Author：Liu Jian

Time：2020-06-28

机器学习11b-概率图模型举例

1 隐马尔可夫模型 (hidden Markov model, HMM)

1.1 模型描述

隐马尔可夫模型有两组随机变量：$I=(i_1,\cdots,i_T), O=(o_1,\cdots,o_T)$，它们之间的依赖关系如下：

$$\begin{array}{ccccccc} i_1 & \rightarrow & i_2 & \cdots & i_{T-1} & \rightarrow & i_T \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ o_1 & & o_2 & \cdots & o_{T-1} & & o_T \end{array}$$

其中，马尔可夫链 $I=(i_1,\cdots,i_T)$ 称为状态序列 (state sequence)，是隐变量，因此称为隐马尔可夫链； $O=(o_1,\cdots,o_T)$ 称为观测序列 (observation sequence)，每个观测 o_t 依赖于相应的状态 i_t 。马尔可夫链：系统下一时刻的状态仅由当前状态决定，不依赖于以往的任何状态。

注意，i_t **(** $t=1,\cdots,T$ **)，下标** t **指的是第** t **种随机变量，而不是对某个总体** i **的第** t **次独立重复采样，对** o_t **的理解也类似。我们可以将** $(i_1,\cdots,i_T,o_1,\cdots,o_T)$ **视为一个长度为** $2T$ **的随机向量** $\boldsymbol{x}$ **，** i_t,o_t **是随机向量的分量，但和一般的随机向量不同的是，表征长度的量** T **像时间一样可以无限延长，** t **表示第** t **个时刻。类似朴素贝叶斯的条件独立性假设，如上图所示，隐马尔科夫模型也对随机向量** $\boldsymbol{x}$ **各分量间的依赖关系作出了假设，图中所作的假设可以表述为**：

1. 齐次一阶马尔可夫性假设：隐藏的马尔科夫链在任意时刻 t 的状态只依赖于其前一时刻的状态，与其他时刻的状态及观测无关 (“一阶”的含义)；也与时刻 t 无关 (“齐次”的含义)；
2. 观测独立性假设：任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其它观测及状态无关。

记状态的取值空间 $\mathbb{Q}=\{q_1,\cdots,q_N\}$，观测的取值空间 $\mathbb{V}=\{v_1,\cdots,v_M\}$。隐马尔科夫模型的参数为三元组 $\lambda=(A,B,\boldsymbol{\pi})$，其中：

1. $A_{N\times N}=[a_{n_1n_2}]_{N\times N}$ 为状态转移概率矩阵， $a_{n_1n_2}$ 表示任意时刻状态值为 q_{n_1} 而下一时刻的状态值为 q_{n_2} 的概率，这一概率与时刻无关：

$$P(i_{t+1}=q_{n_2}|i_t=q_{n_1})=a_{n_1n_2},\ \ t=1,\cdots,T-1$$

可见，A **给出了状态序列之间的关系。**

2. $B_{N\times M}=[b_{nm}]_{N\times M}$ 为观测概率矩阵， b_{nm} 表示任一时刻状态值为 q_n 的条件下观测值为 v_m 的概率，同样地，这一概率与时刻无关：

$$P(o_t=v_m|i_t=q_n)=b_{nm},\ \ t=1,\cdots,T$$

可见，B **给出了状态序列与观测序列之间的关系。**事实上，**HMM 只要求状态的值为离散值，观测的值可以为连续的，即对于每种状态取值，都对应有一个连续的观测分布，比如高斯分布，此时就类似于时间序列 + 高斯混合模型。因此，HHM 本质上等于时间序列 + 混合模型，是一种动态模型 (dynamic model/state space model)。**

3. $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)^T$ 为初始状态概率向量， π_n 给出了初始时刻 $t=1$ 状态 o_1 取值为 q_n 的概率：

$$P(o_1=q_n)=\pi_n$$

可见，$\boldsymbol{\pi}$ **给出了初始的状态取值情况，就像给多米诺骨牌一个初始力一样。**

给定参数 λ **，我们每做一次实验就得到一条链，我们可以进行多次独立实验，得到多条链，比如进行** S **次独立试验，得到** $\{(O_1,I_1),\cdots,(O_S,I_S)\}$ **，且各链的长度即** T **可以不相等。**

概率模型的因子分解 (概率有向图的因子分解基于条件概率)：

$$P(O,I)=P(i_1)P(o_1|i_1)\sum_{t=2}^{T}P(i_t|i_{t-1})P(o_t|i_t)$$

可以看到，HMM 是一个生成式模型。

我们要解决如下的三个问题：

1. 学习问题：已知观测序列 O 的值，估计参数 λ：

$$\max_{\lambda}P(O|\lambda)$$

2. 推断问题
    1. 概率计算问题：已知参数 λ 和观察序列 O 的值，计算观测序列出现的概率 $P(O|\lambda)$；
    2. 解码 (decoding) 问题：已知参数 λ 和观测序列的值 O ，求最有可能的状态序列：

$$\max_{I}P(I|\lambda,O)$$

可以看到，所谓的标注问题就是解码问题。

下面我们讨论求解这三个问题的方法，可以看到，核心思想为：动态规划 (dynamic programming，递归计算反过来 (也就是递推计算) +保留并利用中间结果) 和 EM 算法。

> 递推：从初值出发反复进行某一运算得到所需结果。-----**从已知到未知**，从小到大。

> 递归：从所需结果出发不断回溯前一运算直到回到初值再递推得到所需结果----**从未知到已知**，从大到小，再从小到大。递归 (Recursion) 是从归纳法 (Induction) 衍生出来的。

1.2 概率计算

概率问题：已知 λ 和 O ，计算概率 $P(O|\lambda)$ 。

我们可以看到，直接计算法理论上可行，但实际计算量太大而不可行，由此，我们基于动态规划算法开发出了前向 (forward) 算法与后向 (backward) 算法。

1.2.1直接计算法

我们有：

$$P(O,I|\lambda)=\pi_{i_1}b_{i_1o_1}a_{i_1i_2}\cdots a_{i_{T-1}i_2}b_{i_To_T}$$
$$P(O|\lambda)=\sum_{i_1\in\mathbb{Q},\cdots,i_T\in\mathbb{Q}}P(O,I|\lambda)$$

计算复杂度 $O(TN^T)$ ，过大，实际计算过程中不可行。

1.2.2 前向算法

前向概率 $\alpha_t(n)$ 表示从 1 到 t 的部分观测序列为 $o_1,\cdots,o_t$ 且 t 时刻状态为 q_n 的概率：

$$\alpha_t(n)=P(o_1,\cdots,o_t,i_t=q_n|\lambda)$$

相应的动态规划算法，也就是前向算法：

(1) 计算初始值：

$$\alpha_1(n)=\pi_n b_{no_1},\ \ n=1,\cdots,N$$

(2) 对 $t=1,\cdots,T-1$ 进行递推计算：

$$\alpha_{t+1}(n)=\left(\sum_{j=1}^{N}\alpha_t(j)a_{jn}\right)b_{no_{t+1}},\ \ n=1,\cdots,N$$

(3) 计算概率：

$$P(O|\lambda)=\sum_{n=1}^{N}\alpha_T(n)$$

上述计算复杂度为 $O(TN^2)$ 。

若将 α 看作是一个 $T\times N$ 的矩阵， $\alpha_t(n)$ 为其第 t 行 n 列的元素，则上述计算过程相当于从上到下依次计算每一层的元素，计算当前元素会用到前面的计算结果，计算完矩阵 α 后将最后一行的元素相加即得我们要求的概率。

1.2.3 后向算法

后向概率 $\beta_t(n)$ 表示在 t 时刻状态为 q_n 的条件下，从 $t+1$ 到 T 的部分观测序列为 $o_{t+1},\cdots,o_T$ 概率：

$$\beta_t(n)=P(o_{t+1},\cdots,o_T|i_t=q_n,\lambda)$$

基于后向概率的动态规划算法即为后向算法：

(1) 置初值：

$$\beta_T(n)=1,\ \ n=1,\cdots,N$$

(2) 对 $t=T-1,\cdots,1$ 进行递推计算：

$$\beta_t(n)=\sum_{j=1}^{N}a_{nj}b_{jo_{t+1}}\beta_{t+1}(j),\ \ n=1,\cdots,N$$

(3) 计算概率：

$$P(O|\lambda)=\sum_{n=1}^{N}\pi_n b_{no_1}\beta_1(n)$$

类似地，若将 β 视为 $T\times N$ 的矩阵，则上述计算过程就是从下到上依次计算每一行的元素，最后使用第一行的元素计算概率。

1.2.4 基于前向概率和后向概率的其他公式

1. 使用前向概率和后向概率，观测序列概率 $P(O|\lambda)$ 可以表述为：

$$P(O|\lambda)=\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_t(j)a_{jk}b_{ko_{t+1}}\beta_{t+1}(k)$$

其中，时刻 t 可以从 $1,\cdots,T-1$ 中任意选取，当取 $t=T-1$ 时，可推得前向算法中概率的计算公式，当取 $t=1$ 时，可推得后向算法中概率的计算公式。

2. 给定参数 λ 和观测序列 O ，时刻 t 状态为 q_n 的概率：

$$\begin{aligned}\gamma_t(n)&=P(i_t=q_n|O,\lambda)\\&=\frac{P(i_t=q_n,O|\lambda)}{P(O|\lambda)}=\frac{\alpha_t(n)\beta_t(n)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\end{aligned}$$

给定参数 λ 和观测序列 O ，时刻 t 状态为 q_{n_1} 且时刻 $t+1$ 状态为 q_{n_2} 的概率：

$$\zeta_t(n_1,n_2)=P(i_t=q_{n_1},i_{t+1}=q_{n_2}|O,\lambda)$$
$$=\frac{P(i_t=q_{n_1},i_{t+1}=q_{n_2},O|\lambda)}{P(O|\lambda)}=\frac{\alpha_t(n_1)a_{n_1n_2}b_{n_2o_{t+1}}\beta_{t+1}(n_2)}{\sum_{j=1}^N\sum_{k=1}^N\alpha_t(j)a_{jk}b_{ko_{t+1}}\beta_{t+1}(k)}$$

1.3 学习算法 -- Baum-Welch 算法

进行 S 次独立试验对应 $\{(O_1,I_1),\cdots,(O_S,I_S)\}$，其中每次实验的时长 T 任意，不一定相等。若我们得到完全数据 $\{(O_1,I_1),\cdots,(O_S,I_S)\}$，则由极大似然估计，通过计算频率即可估计相应的概率参数，这种已知状态数据的学习算法是一种监督学习算法。

但一般地，状态序列 O 为隐变量，我们只有观测序列数据 $(O_1,\cdots,O_S)$，此时，学习算法为 Baum-Welch 算法。事实上，**Baum-Welch 算法实就是 EM 算法在隐马尔科夫模型上的应用**，这种状态数据未知的学习算法也称无监督学习算法。需要说明的是，根据《统计学习方法》中的内容，**Baum-Welch 算法的计算只基于一次实验的数据，即 $(O_1,\cdots,O_S)$ 中的一个，而不是所有的 $(O_1,\cdots,O_S)$，不过，个人猜想可以运行 S 次 Baum-Welch 算法，再对所得结果取平均。**

Baum-Welch 算法的迭代计算过程如下：

(1) 设定初始值 $\lambda^{(0)}=(A^{(0)},B^{(0)},\boldsymbol{\pi}^{(0)})$；

(2) 对 $k=0,1,\cdots$，迭代计算：

$$a_{n_1n_2}^{(k+1)}=\frac{\sum_{t=1}^{T-1}\zeta_t^{(k)}(n_1,n_2)}{\sum_{t=1}^{T-1}\gamma_t^{(k)}(n_1)},\ \ n_1,n_2=1,\cdots,N$$
$$b_n^{(k+1)}(m)=\frac{\sum_{t=1,o_t=v_m}^{T}\gamma_t^{(k)}(n)}{\sum_{t=1}^{T}\gamma_t^{(k)}(n)},\ \ n=1,\cdots,N;\ m=1,\cdots,M$$
$$\pi_n^{(k+1)}=\gamma_1^{(k)}(n),\ \ n=1,\cdots,N$$

1.4 解码问题

解码问题：已知 λ 和 O，求最有可能的状态序列：

$$I^*=\arg\max_I P(I|\lambda,O)$$

存在两种算法，一种是近似算法，另一种是基于动态规划的精确算法--维特比算法。

1.4.1 近似算法

近似算法的思路是：在每个时刻 t 选择在该时刻最有可能出现的状态 $i_t^*=\arg\min_n\gamma_t(n)$，从而得到一个状态序列 $I^*=(i_1^*,\cdots,i_T^*)$。近似算法没有考虑状态间的关系，因此得到的状态序列在实际中可能根本不会发生，即相邻两个状态的转移概率可能为 0。

1.4.2 维特比算法

维特比算法实际上就是用动态规划求概率最大的状态路径 I^*。维特比算法的思路如下：

1. 以时刻 $1,\cdots,T$ 为横坐标，状态 $q_1,\cdots,q_N$ 为纵坐标，每个横坐标 t 对应有 N 种纵坐标取值 (一个时刻对应 N 个结点，共有 $T\times N$ 个结点)，我们就是要在观测序列为给定的 O 的条件下找从 1 到 T 的概率最大的状态路径 (对应一条从左往右通过 T 个结点的折线)；
2. 由反证法易知，若概率最大的状态路径为 $I^*=(i_1^*,\cdots,i_T^*)$，则从结点 i_1^* 到结点 i_t^* 的最优路径一定通过 $(i_2^*,\cdots,i_{t-1}^*)$。由此，对 $t=1,\cdots,T$，我们可以依次计算时刻 t 下各结点 $n=1,\cdots,N$ (也就是状态) 的最优路径的长度 (也就是概率) $\delta_t(n)$，并记录下当前结点 (坐标为 (t,n)) 对应的最优路径上前一个结点的状态 (记为 $\psi_t(n)$) 以便回溯：

$$\delta_t(n)=\max_{(i_1,\cdots,i_{t-1})\in\mathbb{Q}^{t-1}}P(i_1,\cdots,i_{t-1},i_t=n,o_1,\cdots,o_t|\lambda),\ \ n=1,\cdots,N$$

注意上式中的 $o_1,\cdots,o_t$ 是给定的已知量而不是变量。进一步地，其递推公式如下：

$$\delta_t(n)=\max_{1\leqslant n_1\leqslant N}\delta_{t-1}(n_1)a_{n_1n}b_{no_t}$$
$$\psi_t(n)=\arg\max_{1\leqslant n_1\leqslant N}\delta_{t-1}(n_1)a_{n_1n}b_{no_t}$$
$$=\arg\max_{1\leqslant n_1\leqslant N}\delta_{t-1}(n_1)a_{n_1n}$$

其中，$t=2,\cdots,T$；$n=1,\cdots,N$。

3. 在计算完 T 时刻 N 个结点各自的最优路径长度，也就是概率 $\delta_T(n)$ ($n=1,\cdots,N$) 后，概率最大的结点就是我们要求的最优路径的终点 $n_2=\max_{1\leqslant n\leqslant N}\delta_T(n)$；然后，我们基于保存的通过各结点的最优路径的前一个节点依次回溯，就可

以得到最优的状态序列 I^* 。

2 条件随机场 (conditional random field, CRF)

2.1 模型描述

条件随机场考察的是条件概率 $P(Y|X)$ ，其中随机变量 Y 构成一个马尔可夫随机场，也就是说，条件随机场就是把概率无向图模型/马尔可夫随机场中的概率 $P(Y)$ 变成给定随机变量 X 下的条件概率 $P(Y|X)$ 。**条件随机场可看做无向图 (随机变量 Y 的概率图模型) 和有向图 (由条件概率中的条件 X 产生) 的结合。**

接下来，我们考察线性链条件随机场 (linear chain conditional random field，也称之为链式条件随机场，chain-structured CFR)，它常用于标注问题，其中 X 是可观测的 (观测序列)， Y 是不可观测的 (标记序列或状态序列)，如下图所示：

$$\begin{array}{ccccccccc} Y_1 & - & Y_2 & \cdots & Y_{n-1} & - & Y_n \\ \nwarrow & & \uparrow & & \uparrow & & \nearrow \\ & & & X=(X_1,\cdots,X_n) & & & \end{array}$$

进一步地，若 X,Y 具有相同图结构：

$$\begin{array}{ccccccccc} Y_1 & - & Y_2 & \cdots & Y_{n-1} & - & Y_n \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ X_1 & & X_2 & \cdots & X_{n-1} & & X_n \end{array}$$

接下来，我们给出链式条件随机场的**三种数学表达形式：参数化形式，向量形式，矩阵形式。**

2.1.1 链式条件随机场的参数化形式

由前可知，只看 Y 部分，链式条件随机场的节点和边分别为：

$$\begin{aligned} G&=(V,E)\\ V&=(1,2,\cdots,n)\\ E&=\{(i,i+1)\},\ \ i=1,\cdots,n-1 \end{aligned}$$

即每条边 $(i,i+1)$ 对应了一个最大团 $C_i=\{Y_i,Y_{i+1}\}$ 。基于因子分解，我们取 $P(Y|X)$ 为如下的形式：

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})=\frac{1}{Z(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})}\exp\left(\sum_{i=1}^{n-1}\sum_{k=1}^{K_1}\lambda_k t_k(y_i,y_{i+1},\boldsymbol{x},i)+\sum_{i=1}^{n}\sum_{l=1}^{K_2}\mu_l s_l(y_i,\boldsymbol{x},i)\right)$$

其中， t_k,s_l 为给定的特征函数； λ_k,μ_l 为对应的权值，是模型的参数； $Z(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})$ 为规范化因子。再次强调， y_i 是随机向量 $\boldsymbol{y}=(y_1,\cdots,y_n)$ 的分量，而不是一个样本。从因子分解角度，我们可以认为上述模型所取势函数如下 (当然，上述模型对应势函数的选取并不是唯一的，下面只是一种可能)：

$$\Psi_{C_i}(Y_{C_i})=\begin{cases}\exp\left(\sum_{k=1}^{K_1}\lambda_k t_k(y_1,y_2,\boldsymbol{x},1)+\sum_{l=1}^{K_2}\left(\mu_l s_l(y_1,\boldsymbol{x},1)+\frac{1}{2}\mu_l s_l(y_2,\boldsymbol{x},2)\right)\right), & i=1\\ \exp\left(\sum_{k=1}^{K_1}\lambda_k t_k(y_i,y_{i+1},\boldsymbol{x},i)+\sum_{l=1}^{K_2}\frac{\mu_l}{2}(s_l(y_i,\boldsymbol{x},i)+s_l(y_{i+1},\boldsymbol{x},i+1))\right), & i=2,\cdots,n-2\\ \exp\left(\sum_{k=1}^{K_1}\lambda_k t_k(y_{n-1},y_n,\boldsymbol{x},n-1)+\sum_{l=1}^{K_2}\left(\frac{1}{2}\mu_l s_l(y_{n-1},\boldsymbol{x},n-1)+\mu_l s_l(y_n,\boldsymbol{x},n)\right)\right), & i=n-1\end{cases}$$

可以看到，**因为是条件概率，所以特征函数 $t_k(y_i,y_{i+1},\boldsymbol{x},i),s_l(y_i,\boldsymbol{x},i)$ 均与 $\boldsymbol{x}$ 有关；此外，它们还与位置 i 有关，也就是虽然最大团的结构都相同，但位置不同，对应的势函数也不同。 $t_k(y_i,y_{i+1},\boldsymbol{x},i)$ 定义在边 $(i,i+1)$ 上，称为转移特征函数 (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 (注意，与 HMM 不同，这里是无向图模型，因此并没有转移概率的含义)； $s_l(y_i,\boldsymbol{x},i)$ 定义在结点 i 上，称为状态特征函数 (status feature function)，用于刻画观测序列对标记变量的影响；可见，同一特征在各个位置均有定义， t_k,s_l 为局部特征函数。**

上述链式条件随机场是对数线性模型 (log linear model)。

2.1.2 链式条件随机场的向量形式

参数化形式中的特征函数为局部特征函数，共有 $K=K_1+k_2$ 个，在每个位置均有定义。这里，我们对每个特征函数在各个位置求和，将**局部特征函数转化为全局特征函数。**

记局部特征函数 $g_k(y_i,y_{i+1},\boldsymbol{x},i)$ 和对应的权值向量 $\boldsymbol{w}=\langle w_1,\cdots,w_K\rangle^T$ ：

$$g_k(y_i, y_{i+1}, \boldsymbol{x}, i) = \begin{cases} t_k(y_i, y_{i+1}, \boldsymbol{x}, i), & k = 1, \cdots, K_1 \\ s_l(y_i, \boldsymbol{x}, i), & k = K_1 + l,\ l = 1, \cdots, K_2 \end{cases}$$

$$w_k = \begin{cases} \lambda_k, & k = 1, \cdots, K_1 \\ \mu_l, & k = K_1 + l,\ l = 1, \cdots, K_2 \end{cases}$$

则全局特征函数 $f_k(\boldsymbol{y}, \boldsymbol{x})$：

$$f_k(\boldsymbol{y}, \boldsymbol{x}) = \sum_i g_k(y_i, y_{i+1}, \boldsymbol{x}, i)$$

$$= \begin{cases} \sum_{i=1}^{n-1} t_k(y_i, y_{i+1}, \boldsymbol{x}, i), & k = 1, \cdots, K_1 \\ \sum_{i=1}^{n} s_l(y_i, \boldsymbol{x}, i), & k = K_1 + l,\ l = 1, \cdots, K_2 \end{cases}$$

记全局特征函数向量 $\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x})$：

$$\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}) = \langle f_1(\boldsymbol{y}, \boldsymbol{x}), \cdots, f_K(\boldsymbol{y}, \boldsymbol{x}) \rangle^T$$

则概率分布：

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp\left(\sum_{i=1}^{n-1} \sum_{k=1}^{K_1} \lambda_k t_k(y_i, y_{i+1}, \boldsymbol{x}, i) + \sum_{i=1}^{n} \sum_{l=1}^{K_2} \mu_l s_l(y_i, \boldsymbol{x}, i) \right)$$

$$= \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp\left(\sum_{k=1}^{K_1} \lambda_k \sum_{i=1}^{n-1} t_k(y_i, y_{i+1}, \boldsymbol{x}, i) + \sum_{l=1}^{K_2} \mu_l \sum_{i=1}^{n} s_l(y_i, \boldsymbol{x}, i) \right)$$

$$= \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp\left(\sum_{k=1}^{K} w_k f_k(\boldsymbol{y}, \boldsymbol{x}) \right) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp\left(\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}) \right)$$

其中，规范化因子：

$$Z(\boldsymbol{x}, \boldsymbol{w}) = \int_{\mathbb{Y}} \exp\left(\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}) \right) \mathrm{d}\boldsymbol{y}$$

2.1.3 链式条件随机场基于因子分解的矩阵形式

> 《统计学习方法》中这一部分的内容表述模糊，细究之下会发现存在矛盾，这是表述不准确导致的。书中既引入了起点状态标记 $y_0 = \text{start}$ 又引入了终点状态标记 $y_{n+1} = \text{stop}$，事实上起点状态标记和终点状态标记只用引入一个即可，这里我们选择引入的是终点状态标记。无论是同时引入起点状态标记和终点状态标记，还是引入二者中的一种，甚至是均不引入 (比如上文给出的势函数取法)，都是行的通的，只要对势函数的定义自洽没有矛盾最后组合起来等于前面设定的模型即可。但《统计学习方法》中的定义是有矛盾的，组合起来得到的并不是前面假定的模型。

对每个位置 $i = 1, \cdots, n-1$，记

$$W_i(y_i, y_{i+1}, \boldsymbol{x}, \boldsymbol{w}) = \sum_{k=1}^{K} w_k g_k(y_i, y_{i+1}, \boldsymbol{x}, i)$$

事实上，给定 $\boldsymbol{x}$，也就是视 $\boldsymbol{x}$ 为常量，则 $\exp\left(W_i(y_i, y_{i+1}, \boldsymbol{x}, \boldsymbol{w})\right)$ 可以视为最大团 $Y_{C_i} = \{i, i+1\}$ 势函数的一种取法 (与前文的取法不同)。但是按照这种势函数的设定，若不引入终点状态标记 $y_{n+1} = \text{stop} = q_1$，则 i 的取值范围为 $1, \cdots, n-1$，此时，势函数无法囊括原模型的所有函数，我们有：

$$\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}) = \left(\sum_{i=1}^{n-1} W_i(y_i, y_{i+1}, \boldsymbol{x}, \boldsymbol{w}) \right) + \left(\sum_{l=1}^{K_2} \mu_l s_l(y_n, \boldsymbol{x}, n) \right)$$

只能表达右式左端的部分。为此，我们引入终点状态标记 $y_{n+1} = \text{stop}$ 以构建最大团 $\{n, n+1\}$，对应有：

$$W_n(y_n, y_{n+1}, \boldsymbol{x}, \boldsymbol{w}) = \sum_{k=1}^{K} w_k g_k(y_n, y_{n+1}, \boldsymbol{x}, n) = \sum_{l=1}^{K_2} \mu_l s_l(y_n, \boldsymbol{x}, n)$$

其中，$g_k(y_n, y_{n+1}, \boldsymbol{x}, n)$ 当 $k = 1, \cdots, K_1$ 时没有定义。

至此，势函数取为：

$$\Psi_i(Y_{C_i}|\boldsymbol{x}, \boldsymbol{w}) = \exp\left(W_i(y_i, y_{i+1}, \boldsymbol{x}, \boldsymbol{w})\right),\ \ i = 1, \cdots, n$$

概率模型的因子分解 (无向图模型的因子分解基于势函数)：

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{\Psi_1(y_1, y_2|\boldsymbol{x}, \boldsymbol{w}) \Psi_2(y_2, y_3|\boldsymbol{x}, \boldsymbol{w}) \cdots \Psi_n(y_n, y_{n+1}|\boldsymbol{x}, \boldsymbol{w})}{Z(\boldsymbol{x}, \boldsymbol{w})}$$

可以看到，CRF 为判别式模型。

接下来，我们就可以基于上述因子分解得到矩阵表达形式了。记随机变量分量 y_i 的样本空间 $\mathbb{Y}=\{q_1,\cdots,q_m\}$，即 Y 的可取值个数为 m。给定 $\boldsymbol{x},\boldsymbol{w}$，我们可构建 n 个 $m\times m$ 阶的矩阵 $H_i\ (i=1,\cdots,n)$：

$$H_i=[h_{jk}]_{m\times m}$$
$$\text{where}$$
$$h_{jk}=\Psi_i(y_i=q_j,y_{i+1}=q_k|\boldsymbol{x},\boldsymbol{w})$$

第 i 个矩阵 H_i 对应第 i 个势函数 $\Psi_i(y_i,y_{i+1}|\boldsymbol{x},\boldsymbol{w})$。需要说明的是，对于 H_n，因为 y_{n+1} 只能取标记固定标记 stop，则 H_n 实际上是一个 $n\times 1$ 的列向量，不过，我们可以将其余位置补 0，使得 H_n 为一个 $n\times n$ 的矩阵。

可以看到，对序列的某种取值 $\boldsymbol{y}=(q_{i_1},\cdots,q_{i_n})$，我们分别选择矩阵 $H_1,\cdots,H_n$ 的第 $(i_1,i_2),(i_2,i_3),\cdots,(i_{n-1},i_n),(i_n,1)$ 个元素相乘，即可得到分子 $\Psi_1(y_1,y_2|\boldsymbol{x},\boldsymbol{w})\Psi_2(y_2,y_3|\boldsymbol{x},\boldsymbol{w})\cdots\Psi_n(y_n,y_{n+1}|\boldsymbol{x},\boldsymbol{w})$ 的值，而分母即规范化因子的计算公式过程如下：(1) 矩阵连乘 $H=\prod_{i=1}^n H_i$；(2) 矩阵 H 第一列元素的累加即为归一化分母的值：

$$Z(\boldsymbol{x},\boldsymbol{w})=\sum_{i=1}^n[H]_{i1}$$

由此，给定 $\boldsymbol{x},\boldsymbol{w}$，$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$ 就可以用矩阵的形式表达出来。

我们指出，我们也可以引入起点状态标记 $y_0=\text{start}=q_1$ 以构建最大团 $\{0,1\}$，而其对应的最大势函数 $\Psi_0(y_0,y_1|\boldsymbol{x},\boldsymbol{w})$ 可取上面的 $\Psi_1(y_1,y_2|\boldsymbol{x},\boldsymbol{w})$ 中的只含 y_1 的部分，剩下的部分为新的势函数 $\Psi_1(y_1,y_2|\boldsymbol{x},\boldsymbol{w})$，即此时：

$$\Psi_0(y_0,y_1|\boldsymbol{x},\boldsymbol{w})=\exp\left(\sum_{l=1}^{K_2}\mu_l s_l(y_1,\boldsymbol{x},0)\right)$$
$$\Psi_1(y_1,y_2|\boldsymbol{x},\boldsymbol{w})=\exp\left(\sum_{k=1}^{K_1}\lambda_k t_k(y_1,y_2,\boldsymbol{x},1)\right)$$

此时，概率模型：

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})=\frac{\Psi_0(y_0,y_1|\boldsymbol{x},\boldsymbol{w})\Psi_1(y_1,y_2|\boldsymbol{x},\boldsymbol{w})\cdots\Psi_n(y_n,y_{n+1}|\boldsymbol{x},\boldsymbol{w})}{Z(\boldsymbol{x},\boldsymbol{w})}$$

势函数对应的矩阵为 $H_0,H_1,\cdots,H_n$，和前面类似，因为 y_0 只能取固定标记 start，因此 H_0 实际上是 $1\times m$ 的行向量，我们通过在其余位置补 0 扩充 H_0 为 $m\times m$ 阶矩阵。此时概率模型中分子的计算与上面类似，分母 $Z(\boldsymbol{x},\boldsymbol{w})$ 的计算过程：(1) 矩阵连乘 $\prod_{i=0}^n H_i$；(2) 第 1 行第 1 列的元素即为归一化分母的值：$Z(\boldsymbol{x},\boldsymbol{w})=[H]_{11}$。

和隐马尔科夫模型类似，我们要解决的问题有：(1) 学习问题；(2) 推断问题，包括概率计算和解码问题。

2.2 概率计算

问题为给定 $\boldsymbol{x},\boldsymbol{w}$，计算：(1) $P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$；(2) $P(y_i|\boldsymbol{x},\boldsymbol{w})$；(3) $P(y_i,y_{i+1}|\boldsymbol{x},\boldsymbol{w})$。

链式条件随机场的概率计算也是基于其因子分解形式，但和隐马尔可夫模型不同的是，这里不存在隐变量，因此若给定 $\boldsymbol{x},\boldsymbol{w}$ 要计算某个序列 $\boldsymbol{y}=(q_{i_1},\cdots,q_{i_n})$ 的概率，我们无需积分，难度大大降低，直接代入因子分解的公式计算即可，**实际上，也就得到了上面提到的矩阵表达形式及相应的概率计算方法。**

若要计算概率 $P(y_i=q_j|\boldsymbol{x},\boldsymbol{w})$，我们有：

$$P(y_i=q_j|\boldsymbol{x},\boldsymbol{w})=\int_{\mathbb{Y}^{n-1}}P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})\mathrm{d}\boldsymbol{y}_{\sim i}$$
$$=\frac{1}{Z(\boldsymbol{x},\boldsymbol{w})}\int_{\mathbb{Y}^{n-1}}\Psi_0(y_0,y_1|\boldsymbol{x},\boldsymbol{w})\cdots$$
$$\Psi_{i-1}(y_{i-1},y_i=q_j|\boldsymbol{x},\boldsymbol{w})\Psi_i(y_i=q_j,y_{i+1}|\boldsymbol{x},\boldsymbol{w})\cdots$$
$$\Psi_n(y_n,y_{n+1}|\boldsymbol{x},\boldsymbol{w})\mathrm{d}y_1\cdots\mathrm{d}y_{i-1}\mathrm{d}y_{i+1}\cdots\mathrm{d}y_n$$

依次计算积分即可。**同样地，这一计算过程也可以表述为矩阵相乘的形式，从左往右积分即为前向算法，从右往左积分即为后向算法。**记 $m\times 1$ 阶前向向量和后向向量分别为 $\boldsymbol{\alpha}_i(\boldsymbol{x},\boldsymbol{w})$ 和 $\boldsymbol{\beta}_i(\boldsymbol{x},\boldsymbol{w})$，为了简便起见，我们忽略记号中的 $\boldsymbol{x},\boldsymbol{w}$，因为概率计算中它们均为给定的量。

前向向量 $\boldsymbol{\alpha}_i\ (i=0,1,\cdots,n)$：

$$\boldsymbol{\alpha}_0 = \begin{pmatrix} \Psi_0(y_0, y_1 = q_1) \\ \Psi_0(y_0, y_1 = q_2) \\ \cdots \\ \Psi_0(y_0, y_1 = q_m) \end{pmatrix}$$

给出了 $\int \Psi_0(y_0, y_1|\boldsymbol{x}, \boldsymbol{w})\mathrm{d}y_0$ 的各种结果，即将 y_0 积分掉 (起点标记固定为 $\mathrm{start} = q_1$ ，故直接代入即可，无需进行积分)，并列出 y_1 为各种可能取值的结果。而

$$\boldsymbol{\alpha}_i^T = \boldsymbol{\alpha}_{i-1}^T H_i \;\; i = 1, \cdots, n$$

$\boldsymbol{\alpha}_i$ 则给出了：

$$\int \Psi_0(y_0, y_1|\boldsymbol{x}, \boldsymbol{w}) \cdots \Psi_i(y_i, y_{i+1}|\boldsymbol{x}, \boldsymbol{w})\mathrm{d}y_1 \cdots \mathrm{d}y_i$$

当 y_{i+1} 取各种可能值时的结果。

类似地，后向向量 $\boldsymbol{\beta}_i$ $(i = 0, \cdots, n)$ ：

$$\boldsymbol{\beta}_n = \begin{pmatrix} \Psi_n(y_n = q_1, y_{n+1}) \\ \Psi_n(y_n = q_2, y_{n+1}) \\ \cdots \\ \Psi_n(y_n = q_m, y_{n+1}) \end{pmatrix}$$

给出了 $\int \Psi_n(y_n, y_{n+1}|\boldsymbol{x}, \boldsymbol{w})\mathrm{d}y_{n+1}$ 的各种可能结果。而

$$\boldsymbol{\beta}_i = H_i \boldsymbol{\beta}_{i+1}, \;\; i = n-1, \cdots, 0$$

$\boldsymbol{\beta}_i$ 则给出了：

$$\int \Psi_i(y_i, y_{i+1}|\boldsymbol{x}, \boldsymbol{w}) \cdots \Psi_n(y_n, y_{n+1}|\boldsymbol{x}, \boldsymbol{w})\mathrm{d}y_{i+1} \cdots \mathrm{d}y_n$$

当 y_i 取各种可能值时的结果。

总地来说，我们有：

$$[\boldsymbol{\alpha}_i]_{j1} = \int \Psi_0(y_0, y_1|\boldsymbol{x}, \boldsymbol{w}) \cdots \Psi_i(y_i, y_{i+1} = q_j|\boldsymbol{x}, \boldsymbol{w})\mathrm{d}y_1 \cdots \mathrm{d}y_i$$

$$[\boldsymbol{\beta}_i]_{j1} = \int \Psi_i(y_i = q_j, y_{i+1}|\boldsymbol{x}, \boldsymbol{w}) \cdots \Psi_n(y_n, y_{n+1}|\boldsymbol{x}, \boldsymbol{w})\mathrm{d}y_{i+1} \cdots \mathrm{d}y_n$$

隐马尔可夫模型和链式条件随机场中的前向概率和后向概率均基于因子分解，但 HMM 中的因子为条件概率，因此其前向和后向概率有明确的意义，而 CRF 的因子为最大团，因此不具备类似的含义。由上述计算公式可以看到，前向概率 $[\boldsymbol{\alpha}_i]_{j1}$ 并不表示"只考虑随机变量 $y_0, y_1 \cdots, y_{i+1}$ 而不考虑随机变量 $y_{i+2}, \cdots, y_{n+1}$ 时， $y_0, y_1, \cdots, y_i$ 为任何可能取值，且 $y_{i+1} = q_j$ "的非规范化概率；同理后向概率 $[\boldsymbol{\beta}_i]_{j1}$ 并不表示"只考虑随机变量 $y_i, \cdots, y_{n+1}$ 而不考虑随机变量 $y_0, \cdots, y_{i-1}$ 时， $y_{i+1}, \cdots, y_{n+1}$ 为任何可能取值，且 $y_i = q_j$ "的非规范化概率，注意与 HMM 中的情况进行区分。

我们有：

1. 规范化因子：

$$\boldsymbol{\alpha}_n = \begin{pmatrix} Z(\boldsymbol{x}, \boldsymbol{w}) \\ 0 \\ \cdots \\ 0 \end{pmatrix} = \boldsymbol{\beta}_0$$

2. 概率 $P(y_i = q_j|\boldsymbol{x}, \boldsymbol{w})$ ：

$$P(y_i = q_j|\boldsymbol{x}, \boldsymbol{w}) = [\boldsymbol{\alpha}_{i-1}]_{j1} \times [\boldsymbol{\beta}_i]_{j1}$$

3. 同理，概率 $P(y_{i-1} = q_j y_i = q_k|\boldsymbol{x}, \boldsymbol{w})$ ：

$$P(y_{i-1} = q_j y_i = q_k|\boldsymbol{x}, \boldsymbol{w}) = [\boldsymbol{\alpha}_{i-2}]_{j1} \times [H_{i-1}]_{jk} \times [\boldsymbol{\beta}_i]_{k1}$$

2.3 学习算法

学习策略为极大似然估计或正则化的极大似然估计，优化算法有改进的迭代尺度法 IIS、梯度下降法及拟牛顿法，具体细节不再赘述。

2.4 解码问题

即给定 $\boldsymbol{x}, \boldsymbol{w}$，求使 $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ 最大的 $\boldsymbol{y}$。和 HMM 一样，采用维特比算法，即动态规划算法，并保存当前结点最优路径的前一个结点以便回溯，具体细节不再赘述。

3 隐狄利克雷分配模型 (Latent Dirichlet Allocation, LDA)

3.1 模型描述

和隐马尔可夫模型一样，话题模型也是一种生成式 (对联合概率进行建模) 有向图模型，典型代表就是隐狄利克雷分配模型。

说明：

1. 词 (word)：待处理数据的基本离散单元，词典含有 N 个词；
2. 文档 (document)：待处理的数据对象，一篇文档对应一个 N 维向量；设共有 T 篇文档 $W = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T\}$，其中，$w_{tn}$ 表示文档 t 中词 n ($1 \leqslant n \leqslant N$) 的**出现的个数 (也就是词频，即词出现的频次，而不是频率)**，第 t 篇文档共有 V_t 个词。
3. 话题 (topic)：一个话题也对应一个 N 维向量，设共有 K 个话题 $\mathrm{B} = \{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K\}$，其中，$\beta_{kn}$ 表示第 k 个话题下词 n ($1 \leqslant n \leqslant N$) 出现的概率，每个向量 $\boldsymbol{\beta}_k$ ($1 \leqslant k \leqslant K$) 都表示一个概率分布。
4. $\Theta = \{\Theta_1, \cdots, \Theta_T\}$，$K$ 维向量 Θ_t ($1 \leqslant t \leqslant T$) 表示第 t 篇文档中每个话题所占的比例，即概率分布，Θ_{tk} 表示文档 t 中话题 k ($1 \leqslant k \leqslant K$) 所占的比例，即话题 k 出现的概率。
5. $Z = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_T\}$，$N$ 维向量 $\boldsymbol{z}_t$ 表示第 t 篇文档中各词所属的话题，其中 z_{tn} 表示文档 t 中词 n 所属的话题。

LDA 模型中文档 $\boldsymbol{w}_t$ 的生成过程：

$$\left.\begin{array}{r} \text{狄利克雷分布}(\text{参数为}K\text{维向量}\boldsymbol{\alpha}) \xrightarrow{\text{生成}} \text{文档 } t \text{ 的话题分布 } \Theta_t \xrightarrow{\text{生成，也称话题指派}} \boldsymbol{z}_t \\ \text{狄利克雷分布}(\text{参数为}N\text{维向量}\boldsymbol{\eta}) \xrightarrow{\text{生成}} \mathrm{B} \end{array}\right\} \xrightarrow{\text{得到}} \boldsymbol{w}_t$$

上述中的\\生成//也就是采样，而\\得到//对应的具体过程为：

$$\text{分布 } \boldsymbol{\beta}_{z_{tn}} \xrightarrow{V_t\text{次实验，并计词}n\text{出现的次数}} w_{tn}$$

即 **LDA 模型按照如下的步骤生成文档 t 中词 n 的词频 w_{tn}：z_{tn} 给出了文档 t 中词 n 所属的话题，而 $\boldsymbol{\beta}_{z_{tn}}$ 则给出了该话题下每个词出现的概率分布，根据这个概率分布，进行 V_t 次独立重复实验，统计词 n 出现的次数，即得词频 w_{tn} 。**

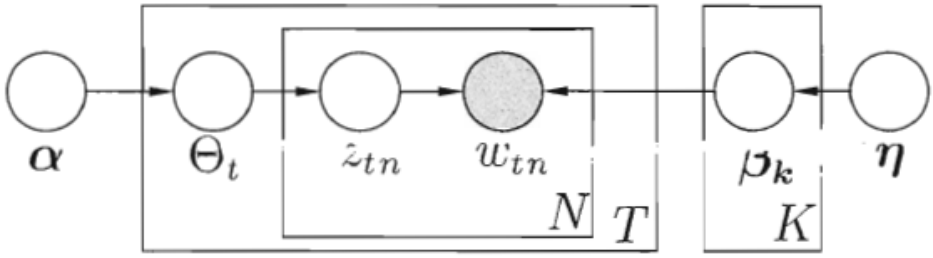


图 14.12 LDA 的盘式记法图

由上述对 LDA 的描述可知，**LDA 模型的概率分布如下**：

$$p(W, Z, \mathrm{B}, \Theta|\boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{t=1}^{T} p(\boldsymbol{w}_t, \boldsymbol{z}_t, \mathrm{B}, \Theta_t|\boldsymbol{\alpha}, \boldsymbol{\eta})$$

$$p(\boldsymbol{w}_t, \boldsymbol{z}_t, \mathrm{B}, \Theta_t|\boldsymbol{\alpha}, \boldsymbol{\eta}) = p(\Theta_t|\boldsymbol{\alpha}) p(\boldsymbol{z}_t|\Theta_t) p(\mathrm{B}|\boldsymbol{\eta}) p(\boldsymbol{w}_t|\boldsymbol{z}_t, \mathrm{B})$$

记$\mathcal{D}$表示狄利克雷分布，$\mathcal{C}$表示类别分布$(\text{Categorical distribution})$，则：

$$p(\Theta_t|\boldsymbol{\alpha}) = p_{\mathcal{D}}(\Theta_t|\boldsymbol{\alpha})$$

$$p(\boldsymbol{z}_t|\Theta_t) = \prod_{n=1}^{N} p(z_{tn}|\Theta_t) = \prod_{n=1}^{N} p_{\mathcal{C}}(z_{tn}|\Theta_t)$$

$$p(\mathrm{B}|\boldsymbol{\eta}) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) = \prod_{k=1}^{K} p_{\mathcal{D}}(\boldsymbol{\beta}_k|\boldsymbol{\eta})$$

$$p(\boldsymbol{w}_t|\boldsymbol{z}_t, \mathrm{B}) = \prod_{n=1}^{N} p(w_{tn}|\boldsymbol{z}_t, \mathrm{B}) = \prod_{n=1}^{N} p(w_{tn}|z_{tn}, \mathrm{B}) = \prod_{n=1}^{N} p(w_{tn}|\boldsymbol{\beta}_{z_{tn}}) = \prod_{n=1}^{N} p_{\mathcal{M}}(w_{tn}|\boldsymbol{\beta}_{z_{tn}}, V_t)$$

类别分布 $p_{\mathcal{C}}(z_{tn}|\Theta_t)$ 表示文档 t 的第 n 个词属于话题 z_{tn} 的概率，即 $p_{\mathcal{C}}(z_{tn}=k|\Theta_t)=\Theta_{tk}$； $p_{\mathcal{D}}(w_{tn}|\beta_{z_{tn}},V_t)$ 表示独立重复的 V_t 次概率分布为 $\beta_{z_{tn}}$ 的类别实验中，类别 n（也就是词 n）出现次数为 w_{tn} 的概率。

可以看到，**文档 t 的词频向量 $\boldsymbol{w}_t$ 是观测量，而 $\boldsymbol{z}_t,\mathrm{B},\Theta_t$ 为隐变量， $\boldsymbol{\alpha},\boldsymbol{\eta}$ 为模型参数。**

3.2 模型学习

给定训练数据 $W=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_T\}$，学习模型的参数 $\boldsymbol{\alpha},\boldsymbol{\eta}$。最大化对数似然函数：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\eta}}\sum_{t=1}^{T}\ln p(\boldsymbol{w}_t|\boldsymbol{\alpha},\boldsymbol{\eta})$$

即含有隐变量的极大似然估计，可用 EM 算法或变分法 (变分推断的学习和推断结合在一起，同时得解) 求解。

3.3 模型推断

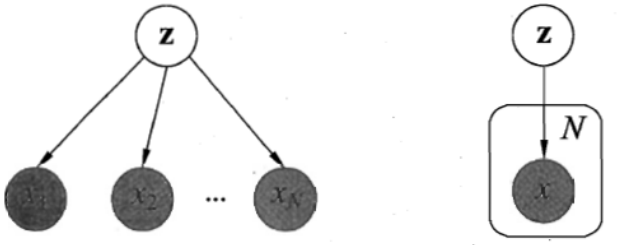
已知参数 $\boldsymbol{\alpha},\boldsymbol{\eta}$，根据词频 $W=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_T\}$ 来推断文档集所对应的话题结构 (推断 Z,B,Θ)，即求解：

$$p(Z,\mathrm{B},\Theta|W,\boldsymbol{\alpha},\boldsymbol{\eta})=\frac{p(W,Z,\mathrm{B},\Theta|\boldsymbol{\alpha},\boldsymbol{\eta})}{p(W|\boldsymbol{\alpha},\boldsymbol{\eta})}$$

因分母的 $p(W|\boldsymbol{\alpha},\boldsymbol{\eta})$ 难以获取，所以上式难以直接求解。实际中，常采用吉布斯采样或变分法进行近似推断。

4 附录-庞氏记法

在学习变分推断之前, 我们先介绍概率图模型一种简洁的表示方法——盘式记法(plate notation) [Buntine, 1994]. 图 14.10 给出了一个简单的例子. 图 14.10(a)表示 N 个变量 $\{x_1,x_2,\ldots,x_N\}$ 均依赖于其他变量 $\mathbf{z}$. 在图 14.10(b)中, 相互独立的、由相同机制生成的多个变量被放在一个方框(盘)内, 并在方框中标出类似变量重复出现的个数 N; 方框可以嵌套. 通常用阴影标注出已知的、能观察到的变量, 如图 14.10 中的变量 x. 在很多学习任务中, 对属性变量使用盘式记法将使得图表示非常简洁.



(a) 普通变量关系图　　(b) 盘式记法

图 14.10　盘式记法的例示