目录

第十四章　概率图模型

- 图模型
  - 有向图模型
    - 静态贝叶斯网络
    - 动态贝叶斯网络
      - 隐马尔可夫模型 (Hidden Markov Model)
      - 卡尔曼滤波器 (Kalman Filter)
      - ···
  - 无向图模型 → 马尔可夫网络
    - 吉布斯/玻尔兹曼机 (Gibbs/Boltzmann Machine)
    - 条件随机场 (Conditional Random Field)
    - ···

概率模型的学习即基于训练样本来估计变量分布的参数。**概率图模型** (probabilistic graphical model) 是一类用图来表达变量相关关系的概率模型。**马尔可夫链** (Markov Chain)：系统下一时刻的状态由当前状态决定，不依赖于以往的任何状态。

0.1　马尔可夫随机场

马尔可夫随机场 (Markov Random Field) 是典型的马尔可夫网 (无向图模型)。图中的每个节点表示一个或一组变量，节点之间的边表示两个变量之间的关系。还有一组定义在变量子集 (一般是团 (clique) 或者极大团 (maximal clique)) 上的**势函数** (potential functions)，亦称为因子 (factor)，势函数是非负的，主要用于定义概率分布函数。

对于 n 个变量 $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$，所有团构成的集合为 $\mathcal{C}$，与团 $Q \in \mathcal{C}$ 对应的变量集合记为 $\boldsymbol{x}_Q$，则**联合概率** $P(\boldsymbol{x})$ 定义为：

$$P(\boldsymbol{x}) = \frac{1}{Z} \prod_{Q \in \mathcal{C}} \psi_Q(\boldsymbol{x}_Q)$$

其中 ψ_Q 为与团 Q 对应的势函数，用于对团中的变量关系进行建模，$Z = \sum_{\boldsymbol{x}} \prod_{Q \in \mathcal{C}} \psi_Q(\boldsymbol{x}_Q)$ 为规范化因子，以确保 $P(\boldsymbol{x})$ 是被正确定义的概率。

为了减少计算开销，可以把联合概率基于极大团定义，$\mathcal{C}^*$ 为所有极大团构成的集合：

$$P(\boldsymbol{x}) = \frac{1}{Z^*} \prod_{Q \in \mathcal{C}^*} \psi_Q(\boldsymbol{x}_Q)$$

如果从节点集 A 中的节点到 B 中的节点都必须经过节点集 C 中的节点，则称节点集 A 和 B 被节点集 C 分离，C 称为“分离集”(separating set)。

马尔可夫随机场满足的性质：

- 全局马尔可夫性 (global Markov property)：给定两个变量子集的分离集，则这两个变量子集条件独立。(在给定分离集的条件下独立：$\boldsymbol{x}_A \perp \boldsymbol{x}_B | \boldsymbol{x}_C$)，令 A,B,C 分别对应单变量 x_A, x_B, x_C，则有：

$$P(x_A, x_B | x_C) = P(x_A | x_C) P(x_B | x_C)$$

- 局部马尔可夫性 (local Markov property)：给定某变量的邻接变量，则该变量条件独立于其他变量。令 V 为图的节点集,$n(v)$ 为节点 v 在图上的邻接节点，$n^*(v) = n(v) \cup \{v\}$，有 $\boldsymbol{x}_v \perp \boldsymbol{x}_{V \setminus n^*(v)} | \boldsymbol{x}_{n(v)}$
- 成对马尔可夫性 (pairwise Markov property)：给定所有其他变量，两个非邻接变量条件独立。令图的节点集和边集分别为 V 和 E，则对图中的两个节点 u 和 v，若 $\langle u, v \rangle \notin E$，$\boldsymbol{x}_u \perp \boldsymbol{x}_v | \boldsymbol{x}_{V \setminus \langle u,v \rangle}$

为了满足非负性，**势函数常被定义为指数函数**，即：

$$\psi_Q(\boldsymbol{x}_Q) = e^{-H_Q(\boldsymbol{x}_Q)}$$

其中 $H_Q(\boldsymbol{x}_Q)$ 是一个定义在变量 $\boldsymbol{x}_Q$ 上的实值函数，常见形式为：

$$H_Q(\boldsymbol{x}_Q) = \sum_{u,v \in Q, u \neq v} \alpha_{uv} x_u x_v + \sum_{v \in Q} \beta_v x_v$$

其中 α_{uv} 和 β_v 是参数，第一个加法项考虑每一对节点关系，第二个加法项考虑单节点。

0.2 条件随机场

条件随机场 (Conditional Random Field，简称 CRF) 是一种**判别式无向图模型**。令 $G = \langle V, E \rangle$ 表示节点与标记变量 $\boldsymbol{y}$ 中元素一一对应的无向图，y_v 表示与节点 v 对应的标记变量，$n(v)$ 表示节点 v 的邻接节点，若图 G 的每个变量 y_v 都满足马尔可夫性 (每个节点都仅受邻接节点影响)，即：

$$P(y_v | \boldsymbol{x}, \boldsymbol{y}_{V \setminus \{v\}}) = P(y_v | \boldsymbol{x}, \boldsymbol{y}_{n(v)})$$

则称 $(\boldsymbol{y}, \boldsymbol{x})$ 构成一个条件随机场。

CRF 的特点是假设输出随机变量构成马尔可夫随机场。条件随机场和马尔可夫随机场均使用团上的势函数来定义概率，两者在形式上没有显著区别；但**条件随机场处理的是条件概率，而马尔可夫随机场处理的是联合概率**。

在条件随机场中，通过选用指数函数并引入特征函数 (feature function)，条件概率被定义为：

$$P(\boldsymbol{y} | \boldsymbol{x}) = \frac{1}{\boldsymbol{Z}} \exp \left(\sum_j \sum_{i=1}^{n-1} \lambda_j t_j(y_{i+1}, y_i, \boldsymbol{x}, i) + \sum_k \sum_{i=1}^{n} \mu_k s_k(y_i, \boldsymbol{x}, i) \right)$$

其中 $t_j(y_{i+1}, y_i, \boldsymbol{x}, i)$ 是定义在观测序列的两个相邻位置上的转移特征函数 (transition feature function)，用于刻画相邻标记变量之间的相关关系以及观测序列对它们的影响。$s_k(y_i, \boldsymbol{x}, i)$ 是定义在观测序列的标记位置 i 上的状态特征函数 (status feature function)，用于刻画观测序列对标记变量的影响，λ_j 和 μ_k 为参数，Z 为规范化因子，用于确保式子是正确定义的概率。

0.3 精确推断

推断是指在基于概率图模型定义的联合概率分布，对目标变量的**边际分布** (marginal distribution) 或某些可观测变量为条件的**条件分布**进行推断。边际分布是指无关变量求和或积分后得到的结果。例如在马尔可夫网中，变量的联合分布被表示成极大团的势函数乘积，给定参数 Θ 求解某个变量 x 的分布，就变成对联合分布中其他无关变量进行积分的过程，称为**边际化** (marginalization)。**推断问题的核心就是如何高效地计算边际分布**。

精确推断是希望获得目标变量的边际分布或者条件分布的精确值；而近似推断是近似解。精确推断实质是动态规划算法，利用图模型所描述的条件独立性来削减计算量。精确推断方法有变量消去法和信念传播法。

0.3.1 变量消去法

变量消去法把多个变量的积的求和问题转化为对部分变量交替进行求积与求和的问题。这样子每次求积和求和都只在局部进行，仅与部分变量有关。但是如果需要计算多个边际分布，重复使用变量消去法会造成大量的冗余计算。

0.3.2 信念传播

信念传播 (Belief Propagation) 算法将变量消去法中的求和操作看做一个消息传递的过程，一个结点仅在接受到来自其他所有结点的消息后才能向另一个结点发送消息，且结点的边际分布正比于它所接收到的消息的承继。

0.4 近似推断

0.4.1 MCMC 采样

通过随机化采样的方式，用样本均值近似期望，因为计算概率的目的，就是基于概率计算期望。概率图模型中最常用的采样技术是：**马尔可夫链蒙特卡罗方法** (Markov Chain Monte Carlo, 简称 MCMC)。

给定连续变量 $x \in X$ 的概率密度函数 $p(x)$，x 在区间 A 中的概率可计算为：

$$P(A) = \int_A p(x)\mathrm{d}x$$

若有函数：$f : X \mapsto \mathbb{R}$，则可计算 $f(x)$ 的期望：

$$p(f) = \mathbb{E}_p[f(X)] = \int_x f(x)p(x)\mathrm{d}x$$

如果 x 不是单变量，而是一个高维多元变量 $\boldsymbol{x}$，且服从一个非常复杂的分布，那么上面的积分就很困难。MCMC 先构造出服从 p 分布的独立同分布随机变量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$，然后得到上面积分式子的无偏估计：

$$\hat{p}(f) = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}_i)$$

MCMC 方法的关键就在于通过构造“**平稳分布为 p 的马尔可夫链**”来产生样本：马尔可夫链运行时间足够长 (即收敛到平稳状态)，则此时产出的样本 $\boldsymbol{x}$ 近似服从于分布 p。判断到达平稳状态的方法：令从状态 $\boldsymbol{x}$ 转移到状态 $\boldsymbol{x}'$ 的概率为 $T(\boldsymbol{x}'|\boldsymbol{x})$，$t$ 时刻状态的分布为 $p(\boldsymbol{x}^t)$，则如果某个时刻马尔可夫链满足平稳条件：

$$p(\boldsymbol{x}^t)T(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t) = p(\boldsymbol{x}^{t-1})T(\boldsymbol{x}^t|\boldsymbol{x}^{t-1})$$

则称 $p(\boldsymbol{x})$ 是该马尔可夫链的平稳分布，且此时以收敛到平稳状态。

也就是说 **MCMC 方法先设法构造一条马尔可夫链，使其收敛至平稳分布恰为待估计参数的后验分布，然后通过这条马尔可夫链来产生符合后验分布的样本，并基于这些样本来进行估计。**

构造马尔可夫链的状态转移概率方法：**Metropolis-Hastings 算法** (简称 MH)。MH 算法会以一定的概率拒绝根据上一轮采样结果 $\boldsymbol{x}^{t-1}$ 采样获得的候选状态样本 $\boldsymbol{x}^*$，而直接选择上一轮的采样结果 $\boldsymbol{x}^{t-1}$ 作为当前这一轮的采样结果。

Algorithm 1 *

Metropolist-Hasting 算法

输入： 先验概率 $Q(\boldsymbol{x}^*|\boldsymbol{x}^{t-1})$

初始化 $\boldsymbol{x}^0$
for $t = 1, 2, \cdots$ **do**
  根据 $Q(\boldsymbol{x}^*|\boldsymbol{x}^{t-1})$ 采样出候选样本 $\boldsymbol{x}^*$;
  根据均匀分布从 (0,1) 范围内采样出阈值 u
  if $u \leq A(\boldsymbol{x}^*|\boldsymbol{x}^{t-1})$ [$A(\boldsymbol{x}^*|\boldsymbol{x}^{t-1})$ 是 $\boldsymbol{x}^*$ 被接受的概率] **then**
    $\boldsymbol{x}^t = \boldsymbol{x}^*$
  else
    $\boldsymbol{x}^t = \boldsymbol{x}^{t-1}$

9: **end if**
10: **end for**
输出: 采样出的一个样本序列：$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$

0.4.2 变分推断

变分推断通过使用已知简单分布来逼近需推断的复杂分布，并通过限制近似分布的类型，从而得到一种**局部最优、但具有确定解的近似后验分布**。

使用变分法时，最重要的考虑如何对隐变量进行拆解，以及假设各变量子集服从何种分布，然后用 EM 算法进行概率图模型的推断和参数估计。

0.5 话题模型

话题模型 (topic model) 是一种生成式有向图模型，主要用于处理离散型的数据 (如文本集合)，在信息检索、自然语言处理等领域有广泛应用。典型代表是**隐狄利克雷分配模型** (Latent Dirichlet Allocation，简称 LDA)。

“词”(word) 是待处理数据的基本离散单元，“文档”(document) 是待处理的数据对象，它由一组词组成。用文档和词进行数据表示的方式成为“词袋”(bag-of-words)。“话题”(topic) 表示一个概念，具体表示为一系列相关的词，以及它们在该概念下出现的概率。

LDA 认为每篇文档包含多个 (K) 个话题，用向量 $\Theta_t \in \mathbb{R}^K$ 表示文档 t 中所包含的每个话题的比例，$\Theta_{t,k}$ 即表示文档 t 中包含话题 k 的比例，具体由话题“生成”文档 t 的步骤:

1. 根据参数为 $\boldsymbol{\alpha}$ 的狄利克雷分布随机采样一个话题分布 Θ_t;
2. 按如下步骤生成文档中的 N 个词;
   - 根据 Θ_t 进行话题指派，得到文档 t 中词 n 的话题 $z_{t,n}$;
   - 根据指派的话题所对应的词频分布 $\boldsymbol{\beta}_k$ 随机采样生成词。