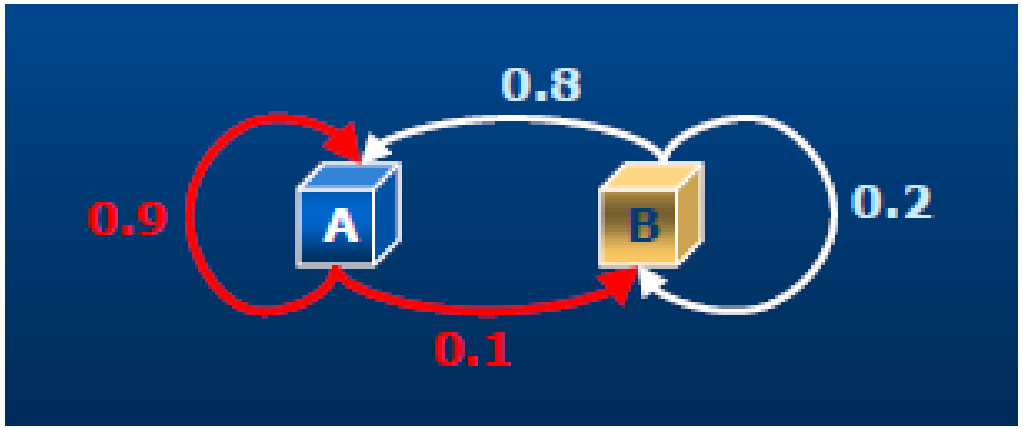
开课吧
kaikeba.com
慧科集团旗下企业

HMM与分词、词性标注、实体识别

隐马尔可夫模型（**hidden Markov model , 记作：HMM**）是马尔可夫模型的进一步发展。语音识别、生物信息识别、自然语言处理。

隐马尔可夫模型的示例—赌场欺诈问题:

某赌场在投骰子，根据点数决定胜负。在多次投掷骰子的时候采取了如下手段进行作弊：准备了两个骰子A和B，其中A为正常骰子，B为灌铅骰子，由于怕被发现，所有连续投掷的时候偶尔使用一下B，A和B之间转换的概率如下：



A 和 B 之间相互转换的概率写成矩阵如下：

	正常骰子 A	灌铅骰子 B
正常骰子 A	0.9	0.1
灌铅骰子 B	0.8	0.2

A 和 B 产生各观测值概率的区别为：

观测值	1	2	3	4	5	6
正常骰子 A	1/6	1/6	1/6	1/6	1/6	1/6
灌铅骰子 B	0	1/8	1/8	3/16	3/16	3/8

骰子作弊问题模型化：

作弊问题由 5 个部分构成：

（1）隐状态空间 S (状态空间)：

$S=\{$正常骰子A，灌铅骰子B$\}$，赌场具体使用哪个骰子，赌徒是不知道的。

（2）观测空间 O：$O=\{1,2,3,4,5,6\}$。正常骰子 A 和灌铅骰子 B 的所有六个面可能取值。

（3）初始状态概率空间 π：

π ={初始选择正常骰子的概率,初始选择灌铅骰子的概率}。

（4）隐状态转移概率矩阵 $P_{2\times2}$：

	正常骰子 A	灌铅骰子 B
正常骰子 A	0.9	0.1
灌铅骰子 B	0.8	0.2

（5）观测值生成概率矩阵 $Q_{2\times6}$：

观测值	1	2	3	4	5	6
正常骰子 A	1/6	1/6	1/6	1/6	1/6	1/6
灌铅骰子 B	0	1/8	1/8	3/16	3/16	3/8

隐马尔可夫模型的定义：

隐马尔科夫模型由以下五部分构成：

（1）隐状态空间 S (状态空间)：

$S=\{S_1,S_2,S_3,\cdots,S_N\}$，其中 N 为状态的数目。

（2）观测空间 O：$O=\{O_1,O_2,O_3,\cdots,O_M\}$，M 为状态对应的观测值的数目。

（3）初始状态概率空间 π：$\pi=\{\pi_1,\pi_2,\pi_3,\cdots,\pi_N\}$，其中

$$\pi_i=P\{X_1=S_i\} \quad (1\le i\le N)$$

（4）隐状态转移概率矩阵 $P_{N\times N}$：

$$P_{N\times N}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix}$$

（5）观测值生成概率矩阵 $Q_{N\times M}$：

$$Q_{N\times M}=\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1M} \\ q_{21} & q_{22} & \cdots & q_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ q_{N1} & q_{N2} & \cdots & q_{NM} \end{bmatrix}$$

记 HMM 为： $\lambda=(S,O,\pi,P,Q)$ **或简写为** $\lambda=(\pi,P,Q)$ **。**

解码问题（decoding）

对于骰子作弊问题中，**解码问题**是：如果确实使用了作弊骰子，这些序列中哪些点时由B投掷出来的。

对于一般的隐马尔可夫模型中，解码问题是给定模型 $\lambda=(\pi,P,Q)$ 和一个观测序列 $v=\{v_1,v_2,v_3,\cdots,v_i,\cdots,v_n\}$，求出模型 $\lambda=(\pi,P,Q)$ 生成 $v=\{v_1,v_2,v_3,\cdots,v_i,\cdots,v_n\}$ 的最有可能状态 $X=\{X_1,X_2,X_3,\cdots,X_i,\cdots,X_n\}$。即推测出隐藏层的状态，也就是解码。

Viterbi算法

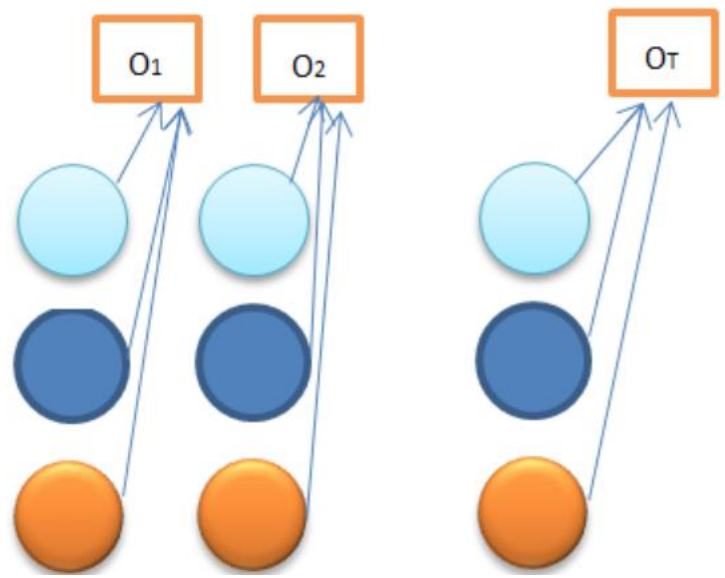
思想：定义一个路径最优变量，然后采取递推的方式迭代，进而降低计算量。

路径最优变量：

$$\delta_t(i) = \max_{X_1 X_2 \cdots X_{t-1}} P(X_1, X_2, \cdots X_{t-1}, X_t = S_i, v_1, v_2, \cdots v_t \mid \lambda)$$

表示在时刻 **t** 沿着一条路径 $X_1, X_2, \cdots X_{t-1}, X_t$，且在 **t** 时刻的状态为 $X_t = S_i$ 产生出观测序列 $v_1, v_2, \cdots v_t$ 的最大概率。

另外，为了寻找路径，我们定义一个 $\phi_t(i)$ 专门记录 **t** 时刻状态 S_i 最有可能由哪个状态转移而来。



状态

观察值

例 10.2 考虑盒子和球模型 $\lambda=(A,B,\pi)$，状态集合 $Q=\{1,2,3\}$，观测集合 $V=$ {红，白}，

$$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.4,0.4)^{\mathrm{T}}$$

已知观测序列 $O=$(红, 白, 红)，试求最优状态序列，即最优路径 $I^*=(i_1^*,i_2^*,i_3^*)$.

初始概率：π=（0.2，0.4，0.4）

解答过程：

①初始化，拿到红球

$$\delta_1 = \pi_1 * P(red|1) = 0.2 * 0.5 = 0.1$$
$$\psi_1 = 0$$
$$\delta_2 = \pi_2 * P(red|2) = 0.4 * 0.4 = 0.16$$
$$\psi_2 = 0$$
$$\delta_3 = \pi_3 * P(red|3) = 0.4 * 0.7 = 0.28$$
$$\psi_3 = 0$$

$$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.4,0.4)^{\mathrm{T}}$$

已知观测序列 $O=$(红, 白, 红)，试求最优状态序列，即最优路径 $I^*=(i_1^*,i_2^*,i_3^*)$.

$$\begin{aligned}\delta_2(1)&=\max_{1\leqslant j\leqslant 3}[\delta_1(j)a_{j1}]b_1(o_2)\\&=\max_j\{0.10\times0.5,0.16\times0.3,0.28\times0.2\}\times0.5\\&=0.028\end{aligned}$$

$$\psi_2(1)=3$$

$$\delta_2(2)=0.0504,\quad \psi_2(2)=3$$

$$\delta_2(3)=0.042,\quad \psi_2(3)=3$$

$$A=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix},\quad B=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix},\quad \pi=(0.2,0.4,0.4)^{\mathrm{T}}$$

已知观测序列$O=$(红, 白, 红)，试求最优状态序列，即最优路径$I^*=(i_1^*,i_2^*,i_3^*)$.

同样，在$t=3$时，

$$\delta_3(i)=\max_{1\leqslant j\leqslant 3}[\delta_2(j)a_{ji}]b_i(o_3)$$

$$\psi_3(i)=\arg\max_{1\leqslant j\leqslant 3}[\delta_2(j)a_{ji}]$$

$$\delta_3(1)=0.00756,\quad \psi_3(1)=2$$

$$\delta_3(2)=0.01008,\quad \psi_3(2)=2$$

$$\delta_3(3)=0.0147,\quad \psi_3(3)=3$$

- 输入的是：

- 我是机器学习从业者

- 输出为：
- 我 是 机器 学习 从业者

在HMM模型中文分词中，我们的输入是一个句子(也就是观察值序列)，输出是这个句子中每个字的状态值。 比如:

小明硕士毕业于中国科学院计算所

输出的状态序列为：

BEBEBMEBEBMEBES

根据这个状态序列我们可以进行切词:

BE/BE/BME/BE/BME/BE/S

所以切词结果如下:

小明/硕士/毕业于/中国/科学院/计算/所

我们可以注意到：B后面只可能接(M or E)，不可能接(B or S)。而M后面也只可能接(M or E)，不可能接(B, S)。输入输出都明确了，下文讲讲输入和输出之间的具体过程。

HMM是一个三元组（pi，A，B）：

初始化概率向量：$\Pi = (\pi_i)$

状态转移矩阵：$A = (a_{ij})$ $Pr(x_{i_t} | x_{j_{t-1}})$

混淆矩阵：$B = (b_{ij})$ $Pr(y_i | x_j)$

初始化概率向量：

示例如下：

```
: Pi_dic
```

```
: {'B': 0.5820149148537713, 'E': 0.0, 'M': 0.0, 'S': 0.41798844132394497}
```

示例数值是对概率值取对数之后的结果(可以让概率相乘的计算变成对数相加)，其中-3.14e+100作为负无穷，也就是对应的概率值是0。下同。

也就是句子的第一个字属于{B,E,M,S}这四种状态的概率，如上可以看出，E和M的概率都是0，这和实际相符合，开头的第一个字只可能是词语的首字(B)，或者是单字成词(S)。

状态转移矩阵：

转移概率是马尔科夫链很重要的一个知识点，大学里面学过概率论的人都知道，马尔科夫链最大的特点就是当前T=i时刻的状态Status(i)，只和T=i时刻之前的n个状态有关。也就是: {Status(i-1),Status(i-2),Status(i-3),...Status(i-n)}

更进一步的说，HMM模型有三个基本假设作为模型的前提，其中有个【有限历史性假设】，也就是马尔科夫链的n=1。即Status(i)只和Status(i-1)相关，这个假设能大大简化问题。

回过头看转移矩阵，其实就是一个4x4(4就是状态值集合的大小)的二维矩阵，示例如下：

```
A_dic
```

```
{'B': {'B': 0.0, 'E': 1226466.0, 'M': 162066.0, 'S': 0.0},
 'E': {'B': 575846.0, 'E': 0.0, 'M': 0.0, 'S': 650620.0},
 'M': {'B': 0.0, 'E': 0.0, 'M': 62332.0, 'S': 162066.0},
 'S': {'B': 639270.0, 'E': 0.0, 'M': 0.0, 'S': 834753.0}}
```

矩阵的横坐标和纵坐标顺序是BEMS x BEMS。(数值是概率求对数后的值，别忘了。)

比如A[0][0]代表的含义就是从状态B转移到状态B的概率，由A[0][0] = -3.14e+100可知，这个转移概率是0，这符合常理。由状态各自的含义可知，状态B的下一个状态只可能是ME，不可能是BS，所以不可能的转移对应的概率都是0，也就是对数值负无穷，在此记为-3.14e+100。

混淆矩阵：

混淆矩阵中，每个元素其实也是一个条件概率而已，根据HMM模型三个基本假设(哪三个请看文末备注)里的【观察值独立性假设】，观察值只取决于当前状态值，也就是:

$$P(Observed[i], Status[j]) = P(Status[j]) * P(Observed[i]|Status[j])$$

其中P(Observed[i]|Status[j])这个值就是从混淆矩阵中获取。

混淆矩阵示例如下：

```
{'B': {'逸': 0.0,
  '乘': 0.00012459201516421 66,
  '阅': 5.6174434582710375e-05,
  '泄': 0.0,
  '筏': 0.0,
  '堡': 9.362405763785063e-06,
  '甜': 2.5206477056344398e-05,
  '揽': 0.0,
  '河': 0.00036009252937634854,
  '功': 0.00023910143950589544,
  '漫': 5.0412954112688797e-05,
```