第6章 隐马尔可夫模型与条件随机场

6.1 马尔可夫模型

如果一个系统有 N 个状态 $S_1, S_2, \cdots, S_N$，随着时间的推移，系统从某一状态转移到另一状态，如果用 q_t 表示系统在时间 t 的状态变量，即 q_t 时刻的状态取值为 $S_j (1 \le j \le N)$ 的概率取决于前 $t-1$ 个时刻 $(1, 2, \cdots, t-1)$ 的状态.

概率为 $P(q_t = S_j \mid q_{t-1} = S_k, q_{t-2} = S_l, \cdots)$ (6.1)

假设1：在特定情况下，系统在时间 t 的状态只与前一个时刻的 $t-1$ 状态相关（离散的一阶马尔可夫链）：

$$P(q_t = S_j \mid q_{t-1} = S_k, q_{t-2} = S_l, \cdots) = P(q_t = S_j \mid q_{t-1} = S_k) \quad (6.2)$$

假设2：如果只考虑 (6.1) 独立于时间 t 的随机过程，即不动性假设，状态与时间无关，即 $P(q_t = S_j \mid q_{t-1} = S_i) = a_{ij}$，$1 \le i, j \le N$ (6.3)

(a_{ij} 与时间无关)

a_{ij} 为状态转移概率 $\sum_{j=1}^{N} a_{ij} = 1$

状态图表示：（状态转移上有概率的非确定的有限状态自动机）

{ 0 概率省略
每个节点所有出边的概率之和等于 1

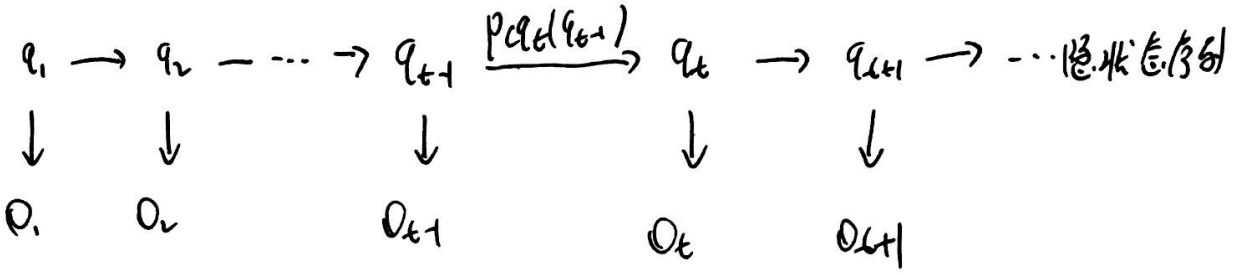
状态序列 $S_1, \cdots, S_T$ 的概率

$$\begin{aligned} P(S_1, \cdots, S_T) &= P(S_1) \times P(S_2 \mid S_1) \times P(S_3 \mid S_1, S_2) \times \cdots \times P(S_T \mid S_1, \cdots, S_{T-1}) \\ &= P(S_1) \times P(S_2 \mid S_1) \times P(S_3 \mid S_2) \times \cdots \times P(S_T \mid S_{T-1}) \\ &= \pi_{S_1} \prod_{t=1}^{T-1} a_{S_t S_{t+1}} \end{aligned}$$

其中，$\pi_i = P(q_1 = S_i)$，为初始状态的概率.

6.2 隐马尔可夫模型.

22重随机过程. 不知道具体的状态序列, 只知道状态转移的概率, 即模型的状态转移过程是不可观察的(隐藏的), 而可观察事件的随机过程是隐藏状态转移过程的随机函数.



HMM图解

组成
- 模型中的状态数为 N (盒子的数量)
- 从每一个状态可能输出的不同的符号数 M (不同颜色球的数目)
- 状态转移概率矩阵 $A=a_{ij}$
  $a_{ij}=P(q_{t+1}=S_j \mid q_t=S_i),\ 1\le i,j\le N$
  $a_{ij}\ge 0$
  $\sum_{j=1}^{N} a_{ij}=1$
- 处状态 S_j 观察到某一特定符号 V_k 的概率分布矩阵 $B=b_j(k)$
  $b_j(k)=P(O_t=V_k \mid q_t=S_j),\ 1\le j\le N,\ 1\le k\le M$
  $b_j(k)\ge 0$
  $\sum_{k=1}^{M} b_j(k)=1$
- 初始状态的概率分布 $\pi=\pi_i$
  $\pi_i=P(q_1=S_i),\ 1\le i\le N$
  $\pi_i\ge 0$
  $\sum_{i=1}^{N}\pi_i=1$

指HMM记为: $\lambda=(A,B,\pi)$ 或 $\lambda=(S,O,A,B,\pi)$.

给定HMM求观察序列:

给定 $\lambda=(A,B,\pi)$, 求 $O=O_1O_2\cdots O_T$

(1). $t=1$

(2). 根据初始状态分布 $\pi=\pi_i$ 选择初始状态 $q_1=s_i$

(3). 根据状态 s_i 的输出概率分布 $b_i(k)$，输出 $O_t=V_k$

(4). 根据状态转移概率 a_{ij}，转移到新状态 $q_{t+1}=s_j$

(5). $t=t+1$，如果 $t<T$，重复(3).(4). 否则结束

问题：(1). 快速计算概率 $P(O|\mu)$（给定 μ 和 O）

(2). 如何选择一定意义下"最优"的状态序列，使得此状态序列"最好地解释"

(3). 给定观察序列 O，如何根据最大似然估计来求模型的参数值？即如何求得使 $P(O|\mu)$ 最大的参数？

6.3 前向算法

解决问题(1). 对于给定的状态序列 $Q=q_1q_2\cdots q_T$，$P(O|\mu)=?$

$$P(O|\mu)=\sum_Q P(O,Q|\mu)=\sum_Q P(Q|\mu)\times P(O|Q,\mu)$$

$$\frac{P(O,Q|\mu)}{P(\mu)}=\frac{P(Q|\mu)P(O,Q|\mu)}{P(\mu)}$$ (乘法公式)

$$P(Q|\mu)=\pi_{q_1}\times a_{q_1q_2}\cdots a_{q_{T-1}q_T}$$

$$P(O|Q,\mu)=b_{q_1}(O_1)\times b_{q_2}(O_2)\cdots b_{q_T}(O_T)$$

困难：搜索路径成指数级爆炸

解决方法：动态规划

定义前向变量 $\alpha_t(i)$：

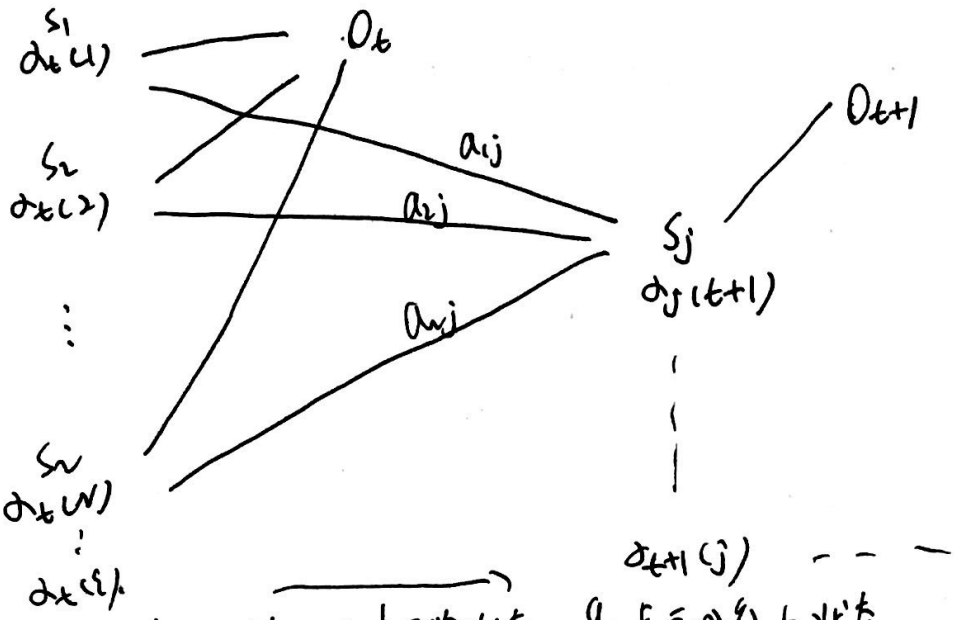
$$\alpha_t(i)=P(O_1O_2\cdots O_t,\ q_t=s_i|\mu)$$

（当前时刻序列概率，由上一步从 O_1 算到 O_t）

$$P(O|\mu)=\sum_i P(O_1O_2\cdots O_T,\ q_T=s_i|\mu)=\sum_i^N \alpha_T(i)$$

动态规划计算 $\alpha_t(i)$：在时间 $t+1$ 的前向变量可以根据时间 t 的前向变量 $\alpha_t(1)\cdots\alpha_t(N)$ 的值递推得

$$\alpha_{t+1}(j)=\left[\sum_i^N \alpha_t(i)a_{ij}\right]\times b_j(O_{t+1})$$



s_i 表示状态，~~q_t 表示隐状态~~ q_t 表示时刻 t 状态.

前向算法：(1) 初始化：$\alpha_1(i) = \pi_i b_i(O_1)$，$1 \le i \le N$

(2) 递推计算：$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \times b_j(O_{t+1})$，$1 \le t \le T-1$

(3) 结束，输出：$P(O|\mu) = \sum_{i=1}^{N} \alpha_T(i)$

时间复杂度分析：从 t 时刻所有 N 个状态转移到 s_i 的可能性，每个时刻 t 要计算 N 个前向变量，因此时间复杂度为 $O(N^2T)$

6.4 后向算法

同样引入向量 (1)

后向变量 $\beta_t(i) = P(O_{t+1} O_{t+2} \cdots O_T | q_t = s_i, \mu)$

(2) 用动态归划计算后向量：

(1) 从时刻 t 到 $t+1$，模型由状态 s_i 转移到状态 j，并从 s_j 输出 O_{t+1}

(2) 在时间 $t+1$，状态为 j 的条件下，序列 $O_{t+2} O_{t+3} \cdots O_T$

第一步的概率：$a_{ij} \times b_j(O_{t+1})$

第二步概率：$\beta_{t+1}(j)$

归纳表示：$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \times \beta_{t+1}(j)$

归纳起始：$\beta_T(1)$, $\beta_{T+1}(2)$, $\cdots \beta_1(N)$.

6.5 Viterbi 搜索算法

解决问题2：如何发现状态序列，能够最好的解释观察序列

状态

一种解释：状态序列中的每个状态都单独地具有概率，对于每个时刻 t ($1 \le t \le T$)，寻找 q_t 使得 $\gamma_t(i) = P(q_t = S_i | O, \mu)$ 最大。

$$\gamma_t(i) = P(q_t = S_i | O, \mu) = \frac{P(q_t = S_i, O | \mu)}{P(O|\mu)}$$

分解过程：(1). 模型在时间 t 到达状态 i，并且输出 $O = O_1 O_2 \cdots O_T$

根据前向变量的定义，实现这一步的概率为 $\alpha_t(i)$

(2). 从时间 t，状态 i 出发，模型输出 $O = O_1 O_2 \cdots O_T$

根据后向变量定义，实现这一步的概率为 $\beta_t(i)$

于是 $P(q_t = S_i, O | \mu) = \alpha_t(i) \times \beta_t(i)$

$P(O|\mu)$ 与时间无关，因此 $P(O|\mu) = \sum_{i=1}^{n} \alpha_t(i) \times \beta_t(i)$

即 $$\gamma_t(i) = \frac{\alpha_t(i) \times \beta_t(i)}{\sum_{i=1}^{n} \alpha_t(i) \times \beta_t(i)}$$

因此 t 时刻的最优解状态为 $\hat{q}_t = \arg\max_{1 \le i \le n} (\gamma_t(i))$

问题：两个最优状态之间的转移概率可能为0，因此每一个状态单独最优不能保证整体状态序列最优

另一种解释：在给定 μ 和 O 的条件下求概率最大的状态序列

$$Q = \arg\max P(Q | O, \mu)$$ (防止小数溢出，取对数)

Viterbi 算法：动态搜索最优状态序列

定义：Viterbi 变量 $\delta_t(i)$ 是在时间 t 时，模型沿着某一条路径到达 S_i，输出观察序列 $O = O_1 O_2 \cdots O_t$ 的最大概率为：

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P(q_1, q_2, \cdots, q_t = S_i, O_1 O_2 \cdots O_t | \mu)$$

递归计算 $\delta_{t+1}(j) = \max_i [\delta_t(i) \cdot a_{ij}] \cdot b_j(O_{t+1})$

算法描述

(1). 初始化 $\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N$

$\psi_1(i) = 0$

(2). 递推计算：

$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(O_t) \quad 2 \le t \le T, \ 1 \le j \le N$

$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(O_t), \quad 2 \le t \le T, \ 1 \le i \le N$

(3). 结束，回溯得到状态序列

$q_t = \psi_{t+1}(q_{t+1})$

时间复杂度：$O(N^2 T)$

6.6 参数学习

解决问题3：如何调节模型 λ 的参数，使得 $P(O|\lambda)$ 最大。

前向后向算法（Baum-Welch or forward-backward procedure）

（如果状态序列已知）

$\pi_i = \delta(q_1, s_i)$

$\bar{a}_{ij} = \dfrac{\text{Q中从状态 } q_i \text{ 转移到 } q_j \text{ 的次数}}{\text{Q中从状态 } q_i \text{ 转移到另一状态（包括自己）的次数}}$

$= \dfrac{\sum_{t=1}^{T-1} \delta(q_t, s_i) \times \delta(q_{t+1}, s_j)}{\sum_{t=1}^{T-1} \delta(q_t, s_i)}$

$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & \text{else} \end{cases}$ 在 kronecker 克罗内克积 处取 1

类似地，$\bar{b}_j(k) = \dfrac{\sum_{t=1}^{T} \delta(q_t, s_j) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \delta(q_t, s_j)}$

否则，如果不存在大量标注的样本（实际很少用，效果较差）：

① EM（期望值最大化）：随机赋值迭代计算。

6.7 HMM 应用举例

1. 汉语的自动分词与词性标注.

汉语分词结果：$O=o_1o_2\cdots o_T$

即求解 $\theta=\arg\max_{\theta} P(O|\mu)$

词性标注：求解 $Q=\arg\max_{Q} P(Q|O,\mu)$

1). 估计 HMM $\mu=(A,B,\pi)$ 的参数

2). 基于任意给定的一个输入句子及其可能的输出序列 O，求出所有可能的 O 中使 $P(O|\mu)$ 最大的解

3). 快速地选择最优状态序列

问题 1). 模型参数

① 观察序列：单词序列

② 状态序列：词类标注序列

③ 状态数目 N

④. 输出符号数 M

如果无标注语料，需要有词性标注词典采用 EM 迭代（无监督）.

应引有指导学习：词语大仙来指计参数.

$$\bar{\pi}_{pos_i}=\frac{pos_i\text{出现在句首}}{\text{语句的首}}$$

$$\bar{a}_{ij}=\frac{pos_i\to pos_j}{pos_i\to pos}$$

$$\bar{b}_j(k)=\frac{pos_j\to w_k}{pos_j}$$

问题 2). 观察序列

维特比出所有可能切分.

~

（老师说非常简单？？可我不会啊QAQ）

6.8 CRFs 及其应用（条件随机场）

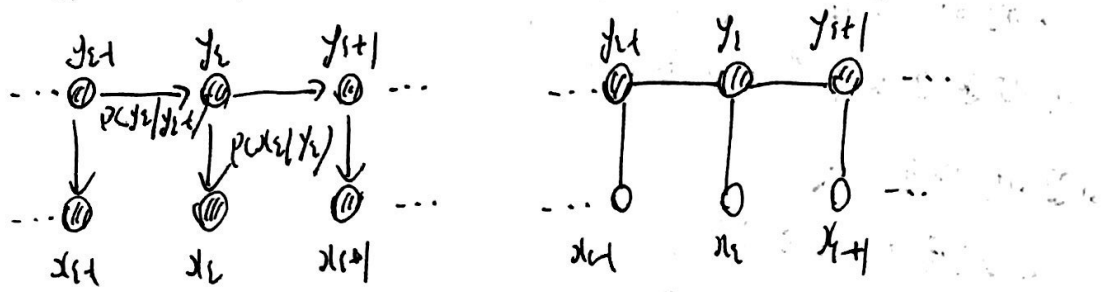
基本思路：给定观察序列 X，输出标记序列 Y，通过计算 $P(Y|X)$ 求解最优标注序列。

设 $G=(V,E)$ 是一个无向图，$Y=\{Y_v \mid v\in V\}$，即 V 中每个结点对应于一个随机变量 Y_v，取值范围为可能的标记集合 $\{y\}$

如果以观察序列 X 为条件，每个随机变量 Y_v 都满足以下马尔可夫特性：

$$P(Y_v|X, Y_w, w\neq v) = P(Y_v|X, Y_w, w\sim v)$$

其中 $w\sim v$ 表示两个结点在图中是邻近结点，即 (X,Y) 为一个条件随机场 G 的图结构是任意的



HMMs　　CRFs

CRFs 的 X 节点并不是由模型生成的，本身存在，只是给打标签而已。

给定观察序列 X，某个特定标记序列 Y 的概率可以定义为：

$$\exp\Big(\sum_j \lambda_j \underbrace{t_j(y_{i-1}, y_i, X, i)}_{\text{转移概率}} + \sum_k \mu_k \underbrace{s_k(y_i, X, i)}_{\text{标记概率}}\Big)$$

λ_j、μ_k 由训练样本估计得出。

$b(X,i)$：X 的位置为某个特定的词

$$b(X,i) = \begin{cases} 1 & \\ 0 & \text{else} \end{cases}$$

$s(y_{i-1}, y_i, X, i)$：y_{i-1} 无法与 y_i 保证形式上一样。

$$t_j(y_{i-1}, y_i, X, i) = \begin{cases} b(X,i) & y_{i-1} \text{ 和 } y_i \text{ 满足某种条件} \\ 0 & \text{else} \end{cases}$$

$$F_j(Y,X) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, X, i)$$

↓ 表示：S 或 T。

$$P(Y|X) = \frac{1}{Z(X)} \exp\Big(\sum_j \lambda_j \cdot F_j(Y,X)\Big)$$

$Z(X)$ 为归一化因子