

概率图模型

1. 基本概念

概率模型 将学习任务归结于计算变量的 概率分布

推断：利用已知变量 推测未知变量 的分布

γ : 关心的变量集合, O : 可观测的变量, R : 其他变量.

$$\begin{cases} \text{生成式} & \rightarrow p(\gamma, R, O) \\ \text{判别式} & \rightarrow p(\gamma | R, O) \end{cases}$$

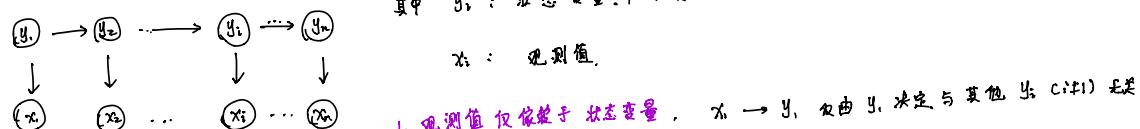
$\gamma \longrightarrow p(\gamma | O)$
推断

2. 利用概率图模型表达变量相关关系

{ 有向无环图：有向图模型、贝叶斯网

} 无向图：无向图模型、马尔可夫网

其中 y_i : 状态变量, 不可被观测, 又称隐变量



2. y_t 取决于 y_{t-1}

→ 所有变量的联合概率分布为 $p(x_1, y_1, \dots, x_n, y_n) = p(y_1) p(x_1 | y_1) \prod_{i=2}^n p(y_i | y_{i-1}) p(x_i | y_i)$

A: 状态转移矩阵

B: 观测矩阵

π : 初始状态概率

$$\lambda = [A, B, \pi]$$

- (1) 令 $t=1$, 根据 π , 选择 y_1
- (2) y_t 和 B 确定 x_t
- (3) y_t 和 A 确定 y_{t+1}
- (4) 若 $t < n$, 令 $t=t+1$, 回到第二步, 否则停止

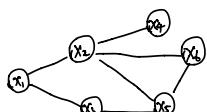
① 已知 $\{\pi, A, B, \dots, y_{n-1}\}$ 求 λ , 即 $p(x_n | \lambda)$

② 已知 λ, x 求 y 隐藏状态

③ 已知 x , 如何调整入量矩阵 $p(x | \lambda)$

3. 马尔可夫随机场

利用势函数定义概率分布函数。

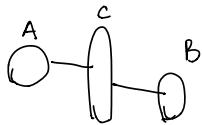


$$p(x) = \frac{1}{Z} \prod_{q \in C} \psi_q(x_q)$$

ψ_q : 势函数

z: 归一化因子

引入最大项，解决因个数过多而问题。



1. 全局马尔可夫性：若空两变量集的分离集，则子集条件独立。

2. 局部马尔可夫性：若空某变量的邻接变量，则该变量条件独立于其它变量。

3. 成对马尔可夫性：若空所有其他变量，两个非邻接变量条件独立。

C: 分离集，分集 A, B.

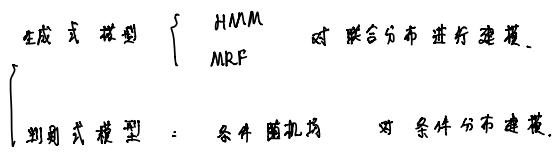
势函数：在偏好的变量上有最大取值，常用指教函数定义。

$$x_A \perp x_B \mid x_C$$

$$\psi_q(x_q) = e^{-H_q(x_q)}$$

$$\text{其中 } H_q(x_q) = \sum_{u \in q, u \neq v} a_{uv} x_u x_v + \sum_{v \in q} b_v x_v$$

4. 条件随机场：CRF



若图中每个变量 y_i 满足马尔可夫性，则 $p(y_0 | x, y_{1:n-1}) = p(y_0 | x, y_{n-1})$

则 (y, x) 构成一个条件随机场。

常用：链式条件随机场。

$$p(y|x) = \frac{1}{Z} \exp \left(\underbrace{\sum_{j=1}^{n-1} \lambda_j \phi_j(y_{i+j}, y_i, x_i)}_{\text{相邻标记变量}} + \underbrace{\sum_{i=1}^n \mu_i s_i(y_i, x_i)}_{\text{单个标记变量}} \right)$$

$\phi_j(y_{i+j}, y_i, x_i)$ ：观测序列中两个相邻标记位置上的转移特征函数，表示相邻标记的关系和观测序列对单影响。

$s_i(y_i, x_i)$ ：观测序列标记位置 i 上的状态特征函数，刻画序列对标记的影响。

5. 学习与推断。

直积化：对联合分布中其他无关变量进行积分

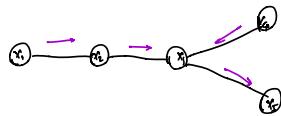
$$p(x_f | y_e) = \frac{p(x_e, x_f)}{\sum_{x_f} p(x_e, x_f)}$$

有效地计算边际分布

精确推断：变量消去、信念传播。

近似基团：MCMC 采样，变分推断。

(1) 变量消去法：



$$\begin{aligned}
 p(x_6) &= \sum_{x_5} p(x_6 | x_5) \sum_{x_4} p(x_4 | x_5) \sum_{x_3} p(x_3 | x_4) \sum_{x_2} p(x_2 | x_3) p(x_4 | x_2) \\
 &= \sum_{x_4} \sum_{x_5} \sum_{x_2} \sum_{x_3} p(x_6, x_5, x_4, x_3, x_2) \\
 &= m_{35}(x_5)
 \end{aligned}$$

变量消去法 把多个变量的积求和问题 转化为对部分变量交替进行求积求和，但会导致重复计算。

(2)

信念传播。

变量消去法： $m_{ij}(x) = \sum_x \psi(x, x_i) \prod_{k \neq i, j} m_{ki}(x_k)$

信念传播法：一个结点必须在接收到来自其他所有结点信息后才能 向另一结点发送信息，且结点的边度分布

正比于它所接收信息的乘积 即 $p(x_i) \propto \prod_{k \in \text{neighbors}} m_{ki}(x_k)$.

(3)

MCMC 采样，随机化近似。

关心概率分布 \rightarrow 计算期望 \rightarrow 抽样的方式 \rightarrow 常用采样方式 $\rightarrow \hat{P}[f] = \frac{1}{n} \sum_{i=1}^n f(x_i)$

$$\begin{aligned}
 E_p[f] &= \int f(x) p(x) dx \\
 f &= \frac{1}{n} \sum_{i=1}^n f(x_i) \quad \text{MCMC}
 \end{aligned}$$

构造服从 p 分布的独立同分布 随机变量 x_1, x_2, \dots, x_n

\rightarrow MCMC 构造平稳分布为 p 的马尔可夫链。

\rightarrow 判断链空的条件 $p(x^t) T(x^{t+1} | x^t) = p(x^{t+1}) T(x^t | x^{t+1})$

\rightarrow 产生符合后验分布的样本，利用这些样本进行估计。

x^{t+1} ：上一靴采样结果

x^* ：候选状态

$x^{t+1} \rightarrow x^*$ 移动概率 $\underbrace{Q(x^* | x^{t+1})}_{\text{先验}} \underbrace{A(x^* | x^{t+1})}_{\text{接受}}$

$$A(x^* | x^{t+1}) = \min\left(1, \frac{p(x^*) Q(x^{t+1} | x^*)}{p(x^{t+1}) Q(x^* | x^{t+1})}\right)$$

平均状态 $p(x^{t+1}) Q(x^* | x^{t+1}) A(x^* | x^{t+1}) = p(x^*) Q(x^{t+1} | x^*) A(x^{t+1} | x^*)$

吉布斯采样：

$\mathbf{x} = [x_1, x_2, \dots, x_N]$, 目标分布 $p(\mathbf{x})$

1. 随机选取变量 x_i

2. $p(x_i | x_{-i})$ 其中 $x_{-i} = [x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_N]$

3. 根据 $p(x_i | x_{-i})$ 对 x_i 采样，采样值代替原值。

(4)

蒙特卡洛推断：使用已知简单分布逼近复杂分布，限制近似分布类型，达到局部最优，通过解剖近似后验分布。

$$p(\mathbf{x} | \theta) = \prod_{i=1}^N p(x_i, z_i | \theta)$$

$$\ln p(\mathbf{x} | \theta) = \sum_{i=1}^N \ln \left\{ \sum_z p(x_i, z_i | \theta) \right\} \text{ 求 } p(z_i | x_i, \theta) \text{ 和 } \theta.$$

$$E: \theta^t \rightarrow p(z_i | x_i, \theta^t) \text{ 计算 } p(x_i, z_i | \theta)$$

$$M: \theta^{t+1} = \arg \max_{\theta} Q(\theta; \theta^t)$$

$$= \arg \max_{\theta} \sum_z \underbrace{p(z_i | x_i, \theta^t)}_{\text{与 } z_i \text{ 的真实后验分布相等}} \ln p(x_i, z_i | \theta)$$

与 z_i 的真实后验分布相等时 θ 可近似于对数似然函数。

$$\ln p(\mathbf{x}) = \int L(\theta) + KL(Q || P)$$

$$\text{其中 } L(\theta) = \int q(z) \ln \left\{ \frac{p(x, z)}{q(z)} \right\} dz, \quad KL(Q || P) = - \int q(z) \ln \frac{p(x, z)}{q(z)} dz.$$

$$\text{假设 } z_i \text{ 服从分布 } q(z_i) = \prod_{i=1}^N q_i(z_i) \quad \text{此时} \quad q(z_i) = \underbrace{\frac{\exp(E_{\theta^t} [\ln p(x, z)])}{\int \exp(E_{\theta^t} [\ln p(x, z)]) dz}}_{\text{与 } z_i \text{ 的真实后验分布相等}}$$

1. 对随机量 z_i ，假设子集分布形式。

结合 EM 算法进行模型推断和参数估计

此时 $E_{\theta^t} [\ln p(x, z)]$ 有闭式解，故推断随机量。

$$= \int \ln p(x, z) \prod_{i=1}^N q_i(z_i) dz, \quad \text{联合 } \ln p(x, z) \text{ 在 } z_i \text{ 之外的随机量} \rightarrow \text{平均值方法}.$$

2. 生成模型

LDA: 生成式模型

LDA: 生成狄利克雷分配模型。

$\theta_{i, k}$: 文档 i 中包含类 k 的比例。

语料: K , 文档: T , 字典: N .

生成文档 i : (1) 根据 α 的狄利克雷分布随机采样 θ_i

生成文档 i : (2) 根据 θ_i 选择话题类别，得到文档中词 j 的概率。

(3) 根据话题类别对应词分布 β_j 随机采样。

文档 $\rightarrow T \times N$

话题 $\rightarrow K \times N$

输入: 先验概率 $Q(x^* | x^{t-1})$.
过程:
1: 初始化 x^0 ;
2: for $t = 1, 2, \dots$ do
3: 根据 $Q(x^* | x^{t-1})$ 采样出候选样本 x^* ;
4: 根据均匀分布从 $(0, 1)$ 范围内采样出阈值 u ;
5: if $u \leq A(x^* | x^{t-1})$ then
6: $x^t = x^*$
7: else
8: $x^t = x^{t-1}$
9: end if
10: end for
11: return x^1, x^2, \dots

图 14.9 Metropolis-Hastings 算法

图 14.12 描述了 LDA 的变量关系，其中文档中的词频 $w_{t,n}$ 是唯一的已观测变量，它依赖于对这个词进行的话题指派 $z_{t,n}$ ，以及话题所对应的词频 β_k ；同时，话题指派 $z_{t,n}$ 依赖于话题分布 Θ_t ， Θ_t 依赖于狄利克雷分布的参数 α ，而话题词频则依赖于参数 η 。

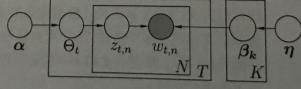


图 14.12 LDA 的盘式记法图

$$p(\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{t=1}^T p(\Theta_t | \boldsymbol{\alpha}) \prod_{k=1}^K p(\beta_k | \boldsymbol{\eta}) \left(\prod_{n=1}^N P(w_{t,n} | z_{t,n}, \beta_k) P(z_{t,n} | \Theta_t) \right), \quad (14.41)$$

其中 $p(\Theta_t | \boldsymbol{\alpha})$ 和 $p(\beta_k | \boldsymbol{\eta})$ 通常分别设置为以 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\eta}$ 为参数的 K 维和 N 维狄利克雷分布，例如

$$p(\Theta_t | \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \Theta_k^{\alpha_k - 1}, \quad (14.42)$$

其中 $\Gamma(\cdot)$ 是 Gamma 函数。显然， $\boldsymbol{\alpha}$ 和 $\boldsymbol{\eta}$ 是模型式(14.41)中待确定的参数。

给定训练数据 $\mathbf{W} = \{w_1, w_2, \dots, w_T\}$ ，LDA 的模型参数可通过极大似然法估计，即寻找 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\eta}$ 以最大化对数似然。

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{t=1}^T \ln p(w_t | \boldsymbol{\alpha}, \boldsymbol{\eta}). \quad (14.43)$$

但由于 $p(w_t | \boldsymbol{\alpha}, \boldsymbol{\eta})$ 不易计算，式(14.43)难以直接求解，因此实践中常采用变分法来求取近似解。

若模型已知，即参数 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\eta}$ 已确定，则根据词频 $w_{t,n}$ 来推断文档集所对应的话题结构(即推断 Θ_t , β_k 和 $z_{t,n}$)可通过求解

$$p(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Theta} | \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{W} | \boldsymbol{\alpha}, \boldsymbol{\eta})}. \quad (14.44)$$

然而由于分母上的 $p(\mathbf{W} | \boldsymbol{\alpha}, \boldsymbol{\eta})$ 难以获取，式(14.44)难以直接求解，因此在实践中常采用吉布斯采样或变分法进行近似推断。