实验二 - 隐马尔科夫模型

李一鸣

2018 年 10 月 12 日

马尔科夫链的生成（隐状态）

假设晴天和雨天的初始概率分别为 p 和 $1-p$。如果前一天是晴天，则第二天晴天和雨天的概率仍然是 p 和 $1-p$。如果前一天是雨天，则第二天晴天和雨天的概率分别为 q 和 $1-q$。

将上述天气变化问题抽象成马尔科夫链，记天气序列为以 $\pi = (\pi_1, \pi_2) = (p, 1-p)$ 为初始状态概率的随机过程 ，其状态空间为 $S = \{1, 2\}$（1 表示晴天，2 表示雨天）。其状态转移矩阵为 P，则有：

$$
\begin{aligned}
P &= \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \\
&= \begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix}
\end{aligned} \tag{1}
$$

为了随机生成前 N 天的天气序列 W，我们可以根据前一天的天气情况再乘以相应的转移概率得到后一天的天气情况。例如：第一天生成晴天的概率为 p，如果第一天生成的天气为 1（晴天），那么第二天生成晴天的概率为 p；如果第二天生成的天气为 2（雨天），那么第二天生成晴天的概率为 q……

实验结果

取 $p = 0.6, q = 0.3, N = 20$，执行 `make 1` 进行实验，得到的结果如下：

```
N = 20
W = [2 1 2 2 1 1 2 2 2 1 1 2 1 1 1 1 1 1 1 2]
```

马尔科夫链的生成（显状态）

某人每天根据天气按以下概率决定当天的活动：

```
emission_probability = {
    'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
    'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5}
}
```

用 1 表示散步，2 表示购物，3表示清理，根据前面生成的 W，我们根据 $W[i]$ 的值，可以得到这个人的活动 $activity[i]$ 为：

$$p(activity[i]=j)=\begin{cases}\begin{cases}0.6, & j=1\\0.3, & j=2\\0.1, & j=3\end{cases}, & W[i]=1\\\\\begin{cases}0.1, & j=1\\0.4, & j=2\\0.5, & j=3\end{cases}, & W[i]=2\end{cases} \tag{2}$$

实验结果

执行 `make 2` 进行实验，得到的结果如下：

```
N = 20
activity = [3 2 1 1 1 1 3 1 1 1 1 2 2 3 1 3 2 2 2 3]
```

隐马尔科夫模型

隐马尔科夫模型的三个问题：

- 概率计算

  计算特定观测序列的概率 - forward/backward 算法

- 预测问题

  给定模型和观测序列，求给定观测序列条件下，最可能出现的对应的状态序列 - viterbi 解码算法，基于动态规划算法

- 学习问题

  给定观测序列，估计模型的参数，使得在该模型下观测序列的条件概率最大 - baum-welch 算法

假设在 N 天内，此人发布的活动状态分别是 $\mathrm{A}=(A_1,A_2,...,A_N)$，推测这 N 天的天气 $\mathrm{T}=(T_1,T_2,...,T_N)$。

求解最可能的隐状态序列是 HMM 的三个典型问题之一，通常用维特比（viterbi）算法解决。维特比算法就是求解 HMM 上的最短路径（-log(prob)，也即是最大概率）的算法。

举例以理解 viterbi 算法的思想

现在假设此人第一天去散步、第二天清理了、第三天购物了（$\mathrm{A}=(1,3,2)$），我们按以下方式估计天气情况：

第 1 步：

$$p(\text{第一天是晴天}|\text{第一天散步})=\frac{p(\text{第一天晴天},\text{第一天散步})}{p(\text{第一天散步})} \tag{3}$$

用数学公式可以表达为：

$$
\begin{aligned}
p(W_1=1|A_1=1) &= \frac{p(W_1=1,A_1=1)}{p(A_1=1)} \\
&= \frac{p(W_1=1)p(A_1=1|W_1=1)}{p(A_1=1)} \\
&\xrightarrow{p=0.6} \frac{0.6\times 0.6}{p(A_1=1)} \\
&= \frac{0.36}{p(A_1=1)}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
p(W_1=2|A_1=1) &= \frac{p(W_1=2)p(A_1=1|W_1=2)}{p(A_1=1)} \\
&= \frac{0.4\times 0.1}{p(A_1=1)} \\
&= \frac{0.04}{p(A_1=1)}
\end{aligned} \tag{5}
$$

我们知道 $p(A_1=1)=0.4$，验算 $p(W_1=1|A_1=1)+p(W_1=2|A_1=1)=1$ 符合全概率公式。

因为 $p(W_1=1|A_1=1)$ 更大，也就是说在朋友第一天去散步时，第一天天气是晴天的概率更大，因此我们得到推测值 $T_1=1$。

第 2 步：

利用动态规划的思想，在已知 $T_1=1$（第一天晴天）和此人第二天清理（$A_2=3$）的情况下，我们再来看看第二天天气的概率情况：

$$
\begin{aligned}
p(W_2=1|W_1=1,A_2=3) &= \frac{p(W_2=1,W_1=1,A_2=3)}{p(W_1=1,A_2=3)} \\
&= \frac{p(W_1=1)p(W_2=1|W_1=1)p(A_2=3|W_1=1,W_2=1)}{p(W_1=1,A_2=3)} \\
&= \frac{0.6\times P_{11}\times 0.1}{p(W_1=1,A_2=3)} \\
&= \frac{0.6\times 0.6\times 0.1}{p(W_1=1,A_2=3)} \\
&= \frac{0.036}{p(W_1=1,A_2=3)}
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
p(W_2=2|W_1=1,A_2=3) &= \frac{p(W_1=1)p(W_2=2|W_1=1)p(A_2=3|W_1=1,W_2=2)}{p(W_1=1,A_2=3)} \\
&= \frac{0.6\times 0.4\times 0.5}{p(W_1=1,A_2=3)} \\
&= \frac{0.12}{p(W_1=1,A_2=3)}
\end{aligned} \tag{7}
$$

同样我们知道 $p(W_1=1,A_2=3)=0.6\times(0.6\times 0.1+0.4\times 0.5)=0.156$，验算 $p(W_2=1|W_1=1,A_2=3)+p(W_2=2|W_1=1,A_2=3)=1$ 符合全概率公式。

因为 $p(W_2=2|W_1=1,A_2=3)$ 更大，也就是说在第一天天晴和朋友第二天清理的情况下，第二天是雨天的概率更大，因此我们得到推测值 $T_2=2$。

第 3 步：

在已知 $T_1=1,T_2=2$ 和 $A_3=2$ 的情况下，我们再来计算第三天的天气概率情况：

$$
\begin{aligned}
p(W_3=1|W_1=1,W_2=2,A_3=2) &= \frac{p(W_1=1,W_2=2,W_3=1,A_3=2)}{p(W_1=1,W_2=2,A_3=2)} \\
&= \frac{p(W_1=1)p(W_2=2|W_1=1)p(W_3=1|W_2=2)p(A_3=2|W_3=1)}{p(W_1=1,W_2=2,A_3=2)} \\
&= \frac{0.6\times 0.4\times 0.3\times 0.3}{p(W_1=1,W_2=2,A_3=2)} \\
&= \frac{0.0216}{p(W_1=1,W_2=2,A_3=2)}
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
p(W_3=2|W_1=1,W_2=2,A_3=2) &= \frac{p(W_1=1,W_2=2,W_3=2,A_3=2)}{p(W_1=1,W_2=2,A_3=2)} \\
&= \frac{p(W_1=1)p(W_2=2|W_1=1)p(W_3=2|W_2=2)p(A_3=2|W_3=2)}{p(W_1=1,W_2=2,A_3=2)} \\
&= \frac{0.6\times 0.4\times 0.7\times 0.4}{p(W_1=1,W_2=2,A_3=2)} \\
&= \frac{0.0672}{p(W_1=1,W_2=2,A_3=2)}
\end{aligned} \tag{9}
$$

我们知道 $p(W_1=1,W_2=2,A_3=2)=0.6\times 0.4\times(0.3\times 0.3+0.7\times 0.4)=0.0888$，验算 $p(W_3=1|W_1=1,W_2=2,A_3=2)+(W_3=2|W_1=1,W_2=2,A_3=2)=1$ 符合全概率公式。

因为 $p(W_2=2|W_1=1,A_2=3)$ 更大，也就是说在第一天天晴和朋友第二天清理的情况下，第二天是雨天的概率更大，因此我们得到推测值 $T_2=2$。

... 第 i 步 ...$(i\in[4,N])$

viterbi 算法的实现

我们在实际计算时，可以不用计算 $(4)(5)(6)(7)$ 中的分母，因为做比较时分母都是相同的。记分子为 $p_1=p(W_1=T_1,...,W_n=1,A_n)$。

算法的伪代码如下：

$for n\in[1,N]:$
  $p_1=p(W_1=T_1|S)p(W_2=T_2|W_1=T_1)...p(W_n=1|W_{n-1}=T_{n-1})p(A_n|W_n=1)$
  $p_2=p(W_1=T_1|S)p(W_2=T_2|W_1=T_1)...p(W_n=2|W_{n-1}=T_{n-1})p(A_n|W_n=2)$
  $if p_1\geq p_2:$
    $T_n=1$
  $else:$
    $T_n=2$

实验结果

前面的例子

前面的例子计算结果如下：

```
N = 3
n = 0, p_1 = 0.36, p_2 = 0.04000000000000001, T[n] = 1
n = 1, p_1 = 0.036, p_2 = 0.12, T[n] = 2
n = 2, p_1 = 0.021599999999999998, p_2 = 0.0672, T[n] = 2
T = [1 2 2]
```

与前文的手算结果完全吻合。

问题 1

假设他连续三天发布的活动状态分别是$(1\ 2\ 3)$，请计算这三天天气序列为$(1\ 2\ 2)$的概率。

计算结果如下：

```
N = 3
n = 0, p_1 = 0.36, p_2 = 0.04000000000000001, T[n] = 1
n = 1, p_1 = 0.108, p_2 = 0.096, T[n] = 1
n = 2, p_1 = 0.0216, p_2 = 0.072, T[n] = 2
T = [1 1 2]
```

也就是说天气序列 $\mathrm{W}=(1,1,2)$ 的概率为 0.072，比其他天气序列的概率都大，因此预测天气序列就是 $\mathrm{T}=(1,1,2)$。

但是如何计算天气序列为$(1\ 2\ 2)$的概率呢？下面分为 3 步进行手算（与上文的例子相同）：

$$
\begin{aligned}
p_1 &= p(W_1=1|A_1=1) \\
&= \frac{p(W_1=1,A_1=1)}{p(A_1=1)} \\
&= \frac{p(W_1=1)p(A_1=1|W_1=1)}{p(A_1=1)} \\
&= \frac{0.6\times 0.6}{p(A_1=1)} \\
&= \frac{0.36}{p(A_1=1)}
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
p_2 &= p(W_2=2|W_1=1,A_2=2) \\
&= \frac{p(W_1=1,W_2=2,A_2=2)}{p(W_1=1,A_2=2)} \\
&= \frac{p(W_1=1)p(W_2=2|W_1=1)p(A_2=2|W_2=2)}{p(W_1=1,A_2=2)} \\
&= \frac{p(W_1=1)p(W_2=2|W_1=1)p(A_2=2|W_2=2)}{p(W_1=1,A_2=2)} \\
&= \frac{0.6\times 0.4\times 0.4}{p(W_1=1,A_2=2)} \\
&= \frac{0.096}{p(W_1=1,A_2=2)}
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
p_3 &= p(W_3=2|W_1=1,W_2=2,A_3=3) \\
&= \frac{p(W_1=1,W_2=2,W_3=2,A_3=3)}{p(W_1=1,W_2=2,A_3=3)} \\
&= \frac{p(W_1=1)p(W_2=2|W_1=1)p(W_3=2|W_2=2)p(A_3=3|W_3=2)}{p(W_1=1,W_2=2,A_3=3)} \\
&= \frac{0.6\times 0.4\times 0.7\times 0.5}{p(W_1=1,W_2=2,A_3=3)} \\
&= \frac{0.084}{p(W_1=1,W_2=2,A_3=3)}
\end{aligned} \tag{12}
$$

如果只看分子我们可以得到其概率为 0.084 可用于与其他长度为 3 的序列的概率直接进行比较。当然，我们也可以再作进一步计算算出分母（实际求解时不必求出）：

$$
p(W_1=1,W_2=2,A_3=3)=0.6\times 0.4\times(0.3\times 0.1+0.7\times 0.5)=0.0912 \tag{13}
$$

真实的结果为：

$$p_3 = \frac{0.084}{0.0912} = \frac{7}{76} \tag{14}$$

同理：

$$p(W_1 = 1, A_2 = 2) = 0.6 \times (0.6 \times 0.3 + 0.4 \times 0.4) = 0.204 \tag{15}$$

$$p(A_1) = 0.6 \times 0.6 + 0.4 \times 0.1 = 0.4 \tag{16}$$

从而

$$p_2 = \frac{0.096}{0.204} = \frac{8}{17} \tag{17}$$

$$p_1 = \frac{0.36}{0.4} = \frac{9}{10} \tag{18}$$

我们要求的概率为在 $A_1 = 1$ 情况下出现 $(1, x, x)$ 的概率乘以在 $(1, x, x)$ 和 $A_2 = 2$下出现 $(1, 2, x)$ 的概率乘以在 $A_3 = 3$ 的情况下出现 $(1, 2, 2)$的概率：

$$\begin{aligned} & p(W_1 = 1, W_2 = 2, W_3 = 2 | A_1 = 1, A_2 = 2, A_3 = 3) \\ =& \frac{p(W_1 = 1, A_1 = 1)}{p(A_1 = 1)} \frac{p(W_1 = 1, W_2 = 2, A_2 = 2)}{p(W_1 = 1, A_2 = 2)} \frac{p(W_1 = 1, W_2 = 2, W_3 = 2, A_3 = 3)}{p(W_1 = 1, W_2 = 2, A_3 = 3)} \\ =& p_1 p_2 p_3 \\ =& \frac{63}{1615} = 0.03901 \end{aligned} \tag{19}$$

问题 2

假设他连续二十天发布的状态是$(2\ 1\ 3\ 2\ 3\ 2\ 2\ 3\ 3\ 1\ 2\ 1\ 1\ 1\ 2\ 3\ 3\ 3\ 3\ 2)$，请推测这 20 天的天气。

计算结果如下：

```
N = 20
n = 0, p_1 = 0.18, p_2 = 0.16000000000000003, T[n] = 1
n = 1, p_1 = 0.216, p_2 = 0.024, T[n] = 1
n = 2, p_1 = 0.0216, p_2 = 0.072, T[n] = 2
n = 3, p_1 = 0.012959999999999998, p_2 = 0.040319999999999995, T[n] = 2
n = 4, p_1 = 0.0030239999999999998, p_2 = 0.03527999999999999, T[n] = 2
n = 5, p_1 = 0.006350399999999998, p_2 = 0.019756799999999995, T[n] = 2
n = 6, p_1 = 0.0044452799999999985, p_2 = 0.013829759999999995, T[n] = 2
n = 7, p_1 = 0.0010372319999999997, p_2 = 0.012101039999999993, T[n] = 2
n = 8, p_1 = 0.0007260623999999997, p_2 = 0.008470727999999995, T[n] = 2
n = 9, p_1 = 0.003049462079999998, p_2 = 0.0011859019199999994, T[n] = 1
n = 10, p_1 = 0.0009148386239999994, p_2 = 0.0008131898879999996, T[n] = 1
n = 11, p_1 = 0.0010978063487999993, p_2 = 0.0001219784831999994, T[n] = 1
n = 12, p_1 = 0.0006586838092799995, p_2 = 7.318708991999996e-05, T[n] = 1
n = 13, p_1 = 0.0003952102855679997, p_2 = 4.391225395199998e-05, T[n] = 1
n = 14, p_1 = 0.0001185630856703999, p_2 = 0.0001053894094879994, T[n] = 1
n = 15, p_1 = 2.3712617134079983e-05, p_2 = 7.904205711359995e-05, T[n] = 2
n = 16, p_1 = 4.742523426815997e-06, p_2 = 5.532943997951996e-05, T[n] = 2
n = 17, p_1 = 3.3197663987711975e-06, p_2 = 3.873060798566397e-05, T[n] = 2
n = 18, p_1 = 2.323836479139838e-06, p_2 = 2.7111425589964774e-05, T[n] = 2
n = 19, p_1 = 4.880056606193659e-06, p_2 = 1.5182398330380275e-05, T[n] = 2
T = [1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 2]
```

问题 3

按照前面生成的活动序列，来推测天气序列，并验证是否与问题一中生成的天气序列相同。

生成的天气序列：

```
N = 20
W = [2 1 2 2 1 1 2 2 2 1 1 2 1 1 1 1 1 1 1 2]
```

生成的活动序列：

```
N = 20
activity = [3 2 1 1 1 1 3 1 1 1 1 2 2 3 1 3 2 2 2 3]
```

推测的天气序列：

```
N = 20
n = 0, p_1 = 0.06, p_2 = 0.2
n = 1, p_1 = 0.072, p_2 = 0.027999999999999997
n = 2, p_1 = 0.0072, p_2 = 0.024
n = 3, p_1 = 0.00432, p_2 = 0.01344
n = 4, p_1 = 0.0030239999999999993, p_2 = 0.009408
n = 5, p_1 = 0.0007055999999999998, p_2 = 0.008231999999999998
n = 6, p_1 = 0.0004939199999999999, p_2 = 0.005762399999999998
n = 7, p_1 = 0.0010372319999999995, p_2 = 0.003226943999999999
n = 8, p_1 = 0.00024202079999999994, p_2 = 0.002823575999999999
n = 9, p_1 = 0.0010164873599999996, p_2 = 0.0003953006399999999
n = 10, p_1 = 0.0006098924159999998, p_2 = 6.776582399999997e-05
n = 11, p_1 = 6.098924159999998e-05, p_2 = 0.00020329747199999992
n = 12, p_1 = 7.318708991999996e-05, p_2 = 2.8461646079999986e-05
n = 13, p_1 = 4.3912253951999975e-05, p_2 = 4.879139327999998e-06
n = 14, p_1 = 1.3173676185599992e-05, p_2 = 1.1709934387199994e-05
n = 15, p_1 = 2.6347352371199985e-06, p_2 = 8.782450790399995e-06
n = 16, p_1 = 1.5808411422719991e-06, p_2 = 4.918172442623998e-06
n = 17, p_1 = 1.1065887995903992e-06, p_2 = 3.442720709836798e-06
n = 18, p_1 = 2.582040532377598e-07, p_2 = 3.0123806211071976e-06
n = 19, p_1 = 1.084457023598591e-06, p_2 = 4.217332869550077e-07
T = [2 1 2 2 2 2 2 2 2 1 1 2 1 1 1 2 2 2 2 1]
```