在线大规模开放中预测学生保留率
使用隐马尔可夫模型的课程

吉里斯·巴拉克里希南（Girish Balakrishnan）

加州大学伯克利分校电气工程与计算机科学系

技术报告编号UCB / EECS-2013-109

http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.html

2013年5月17日

致谢

我要感谢Derrick Coetzee，他是一位非常发人深省的研究合作伙伴，并为我撰写这份报告提供了指导。我还要感谢Armando Fox教授在我的硕士课程期间提供的指导，以及John Canny教授在讨论和完善我们的研究方法方面的所有帮助。最后但并非最不重要的一点，我要感谢我的朋友和家人的无休止的鼓励，如果没有这些鼓励，这将几乎没有什么收获。

使用隐藏功能预测大规模在线公开课程中的学生保留率
马尔可夫模型

德里克·库切（Gerish K.Balakrishnan）

1引言

大规模开放式在线课程（MOOC）是一种基于网络的大规模课程，面向大量参与者。在过去的几年中，由于edX和Coursera等几个设计良好的在线教育网站的出现，以及诸如MIT，Stanford和UC Berkeley等顶尖大学对MOOC的兴起，MOOC受到了越来越多的欢迎。向广大公众开放各种课程。对于最终用户的吸引力在于，无论个人背景如何，都可以从任何地方获得高质量的教育。MOOC通常包含简短的讲座视频（10-15分钟）以及测验和家庭作业，以评估学生对主题的理解。

尽管它们具有优势和受欢迎程度，但开放的性质意味着留学生仍然是一个重大问题，因为几乎任何人都可以注册该课程，并且因课程失败而带来的后果很小。这样一来，大量的学生就开始报名参加该课程，而一开始就没有参加，并且学生在课程的几乎每个阶段都继续辍学（图1中说明了不同大学在不同网站上提供的两种不同的MOOC）。永不参与的学生的问题可能是由于课程本身的外部因素引起的，后一种现象表明，无论出于何种原因，学生都失去了完成课程的意愿。在体育学院和大学中也观察到这种行为，尽管规模较小[1]，但已尝试使用事件历史模型[2]来了解这种行为。事实证明，这样的模型对于推断学生离校的原因以及建议机构减轻问题的干预措施非常有用。

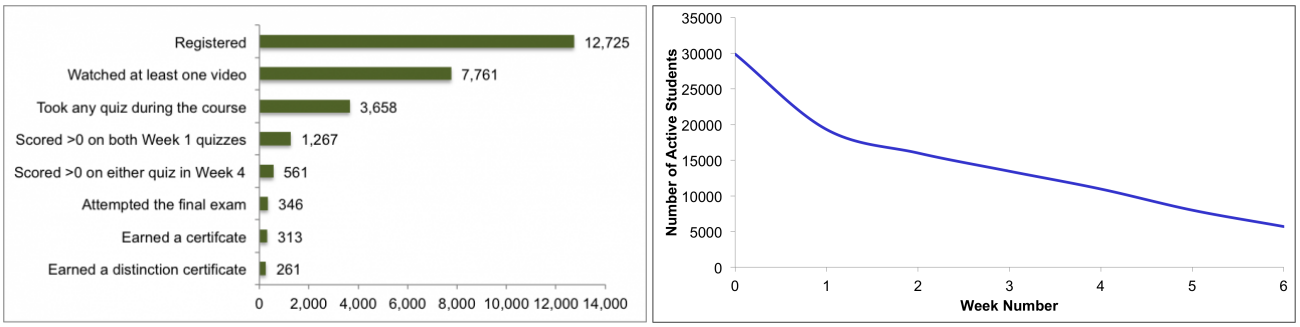

图1：（2012年秋季，杜克大学MOOC）（左）生物电学中的学生持久性[3]，2012年秋季（Berkeley edXs MOOC）（软件）即服务中的学生持久性。在这两种情况下，学生每周都要上课。

在这里，我们尝试使用隐马尔可夫模型（HMM，[4]）作为了解学生随时间变化行为的手段，以更大的规模为在线环境中的学生做同样的事情。HMM被证明是一个合适的选择，因为隐藏状态可以模拟学生的潜在特征，从而影响他们的毅力，然后我们可以从他们与MOOC的可观察交互中推断出这些特征。此外，针对MOOC的HMM允许我们根据学生的先前状态和当前可观察的动作来推断学生下一步的行为。由于有许多不同类型的可观察动作，因此我们探索了两种创建包含多个功能的复合模型的方法。在我们的

第一种方法是，我们为学生在课程的各个时间段内建立一个多维的，连续值的特征矩阵，并使用k均值聚类或跨乘积离散化将该特征空间量化为离散的可观察状态，即离散单流HMM不可或缺的一部分，如[5]中所述。然后使用这种技术，我们可以应用Baum-Welch算法[6] [7]在选定数量的隐藏状态上训练HMM。在第二种方法中，我们使用堆叠集成方法，在该方法中，我们训练几个单独的HMM，每个HMM都考虑一个可观察的特征，然后将其结果传递给Logistic回归器，我们对其进行训练，以预测学生是否会在下一时间被保留当然。

我们将注意力集中在edX的2012年秋季产品上*“CS169.1x-软件即服务”*，这是一门相对稳定的课程，已经在多种产品中成熟，并且包含了MOOC的大部分原型功能。建立模型后，我们想回答以下问题：

- 我们能否准确预测学生在不久的将来是否有可能放弃MOOC？
- 我们是否可以确定最终放弃课程的学生的行为模式，从而提出干预措施以防止这种情况发生？

2数据集

如前所述，我们专注于edX的UC Berkeley的2012年秋季产品*“CS169.1x-软件即服务”*
课程。这是一个为期六周的课程，招收了29,882名学生，并且采用以下格式：

- 11个讲座，每个讲座分为10-20分钟的视频，其中包含未分级的多项选择练习题
- 4个作业，每个作业包含不同数量的编程问题
- 4个不同长度的分级多项选择测验

重要的是要注意，除了第一个星期，由于没有分级作业，课程材料在整个课程的六个星期中平均分配。该课程还设有一个附带的论坛，具有诸如线程跟进和投票等基本功能。

该课程的数据直接从edX获得，后者会为每个课程生成完整的数据转储，并将UC Berkeley课程的数据分发给Berkeley研究人员。数据分为三个不同的实体，如下所示：

1.在整个课程期间，点击流数据均采用JSON格式，包括服务器和浏览器端事件。例如，学生与演讲视频的交互（例如单击“暂停”）被记录为浏览器端JSON事件，而访问的每个页面都被存储为服务器端JSON事件。

2.存储在MySQL数据库中的每个入学学生的作业成绩。

3.存储在MongoDB集合中的论坛主题和评论，以及有关回复次数，编辑次数等的元数据。请注意，被动论坛数据（例如，未收到线程的视图数量）未存储在此处，而必须从主题中推断出来。点击流数据。

3功能集

在定义HMM时，我们首先将课程分为六个时间片，其中一个时间片持续一周。这是一个实际的选择，因为该课程在材料方面平均分布于整个星期，并且可以合理地假设，活跃的学生每周至少访问一次该课程的网站，因为每周都会发布新的讲座和作业。因此，对于定义的每个功能，每个学生在课程的每个星期都将获得该功能的价值。换句话说，如果 $小号$ 是

一群学生 *F* 功能集和 *w ^* 周数，则特征矩阵将具有| *西|西 ×| F |。*

即使要考虑许多可能的功能，我们也只选择了几个跨越MOOC各个方面并且似乎彼此独立的功能，因为相关的观察状态会损害HMM的性能。在以下小节中，我们介绍每个选定的功能。在第7节中提到了我们希望在将来的实现中看到的其他一些功能。

3.1学生“进/出”状态

我们的主要特点是学生是否放弃了课程。这是我们最终希望能够预测的功能，并且被编码为粘性二进制值，即学生可以是活跃的参与者（“处于”状态），或者已经退出（“处于”状态），但是一旦学生放弃课程，他们将无法重新加入课程。这是一个合理的定义，因为我们的目的是尝试预测学生何时最终放弃该课程-稍后恢复课程的学生可能暂时无法参加，而不是放弃了课程。

对于给定的学生，可以通过检查点击流数据以检查他们访问课程的任何元素的最后日期来轻松计算此功能。该定义简单而彻底，可以捕获与课程的任何类型的交互，但是无法捕获行为的细微差别，例如放弃课程但决定以后再访问课程网站的学生赶上材料。根据我们的定义，这样的学生将被认为比实际活动更长的时间。预计这类学生很少。

3.2已观看的可用讲座视频的累计百分比

讲座视频分为10至20分钟，是将知识传递给学生的 主要手段。这些讲座每周发布一次，因此学生无法在课程中访问将来的资料，从而导致与视频的定期互动。此外，提供的点击流数据记录了学生观看的特定视频的秒数。因此，我们不仅可以简单地询问是否观看了讲座视频，还可以提出有关讲座视频的确切完成的更彻底的问题。

一个可能的功能是每周观看一次讲座视频的秒数，但这有一个局限性，即某些星期比其他星期有更长的演讲时段，这是由于一周中的讲座秒数分布不均。取而代之的是，我们考虑从上课开始到本周末观看的可用讲座的累计百分比。这不仅避免了由于使用百分比而导致分配不均的问题，而且该指标还间接衡量了学生在课程中的落后程度。例如，每周的累积百分比显着下降，说明学生落后了。

3.3在论坛上查看的线程数

课程论坛是课程中学生支持的主要手段，而学生可以进行的最基本的互动是查看问题和答案的线索。这是衡量学生参与课程的重要指标，因为它是大多数学生从事的被动指标。但是，点击流数据是每一次学生访问任意一个线程都会收集一次，这意味着，如果我们仅查看学生访问任意一个线程的次数，那么每天多次访问同一线程的学生似乎访问了很多线程比一天访问多个不同线程的学生多。为了克服这个问题，我们施加了一个限制，即给定的线程只能在一天中被学生唯一地查看一次。从而，

3.4在论坛上发帖的数量

与课程论坛的另一种基本交互方式是发布帖子，无论是提问，回复还是评论答案。尽管与发帖的学生数量相比，预计发帖的学生将会少得多，但我们仍将其作为学生参与度和社区意识的重要指标，因为这是学生自愿与社交互动的最积极的互动方式。课程。此外，将这种最活跃的论坛互动与最消极的论坛互动（查看线索）进行比较，可能会引起学生保留的兴趣。

3.5检查课程进度页面的次数

每个edX课程都使学生能够访问课程进度页面，该页面包含有关其完成的演讲模块和作业的信息。可以合理地假设，自从发布材料以来（每周一次），一个积极参与的学生每周都会监视他们的进度几次。在这种情况下，学生在给定的一周内检查其进度的频率表明他们参与了所有课程材料，尤其是作业，因为这可能是有用的功能。

4建模方法

我们所有的模型都是离散的单流HMM [5]，其特征是它们在每个时间片上都有一个离散值的观测变量。我们选择离散的单流HMM，而不是多流对应的HMM，还是连续HMM，是基于这样的事实，即这是最简单的HMM训练方法，并带有大量完善的工具支持。另外，已经发现[5]，当比较单流与多流离散HMM时，最重要的因素是所选功能的独立性，与所采用的实际模型相反。结果，我们用最简单的方法提出了最好的模型。话说回来，

此外，我们使用遍历HMM，在这种状态下每个状态都可以过渡到其他状态，而线性HMM则在状态下对应于各个星期。这提供了两个理想的属性：状态序列的转换（例如，出于相似的原因，使用相同的状态来模拟两个学生分别在第2周和第4周辍学的状态），以及在课程结束后将预测推断为假设的能力课程更长。

以下参数[4]充分描述了这种模型： *ñ* 对于隐藏状态的数量， *中号* 观察变量可以执行的可观察状态数和概率参数 *λ=（A，B，π），* 哪里 *一种* 是状态转移概率矩阵， *乙* 是观察状态概率矩阵，并且 *π* 是初始状态分布向量

大小 *N.* 此外，该组隐藏状态表示为 *H = {H*1，*H*2，...，*HNJ*与 *q t* 当时处于隐藏状态 *t，*
并且可观察状态集表示为 *V = {V*1，*V*2，...，*VMJ*与 *Ø t* 当时是可观察的状态 *t。* 有了这个定义，我们可以表达一个长度序列的概率 *，* 与观察状态序列 *o =（o*1，*Ø*2，...，*ØT），*
和隐藏状态序列 *q =（q*1，*q*2，...，*qT）* 如：

$$p(o,q)=p(q_1)\,(_1|q_1)\prod_{t=2}^{T}(_t|q_{t-1})\,(_t|q_t) \qquad (1)$$

在（1）中，概率 *（t/q t-1）* 和 *（t/q t）* 直接从 *一种* 和 *乙* 分别。

但是，在我们的情况下，我们的所有模型都包含多个功能，因为我们始终将学生“入” /“出”状态的主要特征与一个或多个其他特征结合在一起考虑。此外，尽管“输入” /“输出”状态是离散的，但我们所有其他功能都是连续的。因此，我们探索了几种在单个离散观测变量中对多维连续值特征集进行编码的选项。为了以下讨论的目的，我们表示功能集

如 *F = {F*0，*F*1，*F*2，...}，我们在哪里贴标签 *F*0 作为学生的“入” /“出”状态。

4.1比较不同的模型

在设计模型训练程序时，我们需要调整几个模型参数并选择某些建模方法。因此，在定义不同的模型之前，我们需要一种公平有效的方法来比较这些选择。最终，我们使用二进制分类指标（有关详细信息，请参阅第4.4节）评估模型，以评估模型在预测学生下一步是否会辍学方面的有用性。但是，在最终模型由很多部分组成的情况下（例如，如第4.3节所述，多个HMM后跟Logistic回归器），仅依靠这种方法会很耗时，因为我们必须等待整个模型都经过训练。因此，我们还定义了一个分数， *L，* 可以仅针对当前正在训练的HMM计算得出。 *大号* 是负对数可能性的总和

在训练数据集中为学生观察到的所有从“进入” /“离开”状态的序列。换句话说，如果训练数据集
由一组观察到的“输入” /“输出”状态序列组成，我们对所有可能的状态转移序列 *哦*， 每个观察到的序列在哪里 *Ø ∈ 哦*， 然后每个 *哦*，

问 G 我们

列求和 $q \in \sum$

$$L = - \sum_{Ø \in q \in 问} 日志 (p (o , q)) \quad (2)$$

哪里 *p (o , q)* 按照等式（1）计算。值越高 *L*， 我们的模型分配的预测实际发生的训练次数的概率越高。因此，我们绘制 *大号* 反对我们要做出的决定的选择，然后选择能够产生合理分数的最简单选择。例如，图2展示了一个HMM程序的示例，该程序仅对学生的“输入” /“输出”状态使用了1个附加功能，我们试图确定很多隐藏状态， *N*， 用于模型。在这里，由于生成完整模型非常简单，因此我们还比较了接收器工作特性（ROC）图的曲线下面积（AUC）的二进制分类指标。

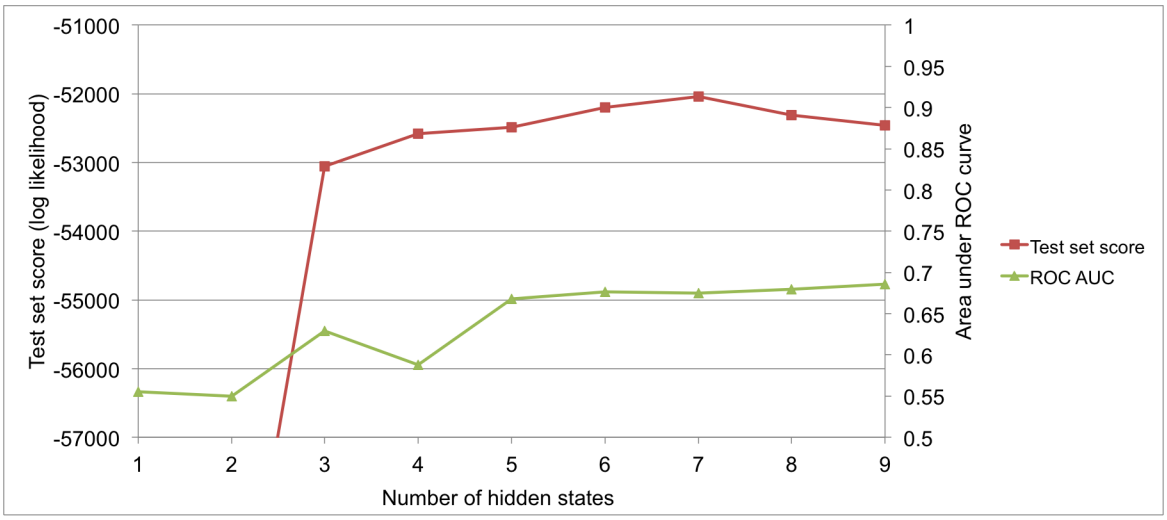

图2： *大号* 分数和AUC相对于隐藏状态的数量， *N*， 对于在学生的“入” /“出”状态下使用一项附加功能的HMM。对于此模型，选择了7个隐藏状态。

4.2具有单个功能且处于“入” /“出”状态的HMM

我们模型中最简单的是结合了 F_0， 具有其他功能的“输入” /“输出”状态， $F_{-世}$。这些
模型使我们能够理解个人特征和留学生率之间的关系。如果 $F_{-世}$ 是连续的，则首先使用以下方法之一离散化：

- k均值聚类-用于将学生聚类为相似行为的组
- 将值分成四分位数-对以百分比编码的要素很有用
- 使特征为二进制，即如果特征的值为0，则为0，否则为1-对于那些仅执行动作即可很好地指示学生与特征交互的特征的特征

方法的确切选择取决于每种功能最适合的方法，我们通过计算分数来比较选择
如第4.1节所述。一旦我们离散化 $F_{-世}$， 我们获得一个新变量 F_{ID} 可以接受一组离散值中的一个值 $D = \{D_0, d_1, ...\}$。然后，我们结合 F_0 与 F_{ID} 通过计算 *d* 与

可能的值 F_0 可以接受{“ in”，“ out”}。这将产生一个状态为{{ $d_{-世}$，“在”），（ $d_{-世}$，“ out”）} $一世 \in [0, |D|)$。

现在，我们可以使用Baum-Welch算法[6] [7]来训练概率参数 λ。 数据均匀地分为训练和测试集，并且仅使用训练数据。培训还要求我们指定 N， 隐藏状态的数量。由于我们不希望将含义附加到这些隐藏状态（尽管可以推断出某些含义），因此这成为我们使用4.1节中的方法选择的可调参数。

然后使用测试集评估模型， 小号，其中包含学生及其相关的特征向量序列。对于单个HMM，特征向量仅包含两个元素，一个用于 F_0 而另一个 $F_{-世}$，但这已扩展到我们的复合模型的多个功能。因此，每个学生 $s \in 小号$ 将具有特征向量序列

（ F_{s1}，F_{s2}，…，F_{s6}），哪里 F_{SW} 是学生的特征向量 s 一周 w。 总共有6个条目，因为该课程有6个星期。首先，我们计算从第一个开始的这些特征向量序列的所有子序列
一周，并在学生停课的那一周结束。我们不考虑自从他们放弃课程后的几周
学生不能重新参加课程，因此他们的价值 F_0 不能改变。称这组子序列 Z. 然后我们删除功能
F_0 从每个 $F_{SW} \in ž \in ž$ 并将它们分开，因为这是我们试图预测的功能。然后，我们使用训练中选择的方法离散化所有特征向量，并将离散化特征向量与“ in” /“ out”组合
该周学生的状态，以获取所有观察状态。让这组最终的观测序列为 $ž_{哦}$，然后是一个子序列 $ž \in ž_{Ø}$ 将包含以下形式的观察状态： $*$， F_0）其中*表示

特征向量和 F_0 =“进出” 。通过这种设置，我们可以质疑学生在下一时间段留在课程中的可能性如何， $t+1$， 得到他们的观察子序列， z 和隐藏状态序列 q 直到时间 t。
该概率可以表示为：

$$(0, t+1 = \text{“在”} \mid z, q) = \frac{p(F = \text{“在”} \cap ž \cap q)\, p(z, q)}{} \quad (3)$$

联合概率 $p(z, q)$ 与（1）中表示的相同，并且易于计算。定义所有可能的集合
状态转换为 $Q = 高 \times H$ 这样 $q \in 问$ 是形式（ $H_{-世}$，H_j），我们可以表达概率
留在下一个学生中的比例 $\sum$ 时间标准 $\sum$ as :

$$(0, t+1 = \text{“中”}) = \sum_{(H_{-世}, H_j) \in Q} \sum_{k=1}^{k} (t+1 = (k, \text{“中”}) \mid q_{t+1} = H_{-世}) \cdot (t+1 = H_{我} q_t = H_j) \quad (4)$$

这两个概率直接来自转换矩阵和发射矩阵 一种 和 B. 这样，根据我们到目前为止的观察结果，我们得出留在课程中的学生百分比。得出一个50％的阈值，然后我们可以对模型是否预测学生将在下一个时间段停留还是进行分类进行分类，然后将模型的答案与实际的答案进行比较，以评估模型的有效性。

4.3具有堆叠功能的HMM

虽然包含单个功能的HMM有助于了解学生针对单个功能的行为模式，但我们可以通过构建包含多个功能的复合模型来改进这些模型所做的预测。一种简单的方法是针对我们要考虑的每个功能分别训练各个HMM（如第4.2节所述），并采用堆叠集成方法，其中将各个HMM的预测概率输入到现成的中逻辑回归器，将权重分配给每个特征模型，并在下一个时间步骤中计算学生租借的总体预测概率。

具体来说，由于我们正在尝试预测下一个步骤的学生保留率，因此我们计算 $P_{说}$（ $0, t+1 =$
“在” $ž_{-世}, q_{-世}$）对于每个HMM， 1个 $\leq$ 一世 $\leq |F|$。 这个值是 一世 HMM的似然值。下一个时间段留在课程中的学生， $t+1$， 给定他们的观察子序列 一世 HMM， $ž_{-世}$，和隐藏状态序列 一世 HMM， $q_{-世}$，直到时间 t。 用于计算这些值的方程式与（3）和（4）相同。物流

回归者有| F |功能，每个HMM都有一个 P，值是逻辑回归器的特征值。这使我们可以训练回归者将学生分类为bein g在下一个时间步骤中输入或输出。

为了评估该模型，对于测试集中的每个学生，每个单独的HMM都会为 $P_{说}$（ $0, t+1 =$
“在” $ž_{-世}, q_{-世}$）。然后，我们简单地喂所有这些 P，到训练有素的后勤回归者，对他们进行分类。学生会

保留在下一个步骤中。如第5节所示，堆叠方法往往比单个HMM产生更好的预测，同时在合并的功能数量方面也非常可扩展。

4.4使用K均值聚类和跨产品离散化的具有多种功能的HMM

在建立复合模型的这种方法中，我们定义了一个特征矩阵， C，尺寸| 西|西 ×| F |，哪里 小号 是个
训练集中的一组学生 w ^是课程的周数，并且 F 是功能集。进一步， $C = \{c_{swj}\}$
哪里 C_j 是功能的价值 F_J 对于学生 s 在一周内 w。 我们首先删除功能 F_0，由于这是我们试图预测的功能，因此需要能够控制哪些观察状态与“处于”状态相对于“观察”状态相对应。
“出”状态。然后，我们转换此子矩阵， $C' \subset C$，使用k均值聚类将其转化为单个变量。输入 C'
进入聚类算法返回 k 聚类以及来自任何特征向量的映射 $f = \{f_1, F_2, ...\}$ 到以下之一
的 k 集群。这些 k 聚类与以下两个结果的每一个结合 F_0，导致我们 中号 可观察的状态
即 $M = 2 K$. 同样，在测试阶段使用来自聚类的映射来获取HMM的观察状态
根据学生在我们的测试集中展示的实际记录的特征序列。 k 是我们使用第4.1节中的过程调整的参数。

因为k均值聚类可能无法为特征向量的某些分布做出较差的选择，所以我们还探索了一种更简单的离散化方法，其中我们为每个特征分别选择了一个即席离散化方法（如4.2节中的离散化方法），然后将这些单独的离散量组合在一起使用简单叉积的一组值，产生的观察状态总数等于每个特征的离散状态数目的乘积。这往往会产生更好的预测，但是由于产生大量的观察状态，因此在缩放方面存在问题，其中大多数状态在训练中不会遇到。由于可以对同一特征进行多个特定的离散化，因此我们通过使用每个离散化构建各种HMM并比较它们的预测性能来选择最佳的离散化。

这些模型的评估类似于第4.2节中规定的评估程序。

5结果

正如导言所述，我们的目标是双重的：在与MOOC的互动以及他们留在课程中的倾向方面，确定学生行为的定义模式，并预测学生是否有可能留在学校学习。下个星期。因此，我们将结果作为单独的部分介绍给两个目标。

5.1学生行为模式

首先，我们研究一些有趣的单个HMM，这些HMM使用单个功能结合“入” /“出”状态。为了生成这些预测，我们预测学生在课程中达到某一星期而表现出某种行为模式时不会掉落的可能性，然后将其与学生在表现出该行为然后降落该点时达到该点的可能性进行比较。下周。这使我们可以检查整个课程期间特定行为的影响。

5.1.1关于学生检查课程进度的频率的模式

图3a显示了每周不断检查自己的进度几次的学生与下周放弃课程的可能性之间的关系。很少或从未检查过进度的学生在上周下降最多，而在下周，从未检查过进度的学生下降频率则更高，下降率接近40％。可以用以下事实来解释：如果一个学生在第4周和第5周还没有检查他们的进度，这时应该进行几次评分作业，那么他们更有可能甚至没有尝试过这些作业，因此倾向于放弃课程。

毫不奇怪，那些每周不断检查4次或更多次进度的学生，放弃课程的可能性很小（ 一2％）。

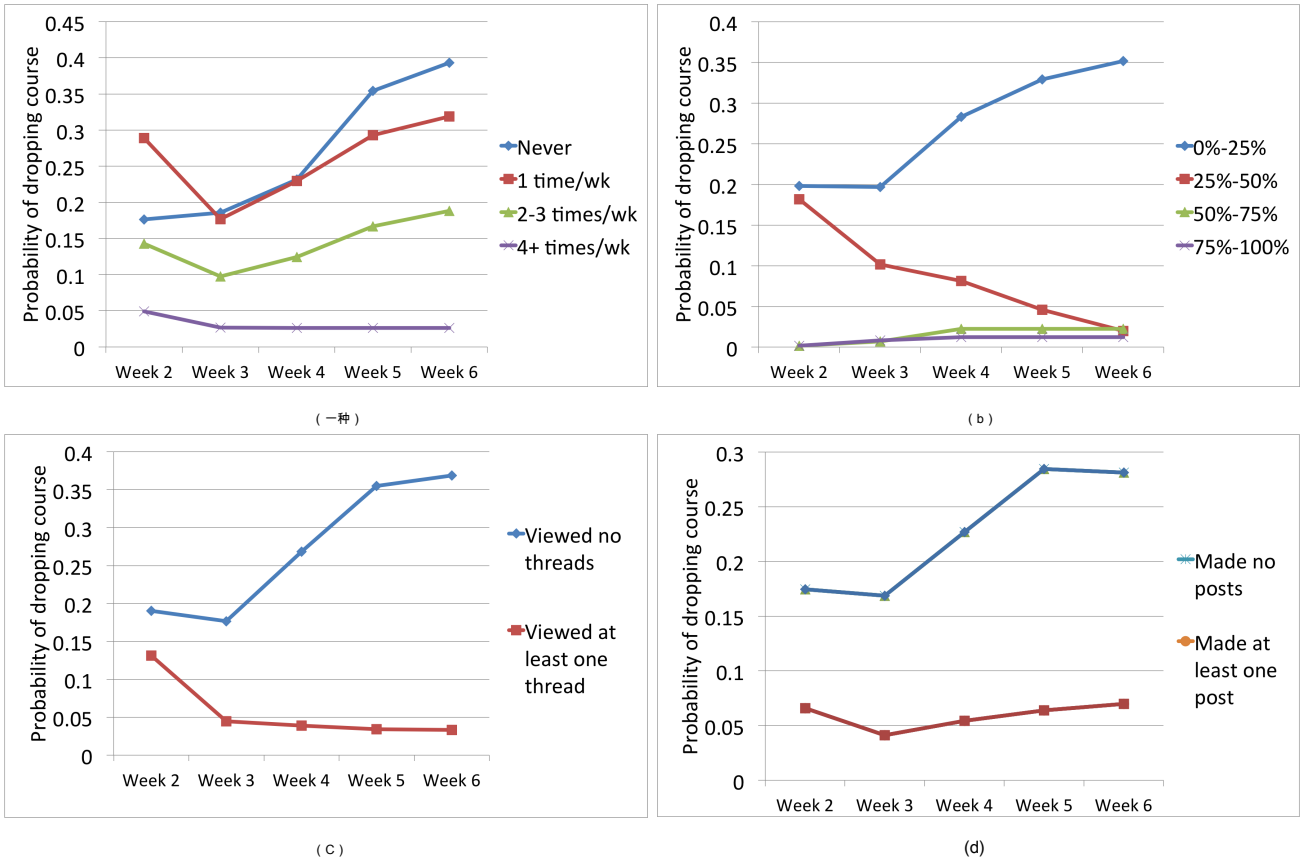

图3： 时间流失的学生：

(a) check their course progress with a consistent frequency every week

(b) view a consistent percentage of lecture videos each week, view a consistent percentage of lecture videos each week (c) consistently view or do not view forum threads each week

(d) consistently post once per week or consistently do not post on the forum

As an example, the green line in (a) shows the probability that a student who checks their course progress 2-3 times a week will dropout of the course each week. So, if a student maintained such consistent behavior for the first 4 weeks, their likelihood of dropping the course in week 5 is about 17%.

5.1.2 Patterns with regards to the cumulative percentage of lecture videos watched

Figure 3b shows the relationship between students who consistently view a certain percentage of lecture minutes each week, and their likelihood to drop the course in the next week. As would be expected, students who watch more of the lectures are less likely to drop - students who watch at least 50% of lecture minutes are extremely unlikely to drop (<0.1% throughout). Interestingly, as the course continues, students who watch no lectures become more likely to drop, exceeding a 35% drop rate, and students who watch only 25%-50% of lectures become very unlikely to drop, reaching 2.2%. This suggests that watching lectures is important, but watching them in their entirety becomes less important toward the end of the course.

5.1.3 Patterns with regards to the number of forum threads viewed

Figure 3c shows the relationship between whether students view forum threads each week, and their likelihood to drop the course in the next week. In the first week the difference is relatively small, but towards the end of the course students who never view the forum become very likely to drop (37%), whereas those who view at least one thread a week are very unlikely to drop (4%). This suggests even minimal interaction with the forum can be a crucial factor for retention.

5.1.4 Patterns with regards to the number of forum posts made

Figure 3d shows the relationship between whether students post forum threads each week, and their likelihood to drop the course in the next week. Students who post on a weekly basis are very unlikely to drop (consistently in the rage of 4-7%). The probability of dropping for students who do not post is lower for this feature than for others considered above, even in week 6 (about 28% as compared to 35-38%), consistent with the notion that students can participate actively in the course without actively contributing to the forum.

5.2 Immediate Student Retention Predictions

We explored individual HMMs that incorporate a single feature as well as two alternatives for composite HMMs, all of which allowed us to classify students as being in or out of the course at every time step subsequent to the first week of the course. For every student in the test set, we examine all subsequences of actions leading up to the point at which they dropped, and predict whether they will drop in the next timestep (we do not examine subsequences following the drop point because these are trivial to predict - students who are dropped remain dropped).

We evaluate our predictions using standard binary classification metrics, such as precision, recall, and F1 scores, as well as the Area Under the Curve (AUC) score for the Receiver Operating Characteristic (ROC) plot, and the Matthews correlation coefficient. In figure 4, we show the ROC plots for the best composite model, in table 1 we summarize the results for both composite models, and in table 2 we summarize the AUC scores for the individual HMMs.

	Composite model using cross-product discretization	Composite model using stacking
Accuracy	0.801	0.805
Matthews Correlation Coefficient	0.137	0.119
ROC AUC Score	0.710	0.696
Precision	Positive: 0.807 Negative: 0.558	Positive: 0.807 Negative: 0.647
Recall	Positive: 0.987 Negative: 0.064	Positive: 0.995 Negative: 0.036
F_1 Score	Positive: 0.888 Negative: 0.115	Positive: 0.891 Negative: 0.068

Table 1: Summary of binary classification evaluation metrics for the composite HMMs. For precision, recall and F1 scores, positive queries are where we ask if a student is staying in the course, and negative queries are for students dropping out.

Feature considered by individual HMM	AUC Score
Number of times the course progress page was visited Cumulative	0.692
percentage of lecture videos watched Number of threads viewed	0.656
	0.657
Number of posts made	0.609

Table 2: Summary of AUC Scores for the different individual HMMs. Each individual HMM is identified by the single feature, in addition to the “in”/“out” state, that is incorporated into the model.

In general, the precision, recall and F-1 values show that our model is relatively poor at predicting negative queries, al- though this may be mostly due to the fact that there are many more positive queries than negative queries, since a sequence of “in”/“out” states for a given student contains several “in” states (positive queries), but only one “out” state (negative

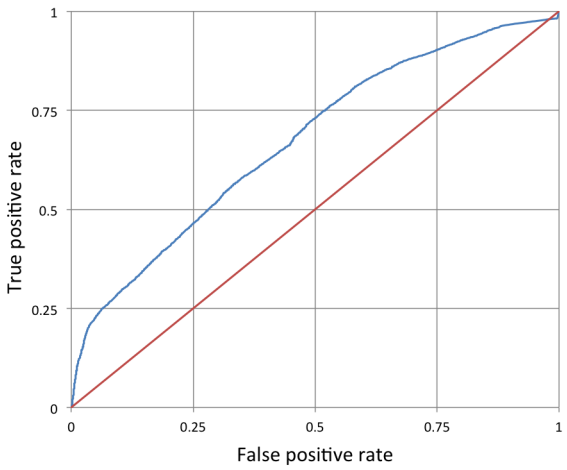

Figure 4: ROC Plot for composite HMM that employs the cross-product discretization method. The line through the center of the plot has area under the curve of 0.5, and denotes a classifier that is no better than random.

queries). As a result, it may not be too instructive to pay attention to these figures. A much better metric is the ROC curve, which indicates that our best classifier is quite a bit better than random, with an AUC score of 0.710.

Both composite models exhibit higher ROC AUCs than the HMMs using single features (comparing table 1 and table 2), suggesting that both techniques somewhat effectively incorporate information from multiple features. The relatively small gain suggests considerable dependency between features.

The cross-product discretization method yielded higher overall ROC AUC than the stacking method. This can be explained in part by the HMMs ability to perform state transitions based on combinations of multiple feature values. However, this advantage is offset by a much longer training time due to the large size of the observed state space.

6 Conclusion

Overall, we were able to design effective Hidden Markov Models to help predict student retention as well as to infer general patterns of behavior between those students that complete the course, and those that drop at different points in time. Our composite HMM that incorporated multiple features produced a reasonable ROC curve with an AUC value of 0.710. Lastly, our individual HMMs offered insight into certain patterns of student behavior, such as the fact that a student who never checks their course progress dramatically increases their probability of dropping out of the class only after the fourth week of the course. While this is purely correlational, it does offer some interesting insight into how students interact with the MOOC and can be used to suggest behavior changes to students that are headed towards dropping the course.

7 Future Work

The most obvious extension to our current methods is the inclusion of more features from theMOOC to enrich our composite model, as well as bring to light more insightful patterns in student behavior. Some of these features include:

- Scores received on homeworks/quizzes
- Number of posts followed and/or upvoted on the forum
- Number of replies or upvotes received on posts made
- Percentage of only this week's lecture videos watched

In addition, it would be instructive to model the course using Kalman filters, as opposed to our current approach of quantiz- ing a multidimensional feature matrix of continuous values. This would preserve some of nuances in the data that is lost in the current process, and may yield better predictions.

Finally, one could explore alternative definition of what it means for a student to be an active participant of the course. Currently, our model does not take into account complex patterns of behavior, such as those students who leave the course for one or two weeks but come back to finish the course. A definition that incorporates these subtleties would enable us to gauge how good the adaptable the course is to serve such individuals.

Tools Used

- Python - The primary language used for feature extraction, and model creation as well as inference.
- GHMM Library - General Hidden Markov Model LIbrary implemented in C with a Python interface, used for creating our models (http://ghmm.org/).
- Scikit-Learn - Python machine learning library, used for k-means clustering (http://scikit-learn.org/). [8]

References

[1] V. Tinto, "Dropout from Higher Education: A Theoretical Synthesis of Recent Research". *Review of Educational Research,* vol. 45, no. 1, pp. 89-125, 1975.

[2] S. L. DesJardines, D. A. Ahlburg, B. P. McCali "An event history model of student departure". *Economics of Education Review,* vol. 18, issue 3, pp. 375-390, 1999.

[3] Y. Belanger, J. Thorton, "Bioelectricity: A Quantitative Approach - Duke Universitys First MOOC", *Duke University Report,* February 2003.

[4] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in *Proceedings of the IEEE,* 77(2):257-285, February 1989.

[5] J. Schenk, S. Schwarzler, G. Rigoll, "Discrete Single Vs. Multiple Stream HMMs: A Comparative Evaluation of Their Use in On-Line Handwriting Recognition of Whiteboard Notes", in *Proceedings of the International Conference on Frontiers in Handwriting Recognition,* August 2008.

[6] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models". *Technical Report TR-97-021,* International Computer Science Institute, Berkeley, CA., 1998.

[7] L. Baum, T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". *Annals of Mathematical Statistics,* 37:1554-1563, 1966.

[8] Pedregosa et al., "Scikit-learn: Machine Learning in Python", JMLR 12, pp. 2825-2830, 2011.