参考文章： <http://www.cnblogs.com/zhizhan/p/4110868.html>

**Hidden Markov Models （隐马尔科夫模型）**

Intuition informs much of what we do: for example, it tells us that certain words tend

to be a certain part of speech, or that if a user visits a signup page, she has a higher

probability of becoming a customer. But how would you build a model around

intuition?

我们做很多事情时常常会依赖直觉：例如，直觉会告诉我们，某些词往往是某些特定词类的一部分，或者说，当一个用户访问一个某购物网站注册页面时，其成为最终客户的概率会较高。但是你思考过如何围绕这些直觉来建立一个具体的模型吗？

Hidden Markov models (HMMs) are well versed in finding a hidden state of a given

system using observations and an assumption about how those states work. In this

chapter, we will first talk about how to track user states given their actions, then

explore more about what an HMM is, and finally build a part-of-speech tagger using

the Brown Corpus. The part-of-speech tagger will tag words in sentences as nouns,

pronouns, or any part of speech in the Brown Corpus.

Hidden Markov模型（HMM）擅长于寻找和预测隐藏和不能被直接观测到的指定系统的状态。其原理是通过一个可以观察到的状态和与其相关的隐藏状态序列的相关概率来进行推演和预测。在这一章中，我们将首先讨论如何跟踪用户状态，进而探讨更多关于HMM的话题，并最终通过使用布朗语料库来建立一个词性标注系统。该语音标注系统会给句中名词、代词打上标签，甚至可以标注任何演讲内容。

HMMs can be either supervised or unsupervised and also are called Markovian due

to their reliance on a Markov model. They work well where there doesn’t need to be a

lot of historical information built into the model. They also work well for adding

localized context to a classification. Unlike what we saw with Naive Bayesian Classification,

which relied on a lot of history to determine whether a user is spammy or not,

HMMs can be used to predict changes over time in a model.

HMM模型支持监督或无人监督，由于其对马尔可夫的依赖因此也被称为马尔可夫模型。

该模型不需要大量的历史数据信息也能正常工作，也能很好地为在小范围内为分类添加上下文。不像我们之前所看到的依赖于大量的历史数据信息才能确定用户是否为垃圾邮件的朴素贝叶斯模型，HMM模型可以用来预测随着时间而变化模型状态。

**Tracking User Behavior Using State Machines 使用状态机来跟踪用户行为**

Have you ever heard of the sales funnel? This is the idea that there are different levels

of customer interaction. People will start as prospects and then transition into more

engaged states (see Figure 6-1).

有人听说过销售漏斗理论吗？ 该理论认为不同层次的客户与销售间存在不同的互动。

人们通常会从观望开始，然后逐步过渡并成为最终的大客户（见图6-1）。

Prospects are “lurkers” who visit the site once or twice but usually don’t engage.

Users, on the other hand, like to browse and occasionally make purchases. Customers

are quite engaged and have bought something but usually don’t buy a lot in a short

time, and thus go back to being users temporarily.

观望者访问购物网站的频率很低，通常只有一次或两次，即使有活动也通常不参与。

另一方面，如果是普通用户，则喜欢经常浏览该购物网站并偶尔购买商品。作为最积极的大客户，一旦有活动就会购买商品，但大客户通常不会在短时间内购买很多东西，因此也会暂时蜕变成普通用户。

Let’s say that we have an online store and determine that out of prospects that visit

the site, 15% will sign up, and 5% will become customers right away. When the visitor

is already a user, he will cancel his account 5% of the time and buy something 15% of

the time. If the visitor is a customer, he will cancel his account only 2% of the time

and go back to being a user 95% of the time instead of continually buying things.

假设我们有一个网上商店，并假定出访问该网站的观望者中，15%将注册成用户，其中5%的用户将立刻成为大客户。当访问者已经是一个用户时，他有5%概率会销户，并且有15%的概率会购买东西。如果访问者已经是一个大客户，他注销的帐户概率只有2%，有95%的概率蜕化成普通用户，而不是不断的购买商品。

We could represent the information we have collected in a transition matrix, which

shows the probability of going from one state to another, or remaining in the same

state (Table 6-1).

我们可以用状态转换矩阵表来描述上述行为，该表显示了一个状态转换成另一个状态的概率，或停留在原先的状态（表6-1）。

What the transition probability defines is known as a state machine (see Figure 6-2).

It also tells us a lot about how our current customers behave. We can determine the

conversion rate, attrition rate, and other probabilities. Conversion rate is the probability

of a prospect signing up, which would be 20%—the probability of going from

prospect to user plus the probability of prospect to customer (15% + 5%). You could

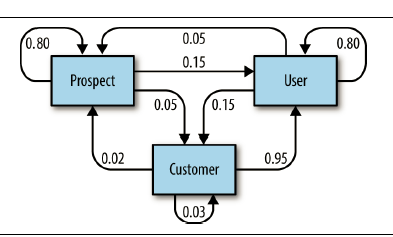
also determine the attrition rate by taking the average of 5% and 2%, which is 3.5%.

转移概率常常被称为一个状态机（见图6-2）。

它告诉了我们很多当前的客户行为。我们可以确定转换率，磨损率和其他概率。

转换率是代表一个观望者变成用户的概率，此处为20%，该概率包括了从观望者变成普通用户以及大客户两部分（15% + 5%）。

你也可以通过5%和2%的平均值来确定磨损率，此处是3.5%。



This is an uncommon way of displaying user behavior in analytics, because it is too

explanatory. But it has one advantage over traditional conversion rate calculations:

the ability to look at how a user operates over time. For instance, we could determine

the probability of a user being a prospect given the last four times he was in fact a

prospect. This is the probability of being a prospect (say 80%) multiplied by the four

times they were a prospect before, which were all 80%. The probability that someone

keeps viewing the site and never signs up is low, because eventually he might sign up.

(下面一段的翻译有点不太确定)

在分析中显示用户行为的并非常见做法，因为它比较难以解释。

但与传统的转换概率计算模型相比，其优势在于该模型拥有可以查看用户如何随着时间的推移而进行角色状态转换的能力。

例如，我们可以通过概率模型计算出过去四次的时间观测点上一个普通用户变成观望者的概率，假设他目前还只是一个观望者。

根据转换概率模型(假设每次的转换概率都是80%)随着时间的推移，在每次的时间观测点上的转换概率仍然是80%。这显然不符合实际情况，

因为某人从来没有注册过该购物网站，却在每次的时间观测点上都出现了不停的查看该购物网站的行为，这样的可能性是非常低的。

But there is also one major problem with this model: there is no way for us to reliably

determine these states without asking each user individually. The state is hidden from

our observation. A user can view the site anonymously.

但是，这个模型也有缺陷：在不要求用户登录的情况下无法准确的确定用户的级别。由于每个人无须登录都查看购物网站的所有商品目录，因此其真实用户级别是我们无法观测到的。（当前匿名浏览者可能只是一个观望者，也可能是最高级别的大客户）

That is actually fine, as you will soon see. As long as we are able to observe interaction

with the site and make a judgment call about the underlying transitions from

other sources (think Google Analytics), then we can still solve this problem.

这个缺陷是可以被克服的，正如你所将看到的那样，只要我们能够观察到用户与网站的互动，并通过调用某些分析工具（如Google Analytics），那么我们仍然可以解决这个问题并鉴别用户的实际身份

We do this by introducing another level of complexity called emissions.

我们将通过引入另外一个层次的复杂性（也叫排放）来解决上述缺陷