mathrm{MSE} = \mathbb{V}_{\theta} + \mathrm{bias}_{\theta}^2(\widehat{\theta})$$

23.6.0 点估计

点估计指对感兴趣的某一单点提供“最优估计”。感兴趣的点可以是参数模型、分布函数 F、概率密度函数 f 和回归函数 r 等中的某一参数，或者可以是对某些随机变量的未来值 Y 的预测。

一般地，令 $X_1, \ldots, X_n$ 为从某分布得来的 n 个 IID 数据点，参数 θ 的点估计 $\hat{\theta}_n$ 是 $X_1, \ldots, X_n$ 的函数：

$$\hat{\theta}_n = g(X_1, \ldots, X_n), \tag{2}$$

估计量的偏差定义为

$$bias(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta. \tag{3}$$

如果 $\mathbb{E}_\theta(\hat{\theta}_n) = \theta$，则称 $\hat{\theta}_n$ 是**无偏的。**

23.6.0 点估计

定义 2

如果 $\hat{\theta}_n \xrightarrow{P} \theta$，则参数 θ 的点估计 $\hat{\theta}_n$ 是**相合的。**

$\hat{\theta}_n$ 的分布称为**抽样分布**，$\hat{\theta}_n$ 的标准差称为**标准误差**，记为 se，

$$se = se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}. \tag{4}$$

通常，标准误差依赖于未知分布 F，在另外一些情况下，se 是未知量，但通常去估计它, 估计的标准误记为 $\hat{se}$。

23.6.0 点估计举例

例 6

令 $X_1, \ldots, X_n \sim Bernoulli(p)$，$\hat{p}_n = n^{-1}\sum_i X_i$，则 $\mathbb{E}(\hat{p}_n) = n^{-1}\sum_i \mathbb{E}(X_i) = p$，所以 $\hat{p}_n$ 是无偏的，标准误差为 $se = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$，估计的标准误差为 $\hat{s}e = \sqrt{\hat{p}(1-\hat{p})/n}$

点估计的质量好坏有时用**均方误差**或 MSE 来评价，均方误差定义为

$$MSE = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2. \tag{5}$$

需注意 $\mathbb{E}_\theta(\cdot)$ 是关于如下分布的期望而不是关于 θ 分布的平均，该分布由数据得来，具体见下

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta). \tag{6}$$

23.6.0 MSE

定理 1

MSE 可写成如下形式:

$$MSE = bias^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n).\mathrm{R} \tag{7}$$

证明.

令 $\bar{\theta}_n = E_\theta\left(\hat{\theta}_n\right)$，则

$$\begin{aligned}
\mathbb{E}_\theta\left(\hat{\theta}_n - \theta\right)^2 &= \mathbb{E}_\theta\left(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta\right)^2 \\
&= \mathbb{E}_\theta\left(\hat{\theta}_n - \bar{\theta}_n\right)^2 + 2\left(\bar{\theta}_n - \theta\right)\mathbb{E}_\theta\left(\hat{\theta}_n - \bar{\theta}_n\right) + \mathbb{E}_\theta\left(\bar{\theta}_n - \theta\right)^2 \\
&= \left(\bar{\theta}_n - \theta\right)^2 + \mathbb{E}_\theta\left(\hat{\theta}_n - \bar{\theta}_n\right)^2 \\
&= \mathrm{bias}^2\left(\hat{\theta}_n\right) + V_\theta\left(\hat{\theta}_n\right)
\end{aligned}$$

□

23.6.0 点估计

定理 2

如果 $bias \to 0$ 且当 $n \to \infty$ 时 $se \to 0$，则 $\hat{\theta}_n$ 是相合的，即 $\hat{\theta}_n \xrightarrow{P} \theta$。

证明.

如果 $bias \to 0$ 且 $se \to 0$, 则根据定理 6.9 有 $MSE \to 0$, 推出 $\hat{\theta}_n \xrightarrow{qm} \theta$(定义 5.2)，再根据定理 5.4 的 (a) 即得证。 $\square$

23.6.0 点估计

例 7

回到抛硬币的例子中，因为 $E_p(\hat{p}_n) = p$，所以 *bias* $= p - p = 0$，$se = \sqrt{p(1-p)/n} \to 0$，因此 $\hat{p}_n \xrightarrow{P} p$，即 $\hat{p}_n$ 是一致估计量。

定义 3

$$\frac{\hat{\theta}_n - \theta}{se} \sim +N(0,1)$$

则称估计量 $\hat{\theta}_n$ 是渐进正态的。

23.6.0 置信区间

参数 θ 的 $1-\alpha$ 置信区间为区间 $C_n=(a,b)$，其中，$a=a\left(X_1,\cdots,X_n\right), b=b\left(X_1,\cdots,X_n\right)$ 是数据的函数，满足

$$P_\theta\left(\theta\in C_n\right)\geqslant 1-\alpha,\quad \theta\in\Theta$$

其含义为 (a,b) 覆盖参数的概率为 $1-\alpha$，称 $1-\alpha$ 为置信区间的覆盖。

23.6.0 置信区间举例

例 8

报纸每天都会报道民意调查地结果。例如，报道称“有 *83%* 的公众对飞行员随身配备真枪飞行的做法表示赞同”，通常你还会看到诸如这样的陈述“该调查有 *95%* 的概率在 *4* 个百分点的范围内变动”。意思就是赞同飞行员随身配备真枪飞行的做法的人数所占的比例 p 的 *95%* 的置信区间是 $83\% \pm 4\%$，如果以后都按这种方式建立置信区间，则有 *95%* 的区间将包括真实的参数值，即使每天估计的量不同（不同的民意测验），这一结论也是正确的。

23.6.0 置信区间举例

例 9

置信区间不是参数 θ 的概率陈述容易让人迷惑，考察（*Berger and Wolpert*，*1984*）中的一个例子，令 θ 为一固定且已知的实数，X_1, X_2 为独立随机变量，满足 $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$，定义 $Y_i = \theta + X_i$ 并假设只观察到了 Y_1 和 Y_2，定义如下 “置信区间”（该区间其实只包括了一个点）：

$$C = \begin{cases} Y_1 - 1, & Y1 = Y2 \\ (Y_1 + Y_2)/2, & Y_1 \neq Y_2 \end{cases}$$

可以验证不管 θ 为多少都有 $\mathbb{P}_\theta(\theta \in C) = 3/4$，所以这是一个 *75%* 的置信区间，假设重做试验得到 $Y_1 = 15, Y_2 = 17$，则以上的 *75%* 的置信区间为 *{16}*，然而可以确信 $\theta = 16$，如果希望对 θ 进行概率陈述，可能有 $\mathbb{P}(\theta \in C | Y_1, Y_2) = 1$，这与称 *{16}* 是 *75%* 的置信区间并没有什么矛盾，但它并不是关于 θ 的置信区间。第 *11* 章将介绍当 θ 为随机变量时的贝叶斯方法以及关于 θ 的概率陈述，特别地，将做这样的陈述 “在给定数据的情况下，θ 在 C_n 中的概率为 *95%*”，然而，贝叶斯区间指的是可信度的可能性，一般来讲，贝叶斯区间不满足有 *95%* 的概率会覆盖参数。

23.6.0 置信区间举例

例 10

在抛硬币的试验中，令 $C_n = (\hat{p}_n - \epsilon, \hat{p}_n + \epsilon)$，其中 $\epsilon^2 = \log(2/\alpha)/(2n)$，由霍夫不等式得，对任意 p

$$\mathbb{P}(p \in C_n) \geq 1 - \alpha.$$

因此，C_n 是 $1-\alpha$ 置信区间。
就像前面提到的那样，点估计通常具有极限正态分布的，这意味着式成立，即 $\hat{\theta}_n \approx N(\theta, \hat{se}^2)$，在这种情况下，可以通过如下方式建立（近似）置信区间。

23.6.0 基于正态的置信区间

定理 3

(基于正态的置信区间) 假设 $\hat{\theta}_n \approx N(\theta, \hat{se}^2)$，令 Φ 为标准正态分布的 *CDF*，$z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$，即 $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2, \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$，其中 $Z \sim N(0,1)$，令

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{se}, \hat{\theta}_n + z_{\alpha/2}\hat{se}),$$

则

$$\mathbb{P}_\theta(\theta \in C_n) \sim 1-\alpha.$$

23.6.0 基于正态的置信区间证明

证明.

令 $Z_n = (\hat{\theta}_n - \theta)/\hat{se}$，根据假设由 $Z_n \to Z$，其中，$Z \sim N(0,1)$，因此

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta(\hat{\theta}_n - z_{\alpha/2}\hat{se} < \theta < \hat{\theta}_n + z_{\alpha/2}\hat{se}) && (8)\\
&= \mathbb{P}_\theta(-z_{\alpha/2} < \frac{\hat{\theta}_n - \theta}{\hat{se}} < z_{\alpha/2}) && (9)\\
&\to \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) && (10)\\
&= 1 - \alpha && (11)
\end{aligned}$$

□

对于 95% 的置信区间，$\alpha = 0.05, z_{\alpha/2} = 1.96 \approx 2$，可以得到 95% 的置信区间为 $\hat{\theta}_n \pm 2\hat{se}$

23.6.0 假设检验

在**假设检验**中，从缺省理论，即**原假设**开始，通过数据是否提供显著性证据来支持拒绝该假设，如果不能拒绝，则保留原假设。

例 11

(检验硬币是否均匀）令

$$X_1, \cdots, X_n \sim Bernoulli(p)$$

为 n 此独立的硬币投掷结果，假设要检验硬币是否均匀，令 H_0 表示硬币是均匀的假设，H_1 表示硬币不是均匀的假设，H_0 称为**原假设**，H_1 称为**备择假设**，可以将假设写成

$$H_0 : p = 1/2 \qquad H_1 : p \neq 1/2.$$

如果 $T = |\hat{p}_n - (1/2)|$ 的值很大，则有理由拒绝 H_0，当详细讨论假设检验的时候，将会确定出拒绝 H_0 的精确 T 值。