隐马尔可夫模型

一、隐马尔可夫模型HMM

隐马尔可夫模型（Hidden Markov Model，HMM）是可用于标注问题的统计学模型，描述了由隐马尔可夫链随机生成观测序列的过程，属于生成模型。

隐马尔可夫模型一般应用于时间序列建模，例如语音识别、自然语言处理、生物信息、机器学习等。

- 隐马尔可夫模型是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态序列，再由各个状态生成一个观测而产生观测序列的过程。
- 隐藏的马尔可夫链随机生成的状态的序列，称为状态序列（state sequence）；
- 每个状态生成一个观测，而由全部状态序列生成全部观测序列，称为观测序列（observation sequence）。

1.1 基本定义

- 设 $Q=\{q_1,q_2,\cdots,q_N\}$ 是所有可能的状态的集合，$V=\{v_1,v_2,\cdots,v_M\}$ 是所有可能的观测的集合，其中状态的总数为 N，观测的总数为 M。
  - 状态序列和观测序列，长度均为固定值 T；
  - 状态序列是不可观测的，观测序列是可观测的；
  - 状态序列是一个马尔可夫链，而观测序列则是由状态序列随机生成的。
- 设 $I=(i_1,i_2,\cdots,i_T)$ 是长度为 T 的状态序列，$O=(o_1,o_2,\cdots,o_T)$ 是对应的观测序列。
  - 设 $A=[a_{ij}]_{N\times N}$ 是状态转移矩阵，其中 $a_{ij}=P(i_{t+1}=q_j|i_t=q_i)$ 是在时刻 t 处于状态 q_i 的条件下在时刻 $t+1$ 转移到状态 q_j 的概率；
  - 设 $B=[b_{ij}]_{N\times M}$ 是观测概率矩阵，其中 $b_{ij}=P(o_t=v_k|i_t=q_j)$ 是在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率；
  - 设 $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$ 是初始状态概率向量，其中 $\pi_i=P(i_1=q_i)$ 是在时刻 $t=1$ 处于状态 q_i 的概率。
- 状态转移概率矩阵 A：

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}$$

其中 $a_{ij}=P(i_{t+1}=q_j|i_t=q_i)$，表示在时刻 t 处于状态 q_i 的条件下，在时刻 $t+1$ 转移到状态 q_j 的概率。

- 状态 b_j 的观测概率：

$$B=\begin{bmatrix} b_1(v_1) & b_1(v_2) & \cdots & b_1(v_M) \\ b_2(v_1) & b_2(v_2) & \cdots & b_2(v_M) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(v_1) & b_N(v_2) & \cdots & b_N(v_M) \end{bmatrix}$$

其中 $b_j(v_k)=P(o_t=v_k|i_t=q_j)$，表示在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率。

- 设 B 是观测概率矩阵，$B=[b_1,b_2,\cdots,b_N]$，其中，$b_j=(b_j(v_1),b_j(v_2),\cdots,b_j(v_M))^T$，$j=1,2,\cdots,N$，表示在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率。
- 隐马尔可夫模型 λ 可以用三元符号表示，即 $\lambda=(A,B,\pi)$。A,B,π 称为隐马尔可夫模型的三要素。

隐马尔可夫模型作了两个基本假设：

- 齐次性假设：隐藏的马尔可夫链在任意时刻 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，也与时刻 t 无关。
- 观测独立性假设：任意时刻的观测只依赖该时刻的马尔可夫链的状态，与其他观测及状态无关。

隐马尔可夫模型可以用于标注问题，状态序列对应着标注序列。

- 这时，观测是一个文字序列，隐藏的状态序列是一个标注序列。
- 可以建立隐马尔可夫模型来进行标注，这时状态的个数等于标注的标记集合中的标记个数。

1.2 生成算法

隐马尔可夫模型可以用于生成一个观测序列的过程。

- 设隐马尔可夫模型 $\lambda=(A,B,\pi)$，长度为 T 的观测序列 $O=(o_1,o_2,\cdots,o_T)$，可按照如下方法随机生成：
  - 按照 π 分布产生初始状态 i_1；
  - 令 $t=1$；
  - 按照状态 i_t 的观测概率分布 $b_{i_t}(k)$ 产生 o_t；
  - 按照状态 i_t 的转移概率分布 $\{a_{i_ti_{t+1}}\}$ 产生 i_{t+1}；
  - $t=t+1$；
  - 如果 $t<T$，则转到步骤2，否则结束。

2、HMM 基本问题

- 隐马尔可夫模型的三个基本问题：
  - 给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，计算在模型下出现观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率 $P(O|\lambda)$，称为概率计算问题（评估问题）；
  - 学习问题：已知观测序列 $O=(o_1,o_2,\cdots,o_T)$，估计模型 $\lambda=(A,B,\pi)$ 的参数，使得在模型下观测序列 O 的概率 $P(O|\lambda)$ 最大，即用极大似然估计的方法估计参数；
  - 预测问题（解码问题）：已知模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，求对应的状态序列 $I=(i_1,i_2,\cdots,i_T)$，使得条件概率 $P(I|O)$ 最大，即利用极大后验概率的准则进行预测。
- 注：生成问题与预测问题，计算问题与学习问题，是三大基本问题的两个对应的一对。例如预测问题是生成问题的反问题，而计算问题是学习问题的子问题。

2.1 概率计算问题

- 给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，计算在模型 λ 下观测序列 O 出现的概率 $P(O|\lambda)$。
- 最直接的方法是按照概率公式直接计算。通过列举所有可能的长度为 T 的状态序列 $I=(i_1,i_2,\cdots,i_T)$，求各个状态序列 I 与观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的联合概率 $P(O,I|\lambda)$，然后对所有可能的状态序列求和，得到 $P(O|\lambda)$。
- 状态序列 $I=(i_1,i_2,\cdots,i_T)$ 的概率是：

$$P(I|\lambda)=\pi_{i_1}a_{i_1i_2}a_{i_2i_3}\cdots a_{i_{T-1}i_T}$$

- 对固定的状态序列 $I=(i_1,i_2,\cdots,i_T)$，观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率是：

$$P(O|I,\lambda)=b_{i_1}(o_1)b_{i_2}(o_2)\cdots b_{i_T}(o_T)$$

- 状态序列 $I=(i_1,i_2,\cdots,i_T)$ 与观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率是：

$$P(O,I|\lambda)=P(I|\lambda)P(O|I,\lambda)=\pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)$$

- 对所有可能的状态序列求和，得到观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率 $P(O|\lambda)$，即：

$$P(O|\lambda)=\sum_I P(O,I|\lambda)=\sum_I P(I|\lambda)P(O|I,\lambda)=\sum_I \pi_{i_1}b_{i_1}(o_1)a_{i_1i_2}b_{i_2}(o_2)\cdots a_{i_{T-1}i_T}b_{i_T}(o_T)$$

- 上式的计算量很大的话，$O(TN^T)$ 阶的。因此需要简化计算。

2.1.1 前向算法

- 给定隐马尔可夫模型 $\lambda=(A,B,\pi)$，定义到时刻 t 部分观测序列 $o_1,o_2,\cdots,o_t$ 且在时刻 t 时处于状态 q_i 的概率为前向概率：$\alpha(t,i)=P(o_1,o_2,\cdots,o_t,i_t=q_i|\lambda)$。
- 前向算法（forward algorithm）是基于“状态序列的马尔可夫性质”的一种递归算法。
  - 输入：隐马尔可夫模型 $\lambda=(A,B,\pi)$，观测序列 $O=(o_1,o_2,\cdots,o_T)$；
  - 输出：观测序列概率 $P(O|\lambda)$；
  - 算法内容：
    - 初值：$\alpha(1,i)=\pi_ib_i(o_1),\quad i=1,2,\cdots,N$
    - 递推：对 $t=1,2,\cdots,T-1$，

$$\alpha(t+1,i)=\left[\sum_{j=1}^N \alpha(t,j)a_{ji}\right]b_i(o_{t+1}),\quad i=1,2,\cdots,N$$

    - 终止：$P(O|\lambda)=\sum_{i=1}^N \alpha(T,i)$
- 前向算法利用前向概率的递推关系，将计算 $P(O|\lambda)$ 的计算量从 $O(TN^T)$ 降低到了 $O(TN^2)$ 阶的计算量，因此是有效的。
- 前向算法的关键是：每一时刻的一个状态，需要该时刻的所有可能的状态转移到这个状态的概率的乘积的和。

2.1.2 后向算法

- 给定隐马尔可夫模型 $\lambda=(A,B,\pi)$，定义到时刻 t 部分观测序列 $o_{t+1},o_{t+2},\cdots,o_T$ 且在时刻 t 时处于状态 q_i 的概率为后向概率：$\beta(t,i)=P(o_{t+1},o_{t+2},\cdots,o_T,i_t=q_i|\lambda)$。
- 后向算法（backward algorithm）是基于“状态序列的马尔可夫性质”的一种递归算法。
  - 输入：隐马尔可夫模型 $\lambda=(A,B,\pi)$，观测序列 $O=(o_1,o_2,\cdots,o_T)$；
  - 输出：观测序列概率 $P(O|\lambda)$；
  - 算法内容：
    - 初值：$\beta(T,i)=1,\quad i=1,2,\cdots,N$
    - 递推：对 $t=T-1,T-2,\cdots,1$，

$$\beta(t,i)=\sum_{j=1}^N a_{ij}b_j(o_{t+1})\beta(t+1,j),\quad i=1,2,\cdots,N$$

    - 终止：$P(O|\lambda)=\sum_{i=1}^N \pi_ib_i(o_1)\beta(1,i)$
- 后向算法利用前向概率的递推关系，将计算 $P(O|\lambda)$ 的计算量从 $O(TN^T)$ 降低到了 $O(TN^2)$ 阶的计算量，因此是有效的。

2.1.3 两个算法的关系

- 前向概率和后向概率的定义可以推广到任意时刻 t 和状态 i 的前向概率和后向概率：

$$P(O,i_t=q_i|\lambda)=\alpha(t,i)\beta(t,i)$$

- 证明：将 O 拆成 $o_1,o_2,\cdots,o_t$ 和 $o_{t+1},o_{t+2},\cdots,o_T$，则前向概率为：

$$\alpha(t,i)=P(o_1,o_2,\cdots,o_t,i_t=q_i|\lambda)$$

- 后向概率为：

$$\beta(t,i)=P(o_{t+1},o_{t+2},\cdots,o_T|i_t=q_i,\lambda)$$

- 则有：

$$P(O,i_t=q_i|\lambda)=P(o_1,o_2,\cdots,o_t,o_{t+1},o_{t+2},\cdots,o_T,i_t=q_i|\lambda)$$

- 利用条件概率公式，则有：

$$P(O,i_t=q_i|\lambda)=P(o_1,o_2,\cdots,o_t,i_t=q_i|\lambda)P(o_{t+1},o_{t+2},\cdots,o_T|o_1,o_2,\cdots,o_t,i_t=q_i,\lambda)$$

- 由于条件独立性，则有：

$$P(O,i_t=q_i|\lambda)=\alpha(t,i)P(o_{t+1},o_{t+2},\cdots,o_T|i_t=q_i,\lambda)$$

- 则有：

$$P(o_{t+1},o_{t+2},\cdots,o_T|i_t=q_i,\lambda)=\beta(t,i)$$

- 则有：

$$P(O,i_t=q_i|\lambda)=\alpha(t,i)\beta(t,i)$$

- 根据前向概率和后向概率的定义，可以推出两种单个观测的概率：

$$P(O|\lambda)=\sum_{i=1}^N \alpha(t,i)\beta(t,i)=\sum_{i=1}^N \alpha(T,i)\beta(T,i)=\sum_{i=1}^N \alpha(1,i)\beta(1,i)$$

- 当 $t=1$ 时，表示总的概率；当 $t=T$ 时，则是一种特殊情况。
- 对应的一个单个观测的概率：

$$P(O,i_t=q_i|\lambda)=\sum_{i=1}^N \alpha(t,i)\beta(t,i)$$

- 这个公式在计算某个特定状态的情况下，计算某个观测的概率，这种情况在后面会用到。

2.2 学习问题

- 给定一个观测序列 $O=(o_1,o_2,\cdots,o_T)$，估计模型 $\lambda=(A,B,\pi)$ 的参数，使得在该模型下观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率 $P(O|\lambda)$ 最大。即用极大似然估计的方法估计参数。
- 隐马尔可夫模型的学习，根据训练数据是包括观测序列和对应的状态序列还是只有观测序列，可以分别由监督学习与无监督学习实现。
- 首先讲监督学习的极大似然估计；然后讲无监督学习的Baum-Welch算法（实际是极大似然估计的应用）；最后讲一般的无监督学习的问题。

2.2.1 监督学习

- 假设训练数据为 $\{(o_1,i_1),(o_2,i_2),\cdots,(o_N,i_N)\}$，其中，$o_j=(o_{j1},o_{j2},\cdots,o_{jT})$，$i_j=(i_{j1},i_{j2},\cdots,i_{jT})$，$j=1,2,\cdots,N$，表示第 j 个观测序列，以及第 j 个观测序列对应的状态序列。
- 这时，可以利用极大似然估计来估计隐马尔可夫模型的参数。
- 监督学习的极大似然估计的步骤：
  - 步骤1：计算转移概率 a_{ij} 的估计值 $\hat{a}_{ij}$：
    - 设样本中时刻 t 处于状态 i 且时刻 $t+1$ 转移到状态 j 的频数为 A_{ij}，则有：

$$\hat{a}_{ij}=\frac{A_{ij}}{\sum_{j=1}^N A_{ij}},\quad i=1,2,\cdots,N;\quad j=1,2,\cdots,N$$

    - 其中，$\sum_{j=1}^N A_{ij}$ 是状态 i 的频数。
  - 步骤2：计算观测概率 $b_j(k)$ 的估计值 $\hat{b}_j(k)$：
    - 设样本中状态为 j 并观测为 k 的频数为 B_{jk}，则有：

$$\hat{b}_j(k)=\frac{B_{jk}}{\sum_{k=1}^M B_{jk}},\quad j=1,2,\cdots,N;\quad k=1,2,\cdots,M$$

    - 其中，$\sum_{k=1}^M B_{jk}$ 是状态 j 的频数。
  - 步骤3：计算初始状态概率 π_i 的估计值 $\hat{\pi}_i$：
    - 设样本中时刻 $t=1$ 处于状态 i 的频数为 C_i，则有：

$$\hat{\pi}_i=\frac{C_i}{N},\quad i=1,2,\cdots,N$$

    - 其中，N 是样本容量。
- 监督学习的极大似然估计的优点：
  - 简单。
  - 但是，这种方法依赖于训练数据是否拥有足够多的信息，以保证能够估计出概率。
  - 对于没有观测的事件，则会出现概率为零的情况。

2.2.2 无监督学习

- 无监督学习的数据，只有观测序列 $O=(o_1,o_2,\cdots,o_T)$，没有对应的状态序列。
- 这时，可以利用极大似然估计来估计隐马尔可夫模型的参数。
- 隐马尔可夫模型的学习，根本方法是极大似然估计。即极大化观测序列 $O=(o_1,o_2,\cdots,o_T)$ 的概率 $P(O|\lambda)$。
- 由于含有隐变量，所以可以考虑用EM算法来求解。
- 隐马尔可夫模型的EM算法的推导：
  - 首先，确定完全数据的对数似然函数：
    - 完全数据是 $(o_1,o_2,\cdots,o_T,i_1,i_2,\cdots,i_T)$，则完全数据的对数似然函数是 $\log P(O,I|\lambda)$。
  - 其次，写出Q函数：
    - 完全数据的对数似然函数关于在给定观测数据和当前参数下对未知数据的条件概率分布的期望，称为Q函数，即 $Q(\lambda,\bar{\lambda})=\sum_I \log P(O,I|\lambda)P(O,I|\bar{\lambda})$，其中，$\bar{\lambda}$ 是HMM的当前参数估计值，λ 是要极大化的HMM的参数。
  - 最后，极大化Q函数：
    - 求出使 $Q(\lambda,\bar{\lambda})$ 极大化的 λ，即 $\bar{\lambda}\leftarrow\arg\max_\lambda Q(\lambda,\bar{\lambda})$，这样产生一个新的HMM参数估计值 $\bar{\lambda}$。
    - 重复上述两个步骤，直到收敛为止。
- 无监督学习的Baum-Welch算法：
  - 输入：观测序列 $O=(o_1,o_2,\cdots,o_T)$；
  - 输出：隐马尔可夫模型 $\lambda=(A,B,\pi)$。
  - 算法内容：
    - 初值：$\pi_i=\frac{1}{N}$，$a_{ij}=\frac{1}{N}$，$b_j(k)=\frac{1}{M}$，其中，$i,j=1,2,\cdots,N$；$k=1,2,\cdots,M$
    - 对 $t=1,2,\cdots,N$ 和观测 $o_t=v_k$，计算：
      - 在模型 $\lambda=(A,B,\pi)$ 下，定义 $\gamma(i,t)=P(i_t=q_i|O)$，表示在给定观测序列 O 的条件下，在时刻 t 处于状态 q_i 的概率；
      - 在模型 $\lambda=(A,B,\pi)$ 下，定义 $\xi(i,j)=P(i_t=q_i,i_{t+1}=q_j|O)$，表示在给定观测序列 O 的条件下，在时刻 t 处于状态 q_i 且在时刻 $t+1$ 处于状态 q_j 的概率；
      - 对 $i=1,2,\cdots,N$，计算 $\pi_i=\gamma(i,1)$；
      - 对 $i,j=1,2,\cdots,N$，计算 $a_{ij}=\frac{\sum_{t=1}^{T-1}\xi(i,j)}{\sum_{t=1}^{T-1}\gamma(i,t)}$；
      - 对 $j=1,2,\cdots,N$，$k=1,2,\cdots,M$，计算 $b_j(k)=\frac{\sum_{t=1,o_t=v_k}^{T}\gamma(j,t)}{\sum_{t=1}^{T}\gamma(j,t)}$；
    - 重复上述步骤，直到收敛为止。
- 无监督学习的Baum-Welch算法的推导：
  - 对于 $t=1,2,\cdots,N$，定义：

$$\gamma(i,t)=\sum_{j=1}^N \xi(i,j)$$

  - 则有：

$$\sum_{t=1}^{T-1}\gamma(i,t)=\sum_{t=1}^{T-1}\sum_{j=1}^N \xi(i,j)$$

  - 对 $t=1,2,\cdots,N$，$i,j=1,2,\cdots,N$，定义：

$$\xi(i,j)=P(i_t=q_i,i_{t+1}=q_j|O)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{P(O|\lambda)}$$

  - 则有：

$$\xi(i,j)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{\sum_{i=1}^N\sum_{j=1}^N P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}$$

  - 则有：

$$\xi(i,j)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{P(O|\lambda)\sum_{i=1}^N\sum_{j=1}^N P(i_t=q_i,i_{t+1}=q_j|O,\lambda)}$$

  - 则有：

$$\xi(i,j)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{\sum_{i=1}^N\sum_{j=1}^N P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}$$

  - 则有：

$$\xi(i,j)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{\sum_{i=1}^N P(i_t=q_i,O|\lambda)}$$

  - 则有：

$$\xi(i,j)=\frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{\gamma(i,t)}$$

  - 则有：

$$P(i_t=q_i,i_{t+1}=q_j,O|\lambda)=\xi(i,j)\gamma(i,t)$$

  - 则有：

$$P(i_{t+1}=q_j,O|i_t=q_i,\lambda)=\frac{\xi(i,j)}{\gamma(i,t)}$$

  - 则有：

$$P(i_{t+1}=q_j|i_t=q_i,O,\lambda)=\frac{\xi(i,j)}{\gamma(i,t)}$$

  - 则有：

$$a_{ij}=\frac{\sum_{t=1}^{T-1}\xi(i,j)}{\sum_{t=1}^{T-1}\gamma(i,t)}$$

  - 则有：

$$\pi_i=\gamma(i,1)$$

  - 则有：

$$b_j(k)=\frac{\sum_{t=1,o_t=v_k}^{T}\gamma(j,t)}{\sum_{t=1}^{T}\gamma(j,t)}$$

  - 则有：

$$\sum_{t=1}^{T}\gamma(j,t)=\sum_{t=1}^{T-1}\gamma(j,t)+\gamma(j,T)$$

  - 则有：

$$\sum_{t=1}^{T}\gamma(j,t)=\sum_{t=1}^{T-1}\gamma(j,t)+\sum_{i=1}^N \xi(i,j)$$

  - 则有：

$$\sum_{t=1}^{T}\gamma(j,t)=\sum_{i=1}^N \xi(i,j)$$

  - 则有：

$$\sum_{t=1,o_t=v_k}^{T}\gamma(j,t)=\sum_{i=1}^N \xi(i,j)$$

  - 则有：

$$b_j(k)=\frac{\sum_{i=1}^N \xi(i,j)}{\sum_{i=1}^N \xi(i,j)}$$

  - 则有：

$$b_j(k)=1$$

  - 则有：

$$\sum_{i=1}^N \xi(i,j)=\sum_{i=1}^N \xi(i,j)

- 实际上，状态 1 倾向于转换到状态 2 ；同时状态 2 也倾向于留在状态 2 。但是由于状态 2 可以转化出去的状态较多，从而使得转移概率均比较小。

  而维特比算法得到的最优路径全部停留在状态 1 ，这样与实际不符。
- `MEMM`倾向于选择拥有更少转移的状态，这就是标记偏置问题。

6. 标记偏置问题的原因是：计算 $P_{\vec{\mathbf{w}}}(i_t \mid i_{t-1}, o_t)$ 仅考虑局部归一化，它仅仅考虑指定位置的所有特征函数。
   - 如上图中，$P_{\vec{\mathbf{w}}}(i_t \mid i_{t-1}=2, o_t)$ 只考虑在 $(i_{t-1}=2, o_t)$ 这个结点的归一化。
     - 对于 $(i_{t-1}=2, o_t)$ ，其转出状态较多，因此每个转出概率都较小。
     - 对于 $(i_{t-1}=1, o_t)$ ，其转出状态较少，因此每个转出概率都较大。
   - `CRF`解决了标记偏置问题，因为`CRF`是全局归一化的：

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=1}^{n-1}\lambda_j t_j(Y_i, Y_{i+1}, \mathbf{X}, i) + \sum_{k=1}^{K_2}\sum_{i=1}^{n}\mu_k s_k(Y_i, \mathbf{X}, i)\right)$$

   它考虑了所有位置、所有特征函数。