

25 | 马尔科夫模型：从PageRank到语音识别，背后是什么模型在支撑？

2019-02-11 黄申

程序员的数学基础课

[进入课程 >](#)



讲述：黄申

时长 14:28 大小 13.26M



你好，我是黄申。

上一节，我们介绍了基于概率的语言模型。概率语言模型的研究对象其实是一个词的序列，以及这个词序列出现的概率有多大。那语言模型是不是也可以用于估算其他序列出现的概率呢？答案是肯定的。

通过上一节我们知道，语言模型中有个重点：马尔科夫假设及对应的多元文法模型。如果我们把这一点进一步泛化，就能引出马尔科夫模型。也就是说，只要序列的每个状态之间存在转移的概率，那么我们就可以使用马尔科夫模型。有时候情况会更复杂，不仅每个状态之间的转移是按照一定概率进行的，就连每个状态本身也是按照一定概率分布出现的，那么还需要用到隐马尔科夫模型。

今天这一节，我们就来学习马尔科夫模型、隐马尔科夫模型，以及它们在 PageRank 和中文分词中的应用。

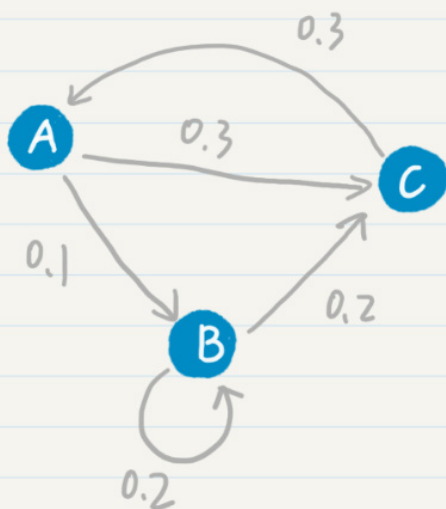
马尔科夫模型

在介绍语言模型的时候，我们提到了马尔科夫假设，这个假设是说，每个词出现的概率和之前的一个或若干个词有关。我们换个角度思考就是，**每个词按照一定的概率转移到下一个词**。怎么个转移呢？我来解释一下。

如果把词抽象为一个状态，那么我们就可以认为，状态到状态之间是有关联的。前一个状态有一定的概率可以转移到下一个状态。如果多个状态之间的随机转移满足马尔科夫假设，那么这类随机过程就是一个马尔科夫随机过程。而刻画这类随机过程的统计模型，就是**马尔科夫模型** (Markov Model) 。

前面讲多元文法的时候，我提到了二元文法、三元文法。对于二元文法来说，某个词出现的概率只和前一个词有关。对应的，在马尔科夫模型中，如果一个状态出现的概率只和前一个状态有关，那么我们称它为**一阶马尔科夫模型**或者**马尔科夫链**。对应于三元、四元甚至更多元的文法，我们也有二阶、三阶等马尔科夫模型。

我们先从最简单的**马尔科夫模型 - 马尔科夫链**开始看。我画了一张示意图，方便你理解马尔科夫链中各个状态的转移过程。



在这张图中，你可以看到，从状态 A 到 B 的概率是 0.1，从状态 B 到状态 C 的概率是 0.2 等等。我们也可以使用状态转移表来表示这张图。

上个状态\下个状态	A	B	C
A	0	0.1	0.3
B	0	0.2	0.2
C	0.3	0	0

我们可以根据某个应用的需要，把上述状态转移表具体化。例如，对于语言模型中的二元文法模型，我这里列出了一个示意表。

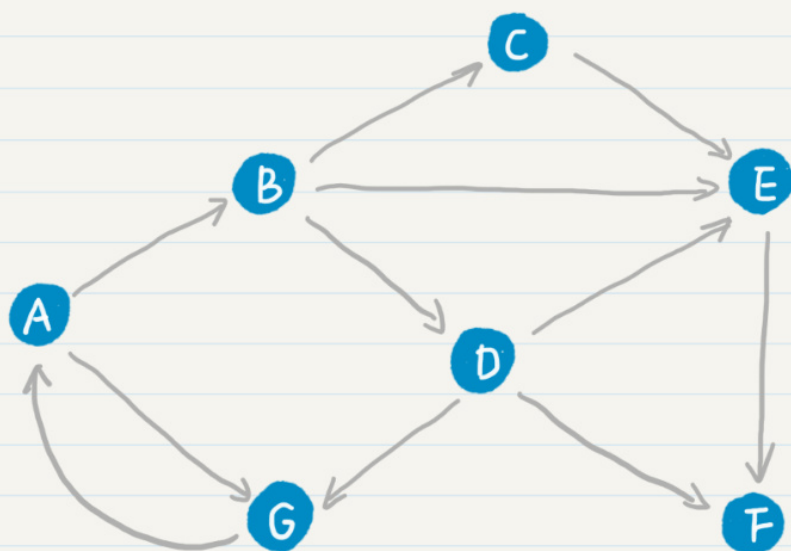
上个词\下个词	我	去	学校
我	0	0.5	0.2
去	0.3	0.05	0.8
学校	0	0.1	0

当然，除了二元文法模型，马尔科夫链还有很多应用的场景。

Google 公司最引以为傲的 PageRank 链接分析算法，它的核心思想就是基于马尔科夫链。这个算法假设了一个“随机冲浪者”模型，冲浪者从某张网页出发，根据 Web 图中的链接关系随机访问。在每个步骤中，冲浪者都会从当前网页的链出网页中随机选取一张作为下一步访问的目标。在整个 Web 图中，绝大部分网页节点都会有链入和链出。那么冲浪者就可以永不停歇地冲浪，持续在图中走下去。

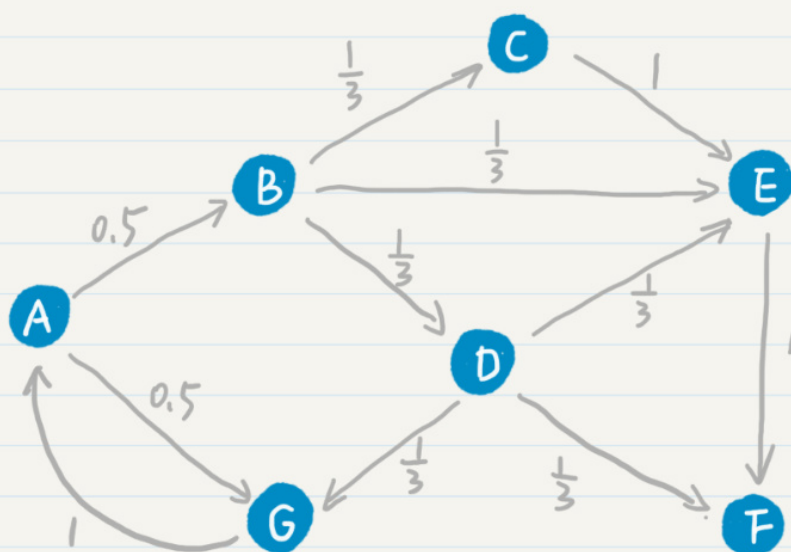
在随机访问的过程中，越是被频繁访问的链接，越是重要。可以看出，每个节点的 PageRank 值取决于 Web 图的链接结构。假如一个页面节点有很多的链入链接，或者是链入的网页有较高的被访问率，那么它也将会有更高的被访问概率。

那么，PageRank 的公式和马尔科夫链有什么关系呢？我先给你看一张 Web 的拓扑图。



其中 A、B、C 等结点分别代表了页面，而结点之间的有向边代表了页面之间的超链接。看了这张图中，你是不是觉得 Web 拓扑图和马尔科夫链的模型图基本上是一致的？我们可以假设每张网页就是一个状态，而网页之间的链接表明了状态转移的方向。这样，我们很自然地就可以使用马尔科夫链来刻画“随机冲浪者”。

另外，在最基本的 PageRank 算法中，我们可以假设每张网页的出度是 n ，那么从这张网页转移到任何下一张相连网页的概率都是 $\frac{1}{n}$ ，因此这个转移的概率只和当前页面有关，满足一阶马尔科夫模型的假设。我在之前的拓扑结构中添加了转移的概率。



PageRank 在标准的马尔科夫链上，引入了随机的跳转操作，也就是假设冲浪者不按照 Web 图的拓扑结构走下去，只是随机挑选了一张网页进行跳转。这样的处理是类比人们打开一张新网页的行为，也是符合实际情况的，避免了信息孤岛的形成。最终，根据马尔科夫链的状态转移和随机跳转，可以得到如下的 PageRank 公式。


$$PR(p_i) = \alpha \sum_{p_j \in M_i} \frac{PR(p_j)}{L(p_j)} + \frac{(1-\alpha)}{N}$$

其中， p_i 表示第 i 张网页， M_i 是 p_i 的入链接集合， p_j 是 M_i 集合中的第 j 张网页。 $PR(p_j)$ 表示网页 p_j 的 PageRank 得分， $L(p_j)$ 表示网页 p_j 的出链接数量， $\frac{1}{L(p_j)}$ 就表示从网页 p_j 跳转到 p_i 的概率。 α 是用户不进行随机跳转的概率， N 表示所有网页的数量。

从最简单的马尔科夫链，到多阶的马尔科夫模型，它们都可以刻画基于马尔科夫假设的随机过程，例如概率语言模型中的多元文法和 PageRank 这类链接分析算法。但是，这些模型都是假设每个状态对我们都是已知的，比如在概率语言模型中，一个状态对应了单词“上学”，另一个状态对应了单词“书包”。可是，有没有可能某些状态我们是未知的呢？下面我们就来详细说说这种情况。

隐马尔科夫模型

在某些现实的应用场景中，我们是无法确定马尔科夫过程中某个状态的取值的。这种情况下，最经典的案例就是语音识别。使用概率对语音进行识别的过程，和语言模型类似，因此我们可以把每个等待识别的词对应为马尔科夫过程中的一个状态。不过，语音识别所面临的困难更大。为什么呢？你先看看下面这个句子。这个句子里全都是拼音，你能看出它表示什么意思吗？

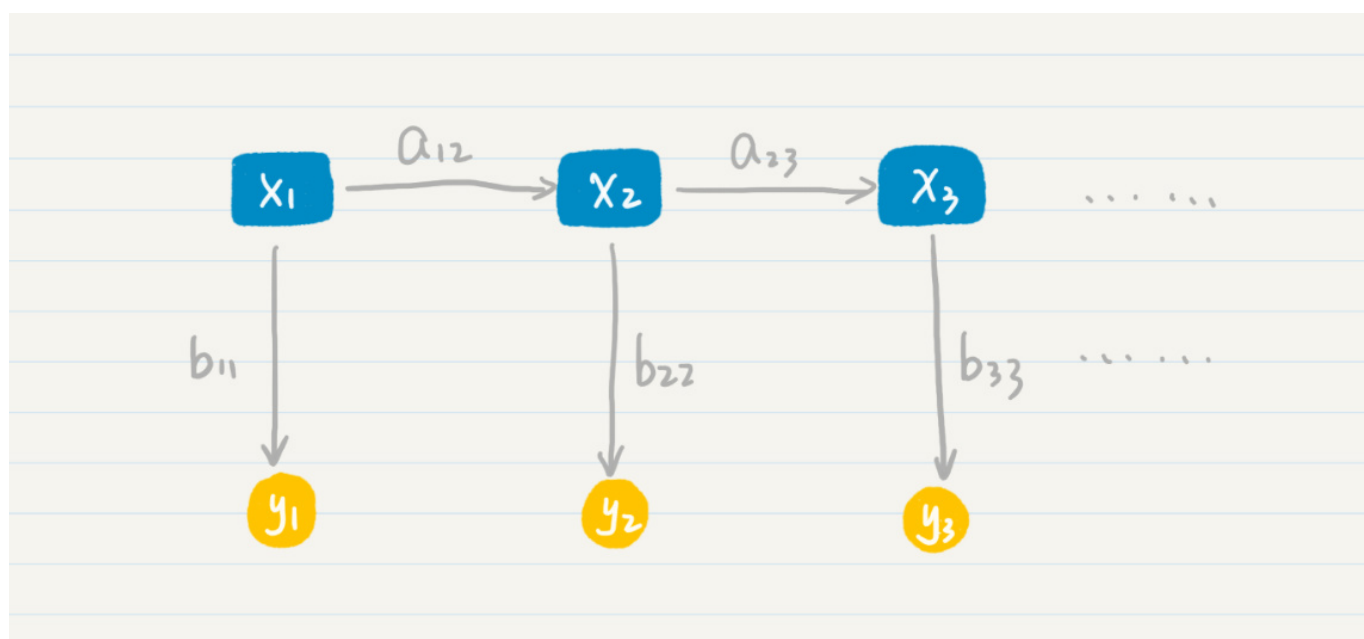
 复制代码

```
1 ni(三声) zhi(一声) dao(四声) wo(三声) zai(四声) deng(三声) ni(三声) ma(一声)
```

中国有句古话说得好，“白纸黑字”，写在文档里的文字对于计算机是确定的，“嘛”“吗”“妈”不会弄错。可是，如果你说一句“你知道我在等你吗”，听众可

能一直弄不明白为什么要等别人的妈妈，除非你给他们看到文字版的内容，证明最后一个字是口字旁的“吗”。另外，再加上各种地方的口音、唱歌的发音或者不标准的拼读，情况就更糟糕了。

计算机只知道某个词的发音，而不知道它具体怎么写，对于这种情况，我们就认为计算机只能观测到每个状态的部分信息，而另外一些信息被“隐藏”了起来。这个时候，我们就需要用隐马尔科夫模型来解决这种问题。隐马尔科夫模型有两层，一层是我们可以观测到的数据，称为“输出层”，另一层则是我们无法直接观测到的状态，称为“隐藏状态层”。我画了一张图方便你理解。




其中， x_1 ， x_2 ， x_3 等等属于隐藏状态层， a_{12} 表示了从状态 x_1 到 x_2 的转移概率， a_{23} 表示了从状态 x_2 到 x_3 的转移概率。这一层和普通的马尔科夫模型是一致的，可惜在隐马尔科夫模型中我们无法通过数据直接观测到这一层。我们所能看到的是， y_1 ， y_2 ， y_3 等等代表的“输出层”。另外， b_{11} 表示了从状态 x_1 到 y_1 的输出概率， b_{22} 表示了从状态 x_2 到 y_2 的输出概率， b_{33} 表示了从状态 x_3 到 y_3 的输出概率等等。

那么在这个两层模型示例中，“隐藏状态层”产生“输出层”的概率是多少呢？这是一系列条件概率决定的，具体的公式我列在这里。

$$\begin{aligned} &P(x_1) \times P(y_1 | x_1) \times P(x_2 | x_1) \times P(y_2 | x_2) \times P(x_3 | x_2) \times P(y_3 | x_3) \\ &= P(x_1) \times b_{11} \times a_{12} \times b_{22} \times a_{23} \times b_{33} \end{aligned}$$

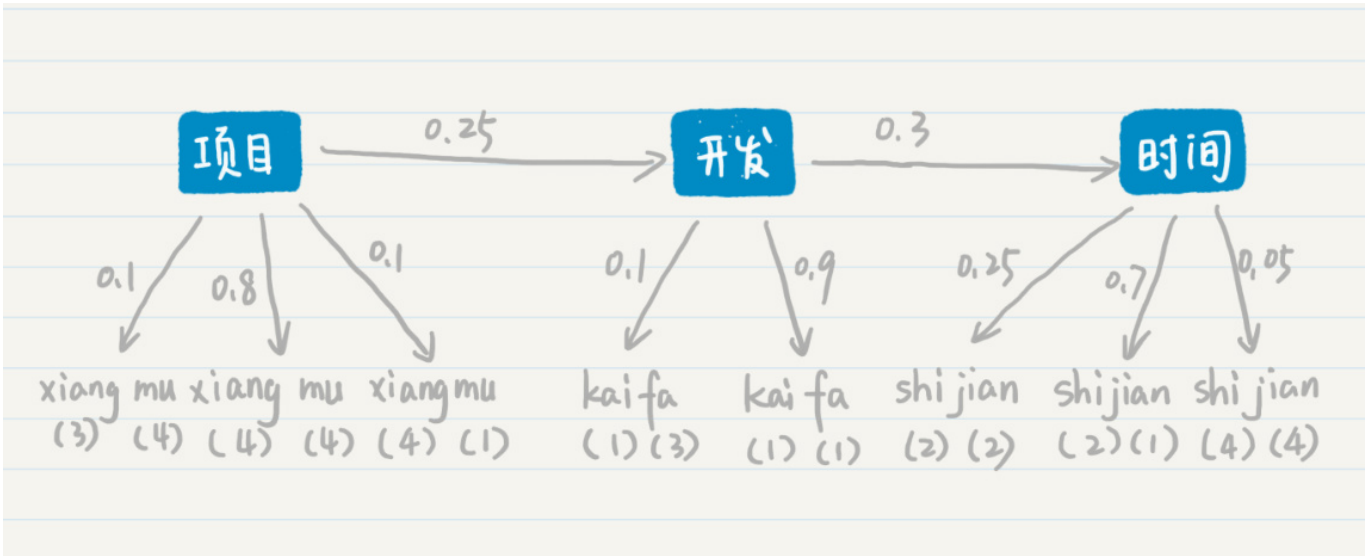
如果你觉得这个两层的模型不太好理解，我来给你说个浅显易懂的例子。假设正在进行普通话语音识别，计算机接受了一个词组的发音。我在下面列出了它的拼音。

 复制代码

```
1 xiang(四声)mu(四声) kai(一声)fa(一声) shi(四声)jian(四声)
```

假设根据我们手头上的语料数据，这个词组有多种可能，我列出两种。

第一种情况



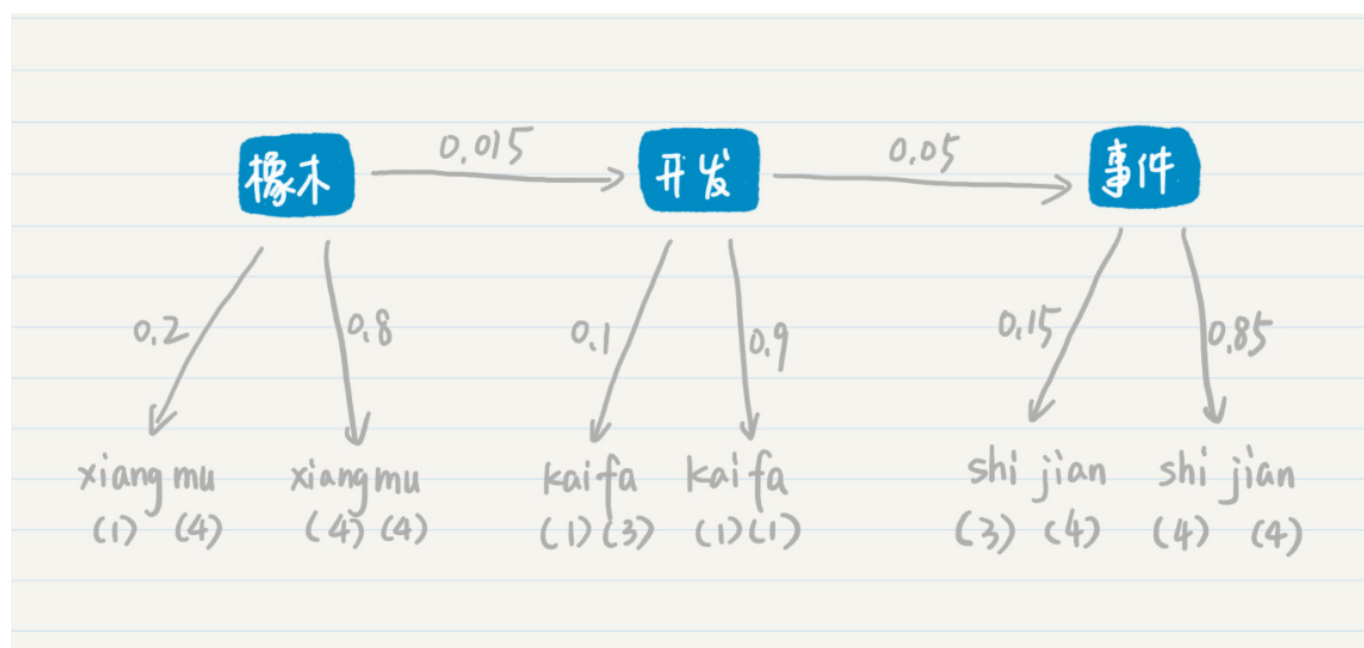
第一种情况下，三个确定的状态是“项目”“开发”和“时间”这三个词。从“项目”转移到“开发”的概率是 0.25，从“开发”转移到“时间”的概率是 0.3。从“项目”输出“xiang (三声) mu (四声)”的概率是 0.1，输出“xiang (四声) mu (四声)”的概率是 0.8，输出“xiang (四声) mu (一声)”的概率是 0.1，“开发”和“时间”也有类似的输出概率。

这个时候你可能会奇怪，“项目”的普通话发音就是“xiang (四声) mu (四声)”，为什么还会输出其他的发音呢？这是因为，前面说的这些概率都是通过历史语料的数据统计而来。在进行语音识别的时候，我们会通过不同地区、不同性别、不同年龄等等的人群，采集发音的样本。如此一来，影响这个发音的因素就很多了，比如方言，口音、误读等等。当然，在正常情况下，大部分的发音还是标准的，所以“项目”这个词输出到“xiang (四声) mu (四声)”的概率是最高的。

好，有了这些概率的分布，我们来看看“项目开发时间”这个词组最后生成的概率是多少。在两层模型的条件概率公式中，我代入了具体的概率值并使用了如下的推导：

$$\begin{aligned} & P(\text{项目}) \times P(\text{xiang}(4)\text{mu}(4) | \text{项目}) \times P(\text{开发} | \text{项目}) \times P(\text{kai}(1)\text{fa}(1) | \text{开发}) \times P(\text{时间} | \text{开发}) \times P(\text{shi}(4)\text{jian}(4) | \text{时间}) \\ &= P(\text{项目}) \times 0.8 \times 0.25 \times 0.9 \times 0.3 \times 0.05 \\ &= P(\text{项目}) \times 0.0027 \end{aligned}$$

第二种情况



在第二种的可能性中，三个确定的状态是“橡木”“开发”和“事件”这三个词。从“橡木”转移到“开发”的概率是 0.015，从“开发”转移到“事件”的概率是 0.05。从“橡木”输出“xiang（一声）mu（四声）”的概率是 0.2，输出“xiang（四声）mu（四声）”的概率是 0.8，“开发”和“事件”也有类似的输出概率。和第一种情况类似，我们可以计算“橡木开发事件”这个词组最后生成的概率是多少，我用下面这个公式来推导：

$$\begin{aligned} & P(\text{橡木}) \times P(\text{xiang}(4)\text{mu}(4) | \text{橡木}) \times P(\text{开发} | \text{橡木}) \times P(\text{kai}(1)\text{fa}(1) | \text{开发}) \times P(\text{事件} | \text{开发}) \times P(\text{shi}(4)\text{jian}(4) | \text{事件}) \\ &= P(\text{橡木}) \times 0.8 \times 0.015 \times 0.9 \times 0.05 \times 0.85 \\ &= P(\text{橡木}) \times 0.000459 \end{aligned}$$

最后比较第一种和第二种情况产生的概率，分别是 $P(\text{项目}) \times 0.0027$ 和 $P(\text{橡木}) \times 0.000459$ 。假设 $P(\text{项目})$ 和 $P(\text{橡木})$ 相等，那么“项目开发时间”这个词组的概率更高。所以“xiang（四声）mu（四声）kai（一声）fa（一声）shi（四声）jian（四声）”这组发音，计算机识别为“项目开发时间”。从中我们可以看出，尽管“事件”这个词产生“shi（四声）jian（四声）”这个发音的可能性更高，但是“橡木开发事件”这个词组

出现的概率极低，因此最终计算机还是选择了“项目开发时间”，隐藏的状态层起到了关键的作用。

总结

马尔科夫模型考虑了 n 个状态之间的转移及其对应的关系。这个状态是比较抽象的含义，在不同的应用领域代表不同的含义。在概率语言模型中，状态表示不同的词，状态之间的转移就代表了词按照一定的先后顺序出现。在 PageRank 这种链接分析中，状态表示不同的网页，状态之间的转移就代表了人们在不同网页之间的跳转。

在马尔科夫模型中，我们知道了每种状态及其之间转移的概率，然后求解序列出现的概率。然而，有些现实的场景更为复杂，比如说我们观测到的不是状态本身，而是状态按照一定概率分布所产生的输出。针对这种情况，隐马尔科夫模型提出了一种两层的模型，同时考虑了状态之间转移的概率和状态产生输出的概率，为语音识别、手写识别、机器翻译等提供了可行的解决方案。

隐马尔科夫模型需要回答的最主要问题是：给定一个模型和某个特定的输出序列，如何找到最可能产生这个输出的状态序列？在本节中，我使用了“项目开发时间”这个例子展示隐马尔科夫模型是如何工作的。不过这个例子很简单，我只比较了两种可能性。但是，实际中可能性是非常多的，如果我们使用穷举法，那么复杂度一定很高。

我们可以把两层的模型看作图结构。其中，状态和输出是结点，转移和输出关系是边，相应的概率是边的权重，这个时候我们就可以对 Dijkstra 算法稍加修改，来找出权重乘积最大的最优路径，提升查找的效率。我们还可以利用状态序列之间存在的先后关系，使用基于动态规划的维特比（Viterbi）算法来找出最优路径。

思考题

机器翻译会使用大量的语料，自动学习不同语言之间词和词的匹配。如果在机器翻译中使用隐马尔科夫进行建模，你认为“隐藏状态层”表示的是什么？“输出层”表示的又是什么？

欢迎留言和我分享，也欢迎你在留言区写下今天的学习笔记。你可以点击“请朋友读”，把今天的内容分享给你的好友，和他一起精进。

程序员的数学基础课

在实战中重新理解数学

黄申

LinkedIn 资深数据科学家



新版升级：点击「👤 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 24 | 语言模型：如何使用链式法则和马尔科夫假设简化概率模型？

下一篇 26 | 信息熵：如何通过几个问题，测出你对应的武侠人物？

精选留言 (8)

写留言



qinggeouy...

2019-03-06

3

机器翻译，比如一个中文句子翻译为英文，中文句子可拆分为多个词，每个中文词可能会匹配多个英文单词，一句中文翻译为英文，可能会有多种翻译结果。那么：

隐藏状态层 -> 多种翻译结果可能性中的一种

输出层 -> 每个中文词可能匹配多个英文单词

...

展开

作者回复: 是的，理解正确



Joe

2019-02-21

👍 2

隐马尔科夫模型在语音中的应用，流程是：

- 1，根据拼音去找到单个对应的词语，不考虑声调的概率。
- 2，再根据词语之间转移的概率，词语对应目标音高的概率，进而求出整个句子输出的概率。概率越大，可能性越高。

因此第一个词可以是xiangmu 对应语料库的所有词，不一定是四声，可以是香木之类的...

展开 ∨

作者回复: 是的👍



唯她命

2019-03-05

👍 1

老师你好，“隐藏状态层”产生“输出层”的概率这句话是什么意思。
还有“隐藏状态层”产生“输出层”的概率那个公式是怎么推导出来的。

作者回复: 我看你另外一个问题，应该是已经理解了“隐藏状态层”和“输出层”的概念，简单的来理解，就是“隐藏层”到“输出层”是有一定概率分布的，不是必然的。如果必然了，就不需要区分这两层了，直接把“输出层”当做“隐藏层”就好了。

也正是因为如此，公式推导使用了条件概率公式。在“隐藏层”的某个状态给定的情况下，出现“输出层”某个输出的概率是多少，以此类推。



李皮皮皮皮...

2019-03-02

👍 1

讲马尔可夫链的第一张图中，就是ABC的那种图，链上的概率是如何得出的？

我的理解是：根据语境中出现的AB，BC这些组合的概率，比如说总共有10个二元组，AB出现了一次，所以A到B的链上概率是0.1。如果按这么理解的话，那么所有的概率之和应该等于1吧。但是我算了一下，图中概率和等于1.1。不知道我理解的对不对，望老师解答🙄

展开 ∨

作者回复: 不是这个意思。实际上每条出链接上的概率，表示给定出发点，到目标点的概率，是一种条件概率。比如，从A到B的概率是0.1，就是 $P(B|A)=0.1$ 。所以，不能把这些条件概率相加。

我这里只是示意图，A可能还会链接到其他我没有画在图中的点，从A到所有目标点的概率加起来应该是1。比如从A出发一共到三个点，B，X和Y， $P(B|A)=0.1$ ， $P(X|A)=0.5$ ， $P(Z|A)=0.4$ ，加起来是1



拉欧

2019-02-12

1

输出层：被翻译语言，隐藏状态层：翻译语言
比如要翻译 get busy living ,or get busy dying
输出层为 get busy living ,or get busy dying
隐藏层可能为：
1、要么忙于生存，要么赶着去死...

展开

作者回复: 思路是对的



□

2019-02-12

1

硬着头皮看完。比如英文翻译目标语言，我认为要翻译的文本的直接看出来的特征（单词包括的意思），单词之间的语法规则和词性，时态是隐藏层。
数学差看起来就是费劲 $\circ(\pi\sim\pi)\circ$ ，英文差看不懂最新论文。不清楚大家啥水平，我建议老师多给几个例子好理解些。深刻体会到没多一个公式，少一分看下去的兴趣。

展开

作者回复: 后面我会多一些关于公式的解释，帮助理解



Thinking

2019-02-12

1

不理解一个地方，由读音推测汉字的过程为什么要算 $P(\text{xiang}(4)\mu(4)|\text{项目})$ 概率而不是 $P(\text{项目}|\text{xiang}(4)\mu(4))$ 概率？隐马尔科夫解决由输出层找到产生输出的隐藏状态层，为什么不换个角度说成由看得到的输入层找到隐藏状态层呢？

展开

作者回复: 这个问题很好！这主要是从统计的可行性出发，以语音识别为例，我们的语料（或者说标注数据、历史数据）都是给定文本，比如说中文，然后收集用户的发音（如果是中文，发音也就对应是拼音），所以可以很自然的拿到 $P(\text{中文}|\text{拼音})$ 这种概率。



唯她命

2019-03-05



老师 语音识别的例子

隐藏层：项目开发时间 或者 橡木开发事件

输出层：xiang(四声)mu(四声) kai(一声)fa(一声) shi(四声)jian(四声)

xiang(一声)mu(-声) kai(一声)fa(一声) shi(四声)jian(四声)

xiang(四声)mu(-声) kai(一声)fa(一声) shi(二声)jian(四声)...

展开 ▾

作者回复: 对文中的例子，这个理解是对的。至于思考题，我暂时不回答，给其他读者更多的思考机会 😊

