Modeling Student Learning Styles in MOOCs

Yuling Shi
Wuhan University
Wuhan, Hubei
sylyjs@whu.edu.cn

Zhiyong Peng
Wuhan University
Wuhan, Hubei
peng@whu.edu.cn

Hongning Wang
University of Virginia
Charlottesville, Virginia
hw5x@virginia.edu

ABSTRACT

The recorded student activities in Massive Open Online Course (MOOC) provide us a unique opportunity to model their learning behaviors, identify their particular learning intents, and enable personalized assistance and guidance in online education. In this work, based on a thorough qualitative study of students' behaviors recorded in two MOOC courses with large student enrollments, we develop a non-parametric Bayesian model to capture students' sequential learning activities in a generative manner. Homogeneity of students' learning behaviors is captured by clustering them into latent student groups, where shared model structure characterizes the transitional patterns, intensity and temporal distribution of their learning activities. In the meanwhile, heterogeneity is captured by clustering students into different groups. Both qualitative and quantitative studies on those two MOOC courses confirmed the effectiveness of the proposed model in identifying students' learning behavior patterns and clustering them into related groups for predictive analysis. The identified student groups accurately predict student retention, course satisfaction and demographics.

CCS CONCEPTS

•**Information systems** → *Data analytics; Clustering;* •**Applied computing** → *Interactive learning environments;*

KEYWORDS

Behavior Modeling; Sequential Data Mining; MOOCs; Probabilistic Modeling

1 INTRODUCTION

Massive open online course (MOOC) have attracted extensive public attention and popularity over the past few years [15]. Learners from all over the world can access enormous course materials in a wide range of subjects without time, location and personal background restrictions. MOOCs empower educators and students to test new modes of teaching or learning at scale. However, due to their open and online characteristics that are distinct from traditional classroom education environment, experiences and pedagogical models from conventional offline education do not fully satisfy the needs in MOOCs [6]. MOOCs have been criticized for its early dropout, low completion rate [10], lack of personalized support and feedback [13], and strong dependency on participants' self-regulation [2]. Increasing research attention has been paid to analyzing student behavior data in MOOCs for better understanding their learning activities, quantifying the quality of teaching and learning, and further improving MOOCs' educational effectiveness.

Existing research on analyzing and mining students' behavior patterns has varied across multiple aspects including identifying and classifying engagement styles [1, 20, 24], predicting dropout, grade of assignments or certificate [3, 9, 16, 19, 26], and recognizing students who need help [23]. A less explored but perhaps a more significant aspect is the ways in which the students prefer to learn. In pedagogy, *learning styles* describe a learner's preferred way of gathering, processing, interpreting, organizing and analyzing information [18]. Learning style can be considered as a *contextual* variable or construct, because what a learner brings to the learning is as much a part of the context as are the important features of the experience itself. Each learner can have distinct but probably consistent preference in perception, organization and retention. These learning styles are characteristic cognitive, affective, and physiological behaviors that serve as good indicators of how learners perceive, interact with, and respond to the learning environment.

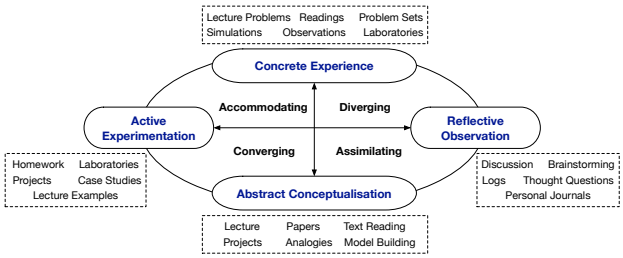


Figure 1: Learning styles in LSI model [4, 22].

Take the Kolb Learning Style Inventory (LSI) [4, 22], which is one of the prominent learning style models, as an example. As shown in Figure 1, LSI asserts a four-mode learning cycle and defines four learning styles – accommodating, converging, diverging and assimilating, where activities from traditional learning environments accommodate each of the mode. An ideal learning process should engage all four modes in response to situation demands; however, learners always exhibit certain preference on several of the four styles. Understanding, analyzing and modeling a student's unique preferred learning style effectively would help students with learning and retaining knowledge, and also help instructors be more aware of the learning progress of a particular student, so as to offer individualized or group instructions to improve students' learning efficiency, depth, retention and enjoyment. However, the LSI learning style assessment is typically performed via surveys that are independent of the specific subject of learning and before the course of learning. It thus lacks the flexibility of tracking students' learning behaviors in a highly dynamic and interactive environment at scale, such as MOOCs.

CIKM'17, November 6–10, 2017, Singapore.

ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: http://dx.doi.org/10.1145/3132847.3132965

Fortunately, the recorded student activities in MOOCs provide us a unique opportunity to model students' learning behaviors, identify their preferred ways of learning and underlying learning intents, and enable personalized assistance and guidance. To capture the homogeneity among students' learning activities, we introduce the concept of *latent student groups*. Each latent group possesses a unique distribution over students' learning behavior patterns observed in the MOOCs activity log data, such as transitional patterns, intensity and temporal distribution of their learning activities. A population of students are therefore modeled as a mixture over the latent student groups, and each student can be best characterized by one of the latent groups, reflecting their learning styles preference. Our generative assumption can be naturally mapped to the Kolb LSI in the MOOC environment: assimilators might repeatedly watch video lectures, while accommodators might take frequent actions transiting between video lectures and homework assignments. We assume these various activity patterns are governed by the students' underlying learning intents, and different latent student groups consist of distinct composition of learning intents and styles. For instance, assimilators switch between video watching and course forum discussions to confirm their understandings, while divergers do these to explore related topics.

To model transitions among students' latent learning intents based on recorded activties, we characterize a latent student group with a Hidden Markov Model (HMM), where hidden states depict students' latent learning intents via their association with the observed learning activities. In addition to modeling transitional behavior patterns in each latent group, we also capture the intensity of activities with a Poisson distribution and the temporal distribution of activities with a Beta distribution in each group. Furthermore, due to the dynamic and nonrestrictive nature of MOOCs, and the diversity of student population, it is infeasible for us to manually exhaust the optimal configuration of latent student groups. Instead, we impose a Dirichlet process prior [5] over the latent student groups to explore the clustering property of students' learning behaviors from data. This leads to a non-parameteric Bayeisan model, which we name as Latent Learning Style model, or L^2S in short.

Knowledge discovered by L^2S from MOOCs' student activity data helps improve students' online learning experience. Characteristics of the learned latent student groups, such as different transitional patterns between students' latent learning intentions and the composition of such patterns inside each group, provide an in-depth understanding of students on MOOCs. We evaluate the developed model on two large MOOC courses with more than ten thousand student enrollments. Promising predictive power of our model is confirmed when compared with several state-of-the-art algorithms for student learning behavior modeling, students completion, dropout and course satisfaction prediction. Meaningful behavioral patterns are also identified through the learned latent student groups, which can guide future efforts in personalizing course content and students' learning experience.

2 RELATED WORK

Analyzing and modeling students' learning behaviors in MOOCs have emerged as an important research topic in data mining community. Such studies help instructors and educators to better understand students' behavior at scale, and enable MOOC providers to improve their online learning environment. Current research effort can be categorized into two major bodies: one focuses on qualitative analysis of students' learning activities, and one focuses on building predictive models about students' learning outcomes.

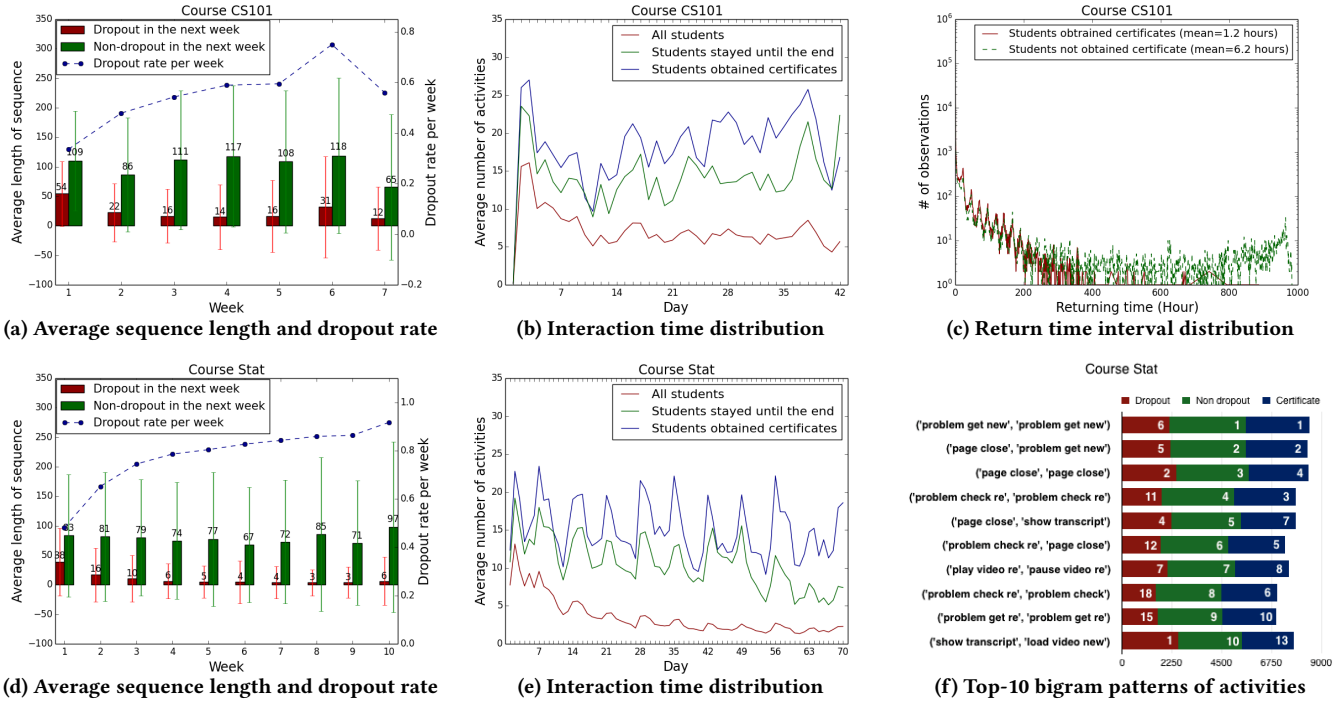
Various behavioral and demographics features have been explored to analyze students' learning activities. Guo et al. [8] investigated how students' navigation strategies in MOOCs vary by demographics. Wilkowski et al. [25] found no correlation between prior skills and course completion rates based on demographic results. Kizilcec et al. [11] integrated different kinds of signals, including students' behavior data, demographic data, geographic data and course enrollment data, and divided students into a few stereotypes based on these features. Seaton et al. [21] examined various activities that help students get certificates, such as time spend on tasks, frequency of accesses, and percentage of accessed course components. Wen et al. [24] viewed students' engagement as a latent variable and focused on their social behaviors such as forum posting and subscribing in addition to other behaviors such as following course materials and completing assessments. Anderson et al. [1] developed a taxonomy of individual engagement style to study the relationship between students' engagement and their grades. However, most of the above works only focus on qualitative analysis of student behavior patterns, and little attempt was made in building a predictive model on top of the identified patterns. Our work differs from them by formalizing the findings from our behavioral analysis and imposing generative assumptions about the behavior patterns for predictive modeling.

On the other hand, increasing attention has been paid to model students' learning behavior for different prediction tasks, such as predicting student dropout, course completion, and grade. Qiu et al. [19] developed a latent dynamic factor model by incorporating students' demographics, forum activities and learning behavior to predict their learning effectiveness. Wang et al. [23] developed a nonlinear state space model to predict students' re-visitation sequence to the course materials. Sequential state models, such as Hidden Markov Models and Recurrent Neural Networks, have been investigated to model students' sequential learning behavior. Ramesh et al. [20] proposed a latent representation model that describes abstract student engagement types and predicts student dropout based on the observed behavior sequences. Coleman et al. [3] developed a probabilistic topic model to cluster students by considering their interactions with courseware as a "bag of interactions." Yang et al. [26] evaluated the social factors that affect dropout by analyzing students' posting behavior in the discussion forum, then proposed a survival model for dropout prediction. Most of these existing research focuses on one specific type of prediction task, and develops supervised models to learn from annotated data. Our solution builds an unsupervised Bayesian model to capture a rich set of students' learning behaviors for predictive modeling. The learned model benefits a set of prediction tasks, including predicting student retention, course completion and satisfaction, although no direct supervision is needed for model estimation.

3 LEARNING BEHAVIOR ANALYSIS

In this study, we collected student learning activities from two edX courses, i.e., "Computer Science 101 (CS101, Summer 2014)" and "Statistical Learning (Stat, Winter 2015)," via a data-sharing agreement with the Stanford University [1]. Example of recorded

[1] http://datastage.stanford.edu



(a) Average sequence length and dropout rate (b) Interaction time distribution (c) Return time interval distribution

(d) Average sequence length and dropout rate (e) Interaction time distribution (f) Top-10 bigram patterns of activities

Figure 2: Behavior pattern analysis in two MOOC courses with large enrollments.

student learning activities in each course include viewing video lectures, attempting graded quizzes and homework assignments, and participating in course forum discussions.

Table 1: Dataset description

Course	Length	# Students	# Logged Activities
CS101	7 weeks	3,078	3,332,743
Stat	10 weeks	8,090	10,328,000

Although in each course there are more than ten thousand students, a small portion (around 30%) of them actually accessed the course content. We extract a subset of enrollments for our analysis and later experimentation, where we require each student has at least two logged actions among five specific categories, i.e., "video", "problem", "forum", "navigate", and "access", in the first week of the course. We define the five categories by aggregating similar click events in the log data. For instance, the "video" category corresponds to the aggregation of events like "play video", "stop video", "pause video", "seek video", "speed change", "load video", "show transcript", and "hide transcript". The final selected datasets consist of 3078 and 8090 students for CS101 and Stat, respectively. Basic statistics about these two courses after selection are reported in Table 1.

Each video, quiz or homework assignment in the log has a unique object id, and the click stream logs record the timestamp of student access to such objects. To differentiate a student's repeated actions on the same object, we label the actions based on the order of each object accessed by a student. For example, if a student watched a video for the first time, the activity event in the click stream is labeled as "play video new"; but if the student watched this video again, we label it as "play video re". This differentiates students' learning and reviewing behaviors. After this labeling, there are in total 53 event types defined in our study.

As student retention remains a serious concern in MOOCs, we start our qualitative analysis of students' learning activities with the statistics about students' dropout rate. In particular, we define a student's dropout status for week t as no learning activities in the coming week. In each week of the two courses, we analyze the average number of actions a student took. From Figure 2a and 2d we can observe that a large number of students dropped out during the period of these two courses. The dropout rate increased from 40% to 92% in both courses. However, it is important to note that in both courses the average behavior sequence length differentiates students who kept being active from those who disappeared in the next week. Given the fact that these two courses belong to different subject areas, attracted different number of students, and have various student population in age, gender and educational levels, it strongly indicates the amount of activities a student takes in an online course is an effective factor to characterize his/her engagement with the course in the near future.

Since students can learn at their own pace and choose when to engage with course content in MOOCs, we then analyze the temporal distribution of students' learning activities during the course span. In Figure 2b, the red curve presents the average number of activities taken per day by the enrolled students in the CS101 course. We can notice the intensity of students' activities peaked at the first week and gradually decayed afterwards. To understand if this is general in all students, we further examine the activity distribution in the students who kept engaging until the end of the course and those who finally got their certificates (as shown in green and blue curves in Figure 2b). It turns out that the activity intensity distributes more uniformly in these two groups of students. In particular, we can notice a spike of activities in the last week of the course, when the students are required to finish certain actions in order to get the certificate. This indicates that students who are

more motivated, e.g., to obtain the completion certificate, behave more consistently. We get similar conclusions in the Stat course in Figure 2e that the intensity of activities remains broadly flat over the course span among the students who kept engaging until the end and those who got the certificates. In particular, because this course released its new course content in a weekly basis, active students' activities peak at the beginning of each week. This suggests that the temporal distribution of students' learning activities reflects their latent learning intents and also characterizes the clustering property of students.

In addition to exploring when students perform actions, the return time interval between two consecutive actions also carries important information about their learning experience. A shorter return time indicates a student is more engaged with the course content, as he/she continues the action flow; but a longer return time might indicate the student is losing his/her interest in the course. Figure 2c shows that the distribution of return time intervals in students who obtained the certificates versus those who failed to do so in the course CS101. Similar results were obtained in the course Stat, but due to space limit we cannot illustrate them. It is clear that the distribution of return time intervals in students who failed to get the certificates has a much longer tail. The average return time of students who got the certificates in course CS101 and Stat are 1.2 hours and 1.19 hours, which is much shorter than that of students who failed to complete the course (6.2 hours and 5.7 hours respectively). These observations suggest that besides how many actions a student takes in an online course, how soon they move onto the next action also differentiates them and indicate their learning outcomes.

To get a more detailed understanding of how students carried out their specific learning activities, we study their action sequences by segmenting them into consecutive sub-sequences. We demonstrate the top-10 most frequent first-order transitions between actions (referred as bigrams) in the course Stat in Figure 2f (CS101 has similar results but due to the space limit the results are omitted here). The left side of the bar chart list the top-10 bigrams, the number on each bar indicates the ranking of this bigram in the corresponding group of students, and length of each bar represents the bigram frequency of in the corresponding student group. We can observe that the first-order activity transitions taken by drop-out students were mainly related to videos and course information pages, while in the non-dropout students and the students who got the certificates, they focused more on problem-based course content. The results further suggest that the detailed action sequence a student has taken reveals his/her latent learning intent, and it further differentiates students who have similar aggregated activities but with different learning or behavior intents.

All these qualitative studies of students' learning activities provide us concrete foundations to build analytic models to characterize students' behaviors, infer their latent learning intents, and predict their learning outcomes. This becomes the focus of our next section.

4 METHODOLOGY

Our goal in this work is to characterize and model students' learning behaviors by capturing the homogeneity among students, which we define as *latent student groups*. Based on the findings in Section 3, we make generative assumptions about a student's learning activity sequence, including its temporal distribution and transitional patterns. The latent student groups are identified by imposing a Dirichlet Process prior over the clustering property of students' learning behaviors. In this section, we describe in detail of how to formulate latent learning styles via computational models and identify the latent student groups via statistical inference.

4.1 Latent Learning Style Model

Formally, we denote a collection of V students as $U = \{u_1, \ldots, u_V\}$, where each student $u \in U$ is associated with an observed behavior sequence $\mathbf{x}^u = (\mathbf{x}^u_1, \ldots, \mathbf{x}^u_W)$ across W weeks in a particular MOOC course. Each sub-sequence in week w is denoted as $\mathbf{x}^u_w = \{x^u_{w,1}, \ldots, x^u_{w,T^u_w}\}$, in which element $x^u_{w,t}$ represents the t-th action taken by student u in week w and T^u_w is the total number of actions observed in week w. The timestamp associated with $x^u_{w,t}$ is denoted as $ts^u_{w,t}$, where the corresponding timestamp sequence of student u is defined as $\mathbf{ts}^u = (\mathbf{ts}^u_1, \ldots, \mathbf{ts}^u_W)$.

As suggested by our findings in Figure 2f, we assume the observed learning action sequences are governed by the students' underlying learning intents, and the transition of learning intents leads to the development of their detailed learning actions. Because the learning intents are not observable in the MOOCs log data, we model them with latent states. Specifically, we denote the latent state sequence as $(\mathbf{y}^u_1, \ldots, \mathbf{y}^u_W)$, where $\mathbf{y}^u_w = \{y^u_{w,1}, \ldots, y^u_{w,T^u_w}\}$ and $y^u_{w,t}$ takes value in $\{s_1, s_2, \ldots, s_N\}$ to indicate student u's latent learning intent at his/her t-th action in week w. Formally, we assume each student's observed learning sequence $\mathbf{x}^u$ is governed by the corresponding latent state sequence $\mathbf{y}^u$ via a first-order Hidden Markov Model (HMM). Since both the learning actions and latent states are discrete random variables, multinomial distributions are used to model the emission and transition probabilities. In particular, the emission probability characterizes the clustering of learning activities under the same learning intent, and the transition probability depicts a student's preference in how to perform those different learning intents in order. This generative design also aligns with the basic idea of learning style in pedagogy [18].

Because a Markov model only has a short-term memory, it cannot capture the temporal distribution of the learning actions over a course span. But as shown in Figure 2b and 2e, such temporal distribution well differentiates students' learning intents. For instance, in the early stage of a course, a student's intent is more likely to be experiencing the course, which is reflected on the concentration of video related actions; while in the later stage of the course, the intent is more likely to shift to practice with the learnt knowledge, reflected on the actions in homework and quizzes. To capture this, we enhance our HMM-based sequence model by introducing a Beta distribution to model the timestamp of each action in the emission probabilities. We refer to this as time-based emission. We assume different latent states are associated with different Beta distributions, i.e., $ts^u_{w,t} \sim Beta(\Omega^{s_i}_1, \Omega^{s_i}_2)$ if the corresponding latent state $y^u_{w,t} = s_i$. Accordingly, we normalize the timestamp $ts^u_{w,t}$ of each action $x^u_{w,t}$ into the range of 0 to 1 by the overall duration of a course. Furthermore, to capture a student's activity intensity, which also differentiates their learning intents as shown in Figure 2a and 2d, we use a Poisson distribution to model student u's action sequence length T^u_w in week w.

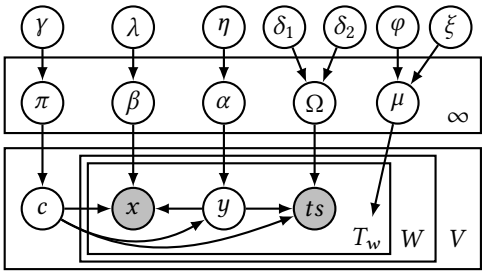


Figure 3: Graphical model representation of L^{2}S model. Light circles denote the latent random variables, and shadow circles denote the observed ones. The outer plate indexed by V denotes the students, the inner plated indexed by T_w denotes the behavior sequences from a student u in week w, the middle plate indexed by W denotes the week, and the upper plate denotes the parameters for the countably infinite number of latent student groups in the collection.

In addition, as suggested by Figure 2c, it is important to model not only when an action is taken, but also how soon the next learning action happens. A delayed return indicates a student's intent might shift. To differentiate the return time intervals in students who are actively engaging with the course content from those who temporally leave the course, we introduce a special learning action named *idle*. Based on the studies in time-based session analysis of search engine logs [14], we insert idle actions into the behavior log when the return time between two consecutive actions exceeds a threshold. Following [14], we fit a power law distribution of return time intervals to determine the timeout threshold. In particular, if a return time interval is m times longer than the threshold, m consecutive idle actions will be inserted with evenly distributed timestamps in this interval to reflect the long gap. To accommodate this special learning action in our HMM model, we introduce an idle state s_{idle}, whose emission probability is confined to the idle action only, and no other latent state can generate the idle action.

Putting these components together, we get a comprehensive behavior model describing students' learning activities on MOOCs. Parameters of each model component summarize students' learning behavior patterns (or learning styles). To construct a fully generative model, we need to specify the generation of those model parameters as well. Given the diversity of MOOC student population, it is unrealistic to assume one model fits all students. But on the other hand, due to the sparsity in each student's log data, it is also impractical to create an independent set of model parameters for each individual student. In this work, we introduce a new concept named *latent student groups*, by assuming students are clustered and those belonging to the same group share the same set of model parameters. This captures the homogeneity of students' learning behavior patterns. At a population level, students are therefore modeled as a mixture over those latent student groups. This further recognizes the heterogeneity of their behaviors.

Based on the above model specifications, we define the set of group indicators as $\mathbf{c} = \{1, \ldots, k, \ldots, K\}$, such that c^u denotes the group membership for user u. In each latent student group k, we denote α^k as a $(N+1)\times N$ transition matrix, where N is the number of latent states, and β^k as a $N \times M$ emission matrix, where M is the number of learning actions. α^k and β^k serve as parameters of the Multinomial distributions for the transition and action-based emission probabilities in HMM. Accordingly, in group k, the time-based emission probability for latent state s_i is denoted as $ts^u_{w,t} \sim Beta(\Omega_1^{k,s_i}, \Omega_2^{k,s_i})$ if $y^u_{w,t} = s_i$, and the Poisson distribution for activity intensity in week w is denoted as $\mathbf{T}^u_w \sim Poisson(\mu^k)$. As a result, a latent student group k can be fully characterized by a set of model parameters $\theta^k = (\alpha^k, \beta^k, \mu^k, \Omega_1^k, \Omega_2^k)$. The probability of observing student u's learning action sequence under the latent student group k is thus computed as,

$$p(\mathbf{x}^u, \mathbf{y}^u, \mathbf{ts}^u, \mathbf{T}^u \mid c^u = k, \theta^{c^u}) \tag{1}$$
$$= p(\mathbf{x}^u, \mathbf{y}^u \mid \alpha^k, \beta^k)\, p(\mathbf{ts}^u \mid \Omega_1^k, \Omega_2^k) \prod_{w=1}^{W} p(\mathbf{T}^u_w \mid \mu^k)$$

To specify the generation of the model parameters in each latent student group, we introduce conjugate priors for them. In particular, we assume the transition and emission matrices are drawn from Dirichlet distributions as $\alpha^k \sim Dir(\eta)$ and $\beta^k \sim Dir(\lambda)$; and the rate parameter in Poisson distribution is drawn from a Gamma distribution as $\mu^k \sim Gamma(\xi, \varphi)$. However, because there is no conjugate prior for Beta distribution, we use exponential distributions as the prior, where $\Omega_1^k \sim Exp(\delta_1)$ and $\Omega_2^k \sim Exp(\delta_2)$.

The only thing left in our generative model for students' learning behavior is the specification of the number of latent student groups in a population of students. Instead of manually determining it, we treat c^u itself as a random variable, and assume each θ^k is drawn from a Dirichlet Process (DP) prior [5]. A Dirichlet Process $DP(G_0, \gamma)$ with a base distribution G_0 and a scaling parameter γ is a distribution over distributions. An important feature of DP is that draws from it often share common attributes which could naturally split data into groups. The numbers of unique draws, i.e., the number of clusters, varies with respect to the data and therefore is random, instead of being pre-specified.

By assuming the distribution of latent student groups in a given population of students follows DP, the joint probability of a student's learning action sequence can be further specified via a stick breaking process,

$$p(\mathbf{x}^u, \mathbf{y}^u, \mathbf{ts}^u, \mathbf{T}^u \mid \gamma, G_0) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^k} \tag{2}$$

where $\theta^k \sim G_0$, G_0 is the base distribution specifying the generation of θ^k, δ_{θ^k} is a distribution of $(\mathbf{x}^u, , \mathbf{y}^u, \mathbf{ts}^u, \mathbf{T}^u)$ concentrated at θ^k as defined in Eq (1), and $\pi = (\pi_k)_{k=1}^{\infty} \sim Stick(\gamma)$. In particular, the stick-breaking process $Stick(\gamma)$ for π is defined as: $\pi_k^{'} \sim Beta(1, \gamma)$, $\pi_k = \pi_k^{'} \prod_{t=1}^{k-1}(1 - \pi_t^{'})$.

In this resulting generative model, each latent student group possesses a unique configuration of the transition patterns and temporal distributions of student learning activities. This captures the diversity of students' learning intents and how they perform different actions to realize those intents. This corresponds to the principles in the theory of learning styles [18]: learners have their preferred ways of gathering, processing, interpreting, organizing and analyzing information. Therefore, we name our model as Latent Learning Style Model, or L^{2}S in short. Using the language of graphical model representation, our L^{2}S model is summarized in Figure 3. In the next section, we will describe how to perform statistical inference in L^{2}S to identify the latent student groups

and corresponding group configurations from students' observed learning activities.

4.2 Posterior Inference

To apply L^{2}S for clustering students into latent groups and identifying the corresponding group configurations, we need to infer the posterior distribution of 1) group-wise model parameters $\{\theta^k\}_{k=1}^{\infty}$ which capture the homogeneity of students associated with each latent student group: the transition and emission probabilities (α^k, β^k) within each latent user group k characterize the latent learning intentions, Ω^k and μ^k represent students' activities distribution over time and corresponding intensities. 2) the latent group membership c^u; and 3) the latent learning state sequence $\mathbf{y}^u$ of student u that represents his/her learning intent in each action. As conjugate priors are imposed over most of the model parameters, collapsed Gibbs sampling becomes a promising choice. Although the Beta distribution does not have a conjugate prior, auxiliary variable based Gibbs sampling scheme proposed in [17] still provides us an efficient inference solution under the employed DP prior.

To facilitate the description of our sampling scheme, we assume that at a particular step in sampling c^u for student u, there are in total K active latent student groups (i.e., groups that associated with at least one student, excluding the current student u), and by permuting the indices, we can index them from 1 to K. In order to apply Gibbs sampling, we need to derive the conditional distribution of each latent variable of interest. In the following, we provide the detailed derivations of those conditional distributions.

• **Sampling c^u:** Given a student u's learning action sequence $\mathbf{x}^u$ and latent state sequence $\mathbf{y}^u$, to compute the likelihood of sampling a new latent student group for these two sequence, we need to take integral of the group configuration θ with respect to its base distribution G_0. Thanks to the conjugacy between multinomial-Dirichlet and Poisson-Gamma distributions introduced in G_0, we can analytically integrate out α, β and μ in the likelihood computation. But as there is no conjugate prior for Beta distribution, we introduce a set of auxiliary random variables of size e [17] to compute the corresponding likelihood from the time-based emission probabilities. In each auxiliary student group, we sample the shape parameters Ω of Beta distribution from the exponential distributions. Denoting the number of students in group k as n_k^{-u} (excluding the current student u), the posterior distribution of c^u can be estimated by,

$$p(c^u = k|\mathbf{x}^u, \mathbf{y}^u, \mathbf{ts}^u, \mathbf{T}^u, \theta^k, G_0, \gamma) \tag{3}$$
$$\propto \begin{cases} n_k^{-u} \prod_{w=1}^{W} p(\mathbf{y}_w^u, \mathbf{x}_w^u, \mathbf{T}_w^u, \mathbf{ts}_w^u|\theta^k) & \text{for } 1 \le k \le K, \\ \frac{\gamma}{e} \int \prod_{w=1}^{W} p(\mathbf{y}_w^u, \mathbf{x}_w^u, \mathbf{T}_w^u, \mathbf{ts}_w^u|\theta^k) p(\theta^k|G_0) d\theta^k & \text{for } K < k \le K+e \end{cases}$$

When an auxiliary group is chosen for c^u, we will sample α^k, β^k and μ^k from their corresponding posterior distribution based on this observation, and append the resulting θ to existing latent student groups $\{\theta^k\}_{k=1}^{K}$ as one extra active latent group.

Because of the introduction of auxiliary variables $\{\Omega^k\}_{k=1}^{e}$, the number of sampled auxiliary variables affects the accuracy of the posterior distribution in Eq (3). To avoid sampling bias, we will draw a new set of auxiliary variables every time when sampling c^u.

• **Sampling $\mathbf{y}^u$:** Given the latent student group assignment of student u, the latent state of each observed action can be estimated from their posterior distribution calculated by the Forward-Backward algorithm. We compute the forward probability $\Phi(y_{w,t}^u)$ of each latent state $y_{w,t}^u$ as $\Phi(y_{w,t}^u) = p(ts_{w,t}^u|y_{w,t}^u)p(x_{w,t}^u|y_{w,t}^u)$ $\sum_{y_{w,t-1}^u} \Phi(y_{w,t-1}^u)p(y_{w,t}^u|y_{w,t-1}^u)$ and the corresponding backward probability $\Psi(y_{w,t}^u)$ as $\Psi(y_{w,t}^u) = \sum_{y_{w,t+1}^u} \Psi(y_{w,t+1}^u)p(x_{w,t+1}^u|y_{w,t+1}^u)$ $p(y_{w,t+1}^u|y_{w,t}^u)p(ts_{w,t+1}^u|y_{w,t+1}^u)$.

As a result, each latent state $y_{w,t}^u$ can be readily sampled from,

$$p(y_{w,t}^u) = p(y_{w,t}^u|\mathbf{x}^u, \mathbf{ts}^u) \sim Multi\Big(\frac{\Phi(y_{w,t}^u)\Psi(y_{w,t}^u)}{p(\mathbf{x}^u)}\Big) \tag{4}$$

where $p(\mathbf{x}^u)$ represents the likelihood function conditioned on the current transition and emission matrices α^k and β^k of group k that u is assigned. Once the student-level latent variables are sampled, we can perform the group-level model parameter inference as follows.

• **Sampling $\{\alpha^k\}_{k=1}^K$, $\{\beta^k\}_{k=1}^K$ and $\{\mu^k\}_{k=1}^K$:** Due to the conjugacy between Multinomial-Dirichlet and Poisson-Gamma distributions, the posterior distributions for these group-level parameters can be efficiently calculated. In particular, we have,

$$p(\alpha_i^k \mid \{\mathbf{y}^u\}_{u\in c_{mem}^k}, \eta) = \frac{\Gamma(N\eta + \sum_{j=1}^N n_{ij}^k)}{\sum_{j=1}^N \Gamma(\eta + n_{ij}^k)} \prod_{j=1}^{N} (\alpha_{ij}^k)^{\eta + n_{ij}^k - 1}$$

$$p(\beta_i^k \mid \{\mathbf{x}^u, \mathbf{y}^u\}_{u\in c_{mem}^k}, \lambda) = \frac{\Gamma(M\lambda + \sum_{o=1}^M m_{io}^k)}{\sum_{o=1}^M \Gamma(\lambda + m_{io}^k)} \prod_{o=1}^{M} (\beta_{io}^k)^{\lambda + m_{io}^k - 1}$$

$$p(\mu^k|\varphi^*, \xi^*) = Gamma(\varphi + \sum_{u\in c_{mem}^k} \sum_{w=1}^{W} \mathbf{T}_w^u, \frac{\xi}{1 + \xi|c_{mem}^k|})$$

where c_{mem}^k is the set of students assigned to group k, n_{ij} is the number of times that the latent state transits from s_i to s_j within group k, m_{io}^k is the number of times action o is associated with latent state s_i, φ^* and ξ^* are the parameters of posterior Gamma distribution where $\varphi^* = \varphi + \sum_{u\in c_{mem}^k} \sum_{w=1}^{W} \mathbf{T}_w^u$ and $\xi^* = \frac{\xi}{1+\xi|c_{mem}^k|}$.

• **Estimating $\{\Omega^k\}_{k=1}^K$:** Since there is no conjugate prior for Beta distribution, we do not have an analytical form of the shape parameters' posterior distribution. In this work, we appeal to the stochastic Expectation Maximization (EM) method to estimate these two non-negative shape parameters of Beta distribution.

Specifically, assume there are K active student groups after the sampling of $\{c^u\}_{u\in U}$, and denote $\mathbf{ts}^{k,s_i} = (ts_1^{k,s_i}, ..., ts_D^{k,s_i})$ is a set of timestamps of learning actions associated with the latent state s_i in group k. The complete-data log-likelihood over the timestamps can be computed as,

$$\ln L(\Omega_1^{k,s_i}, \Omega_2^{k,s_i}; \mathbf{ts}^{k,s_i}, \delta_1, \delta_2) = \tag{5}$$
$$\sum_{j=1}^{D} \Big[\ln\Big(p(ts_j^{k,s_i}; \Omega_1^{k,s_i}, \Omega_2^{k,s_i})\Big) + \ln\Big(p(\Omega_1^{k,s_i}; \delta_1)\Big) + \ln\Big(p(\Omega_2^{k,s_i}; \delta_2)\Big) \Big]$$

By taking partial derivation of Eq (5) with respect to the two shape parameters $(\Omega_1^{k,s_i}, \Omega_2^{k,s_i})$, the gradients can be computed as,

$$\frac{\partial \ln L(\cdot)}{\partial \Omega_1^{k,s_i}} = \sum_{j=1}^{D} \ln(ts_j^{k,s_i}) - D\frac{\partial \ln(B(\Omega_1^{k,s_i}, \Omega_2^{k,s_i}))}{\partial \Omega_1^{k,s_i}} - D\delta_1 \tag{6}$$

$$\frac{\partial \ln L(\cdot)}{\partial \Omega_2^{k,s_i}} = \sum_{j=1}^{D} \ln(1 - ts_j^{k,s_i}) - D\frac{\partial \ln(B(\Omega_1^{k,s_i}, \Omega_2^{k,s_i}))}{\partial \Omega_2^{k,s_i}} - D\delta_2 \tag{7}$$

The optimal setting of $(\Omega_1^{k,s_i}, \Omega_2^{k,s_i})$ can be obtained by setting the partial derivatives to zero. In the resulting updating formula, it is easy to verify that the first part of the right-hand side of Eq (6) and (7) corresponds to the geometric mean of the observed timestamps. We define $\hat{G}_{\mathbf{ts}^{k,s_i}} = \prod_{j=1}^{D}(ts_j^{k,s_i})^{\frac{1}{D}}$ and $\hat{G}_{\mathbf{1-ts}^{k,s_i}} = \prod_{j=1}^{D}(1-ts_j^{k,s_i})^{\frac{1}{D}}$. The gradient of log Beta function is $\frac{\partial \ln(B(\Omega_1^{k,s_i},\Omega_2^{k,s_i}))}{\partial \Omega_2^{k,s_i}} = \Psi(\Omega_1^{k,s_i}) - \Psi(\Omega_1^{k,s_i} + \Omega_2^{k,s_i})$, where $\Psi(\Omega_1^{k,s_i}) = \frac{\partial \ln \Gamma(\Omega_1^{k,s_i})}{\partial \Omega_1^{k,s_i}}$. Because the logarithmic approximation to the digamma function is $\Psi(\alpha) \approx \ln(\alpha - \frac{1}{2})$ [7], two shape parameters of Beta distribution can be approximately estimated as follows,

$$\hat{\Omega}_1^{k,s_i} \approx \frac{1}{2} + \frac{e^{\delta_2}\hat{G}_{\mathbf{ts}^{k,s}}}{2(e^{\delta_1+\delta_2} - e^{\delta_2}\hat{G}_{\mathbf{ts}^{k,s}} - e^{\delta_1}\hat{G}_{\mathbf{1-ts}^{k,s}})}$$

$$\hat{\Omega}_2^{k,s_i} \approx \frac{1}{2} + \frac{e^{\delta_1}\hat{G}_{\mathbf{1-ts}^{k,s}}}{2(e^{\delta_1+\delta_2} - e^{\delta_2}\hat{G}_{\mathbf{ts}^{k,s}} - e^{\delta_1}\hat{G}_{\mathbf{1-ts}^{k,s}})}$$

Putting the above sampling procedures together, the posterior inference for the proposed L^{2}S model can be performed by alternating the sampling steps for $\{c^u\}_{u\in U}$, $\{\mathbf{y}^u\}_{u\in U}$ and $\{\theta^k\}_{k=1}^{\infty}$ iteratively. The sampling of $\{c^u\}_{u\in U}$ and $\{\mathbf{y}^u\}_{u\in U}$ is performed once for every student, and the $\{\theta^k\}_{k=1}^{\infty}$ is performed once for the whole collection. Because the nearby samples from a Gibbs sampling chain are usually correlated, we only kept the posterior samples in every five iterations, i.e., thinning the sampling chain. Besides, samples from the burn-in period (i.e., the beginning of the sampling chain) are discarded as they may not accurately represent the desired distribution (in our experiment, we discarded the first 20% samples).

5 EXPERIMENTAL RESULTS

In this section, we evaluate the proposed L^{2}S model on the student learning activity data extracted in CS101 and Statistical Learning courses, which are studied in Section 3. We first qualitatively demonstrate the latent student groups discovered by the L^{2}S model from students' behavior sequences, i.e., the distribution of latent states, and transition and emission probabilities in the identified groups. Then we perform a series of quantitative evaluations to validate the effectiveness of our L^{2}S model against several state-of-the-art models in dropout prediction, certificate prediction, course satisfaction and demographics-based clustering.

5.1 Experiment Settings

We performed the same pre-processing on the student learning activity data extracted from the two courses as we described in Section 3. Our final evaluation dataset consists of 3078 and 8090 students in course CS101 and Stat respectively.

5.1.1 Baselines. We include several popularly used clustering and classification algorithms in student behavior analysis as our baselines for comparison.

Support Vector Machine (SVM). SVM has been applied for course completion and dropout prediction tasks in [12, 16, 19]. Following their settings, we used a linear SVM classifier fed with bigram action features as a baseline.

Vanilla RNN. Vanilla RNN is a sequence prediction model with recurrent structures to capture the temporal dynamics in sequential data. Recent works [16, 23] have applied vanilla RNN in dropout prediction. We kept the same network and parameters as employed in [16] when implementing this baseline.

K-Means. K-Means is a widely used unsupervised clustering algorithm. We used BIC criterion to determine the optimal number of clusters in this baseline.

K-Means+HMM. This baseline combines HMM model with K-Means to perform sequence clustering. In particular, The K-Means algorithm operates on the likelihood computed by a set of HMMs to cluster the sequences. Different from our L^{2}S model, the HMM model in this baseline only includes the action-based emission probabilities. Therefore, the temporal distribution and intensity of learning actions are not modeled in this baseline.

Basically, there are two kinds of feature representations of student sequential learning activities employed in these baselines. K-Means+HMM, just as our L^{2}S model, directly models the input action sequence. In the other baselines, i.e., SVM, Vanilla RNN and K-Means, feature vectors for each student are constructed based on the frequency of first-order transitions between consecutive actions (referred as bigrams). We chose bigram as the unit to construct the feature vector because the two HMM-based models impose a first-order transition assumption over the latent states, it is meaningful to compare the effectiveness between action-based transition and latent state based transition for behavior modeling.

5.1.2 Evaluation criterion. As our L^{2}S model works in an unsupervised manner to cluster students into latent student groups, we compare all the algorithms based on metrics for clustering evaluations. We should note that although the SVM and Vanilla RNN baselines are supervised classification methods, we treat their predicted class labels as cluster labels in our evaluation. Specificially, we choose the following three metrics that are commonly used to measure clustering results.

NMI. Normalized mutual information (NMI) is an information theoretic measure about how well the computed clusters and the ground-truth clusters predict one another, normalized by the amount of information inherent in the two clustering systems. A higher NMI means a better clustering quality.

Entropy. It uses external information such as class labels to evaluate clustering quality. A lower entropy indicates a better clustering quality.

F1 score. It measures how close a clustering result is to the ground-truth cluster labels. F1 score is evaluated by computing the pairwise precision and recall in all clusters.

5.1.3 Evaluation Settings. In L^{2}S model, we fixed the hyper-parameters as $\gamma = 0.01$, $\lambda = 1.0$, $\eta = 1.0$, $\varphi = 0.01$, $\xi = 0.01$, $\delta_1 = 0.1$, $\delta_2 = 0.1$ and $N = 11$ in the following experiments. We set the total number of sampling iterations to 1000. In addition, based on the return time analysis shown in Figure 2c, and the method proposed in [14], we set the timeout threshold to 2 hours to define the *idle* actions in both courses.

As SVM and vanilla RNN baselines are supervised methods, we split the whole dataset into two parts: the first 80% students for training and the rest 20% for testing, and perform 5-fold cross-validation on them. To ensure the consistence of the comparison results, the other three unsupervised algorithms, i.e., L^{2}S, K-Means and K-Means+HMM, are applied on the same train/test separation.

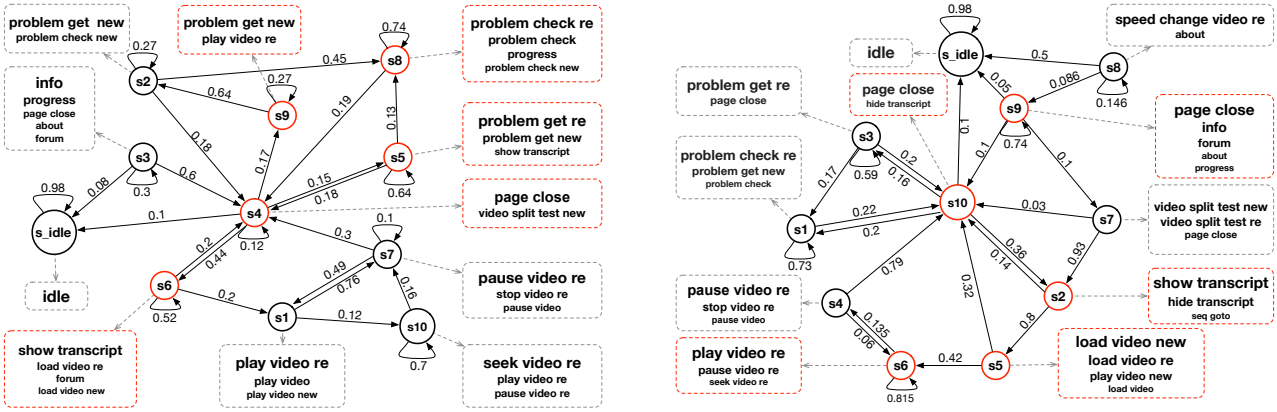


Figure 4: Two distinctive latent student groups and their latent transitions in course CS101.

5.2 Qualitative Evaluation

We first demonstrate the latent student groups identified by our L^{2}S model in the CS101 course. Similar results were also obtained in the Stat course, but due to space limit we cannot show them. Although the group assignments in individual students may change during the sampling iterations, the internal structure of each latent student group is rather stable in the later stage of sampling. We used the group assignments obtained from the last iteration of Gibbs sampling, and analyzed the learned transition and emission probabilities in each group. Multiple latent student groups were created for the course CS101, we selected the two most distinctive groups for illustration as described below.

In each group, we adopted the Viterbi algorithm to estimate the most probable latent state assignments in each student's action sequence based on the learned model. Then we accumulated the occurrences of each latent state among the group members and highlighted the top 5 of them as the dominating latent states. For each latent state, we also extracted the activities from the corresponding emission probabilities. To increase visibility, we selected the transitions with probability larger than 0.05 and discarded those smaller than 0.05. We demonstrate the selected transitions and emissions in Figure 4. In the figure, the higher the emission probability is, the larger font is used for the action displayed in the dashed box.

From the transition graph on the left of Figure 4, we can find many interesting learning patterns identified in this latent student group. The selected most frequent actions associated with the highlighted latent states (the red nodes) are mostly related to "video" and "problem", but rarely related to the course information pages (such as "info", "progress"). We can also discover a strong transition from "video" related latent states to "problem" related states, while for the "problem" related states, they commonly have a higher self-transition probability than that to other latent states. These identified behavior patterns have high similarity to the concept of accommodator developed in the learning style theory. By looking into the final performance of students belonging to this latent group, we found most of them completed the course with certificates.

The transition graph on the right of Figure 4 is extracted from another latent student group that exhibits quite different behavior patterns. The dominating latent states in this group are mostly associated with "video" related activities, and only two latent states (denoted as $s1$ and $s3$ in the figure) are associated with "problem" related activities. Moreover, strong transitions exist among three latent states (denoted as $s2$, $s5$ and $s6$ in the figure) which are strongly associated with "video" related activities, while two "problem" related latent states are less likely to transit to these "video" related latent states. The learning behavior patterns identified in this group are quite similar to the assimilator concept described in the learning style theory. Accordingly, we found most students in this group did not obtain the certificate in the end.

5.3 Quantitative Evaluation

In this section, we conduct a suite of quantitative evaluations from different perspectives to verify the effectiveness of our solution in predictive modeling of students' learning behaviors.

•**Dropout prediction.** Due to the low retention rate in MOOCs, predicting student dropout is an important task in MOOCs data mining. The dropout definition in online learning environment varies [9, 12, 16, 23] due to the distinct goals of capturing the student status in different contexts. Basically, there are two popularly used dropout definitions. The first one is whether a student completes the course and obtains the certificate, which focuses more on the learning outcome. The second one is whether a student will participate the course in the near further, which focuses more on student short-term engagement. The first definition is commonly used in both classroom and online education environment, while the second definition is arguably more important in online learning environment, because of the self-regulated learning and lack of face-to-face supervision from the teacher. In this experiment, we focus on the short-term dropout prediction.

It is important to monitor the change of the students' status during the course and identify students who have higher probability of dropout as early as possible, so that personalized assistance and guidance or early interventions can be applied. To evaluate different algorithms' predictive power in this task, we label a student as dropout in week t if he/she has no logged activity in week t. This will be considered as the ground-truth label in supervised classifier training (i.e., SVM and Vanilla RNN) and clustering result evaluation. All historical data from a student before week t is used for model learning; and for each week's prediction, a new classifier or clustering model is trained independently in each baseline.

Figure 5 presents the mean and standard deviation of the prediction performance of different methods in these two courses. Overall,

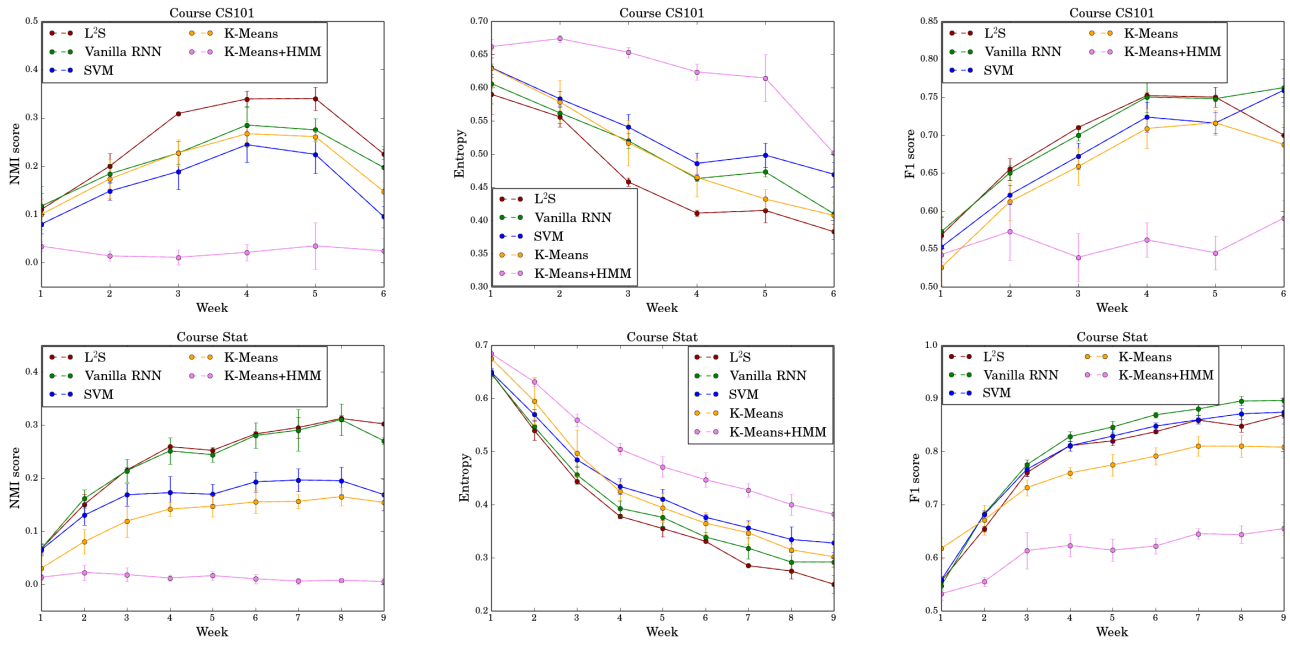


Figure 5: Performance comparison of dropout prediction by different methods in course CS101 and Stat.

L^2S achieved the best performance against all baselines in terms of NMI and entropy measurements, except the F1 comparison in Stat course. When we looked into the detailed clustering results from the K-Means baseline, we found almost all of the students were clustered into one big cluster every week. This result is useless in predicting students' future behavior. In particular, even compared with two supervised models, i.e., SVM and RNN, our L^2S model achieved consistent improvement over them in NMI and entropy metrics. As we mentioned before, these two supervised methods only exploit action-based transitions, which cannot capture the important temporal distribution and relatedness among those actions. Although these two methods can leverage direct supervision from historical data, they still suffer from insufficient representation of students' learning behaviors, and therefore are inept to accurately predict their future behaviors.

•Certificate acquisition prediction. As we mentioned above, certificate acquisition prediction is another important topic in MOOCs data mining and has been explored recently [3, 12, 16, 19]. In our experiment, the predictions from each method are made after the Midterm, one week and two weeks before the end of the courses (denoted as "Mid", "-1w" and "-2w" in Table 2 accordingly). Because the completion rate is very low in both courses (less than 10%), it leads to a very imbalanced label distribution. In this situation, F1 metric is not sensitive in recognizing the quality of different clustering results, as even if one puts all instances into one cluster, the F1 score will still be reasonably high. Because NMI is normalized against the ground-truth label distribution, it emphasizes the clustering of minor clusters. In this experiment, we only use NMI to compare different algorithms' clustering results.

As shown in Table 2, in both courses L^2S outperformed all the baselines except the prediction made in Stat course one week before the end of it. This result reveals our L^2S model has consistent predictive power as in dropout prediction. We can observe that although vanilla RNN performed better than other baselines in dropout prediction, it performed much worse than L^2S and SVM in this task. When we looked into its detailed prediction results, we found that almost all students were predicted as failing to get the certificates, as the prior distribution of training labels strongly favor this prediction.

Table 2: NMI score over certificate acquisition prediction.

Models	CS101			Stat		
	Mid	-2w	-1w	Mid	-2w	-1w
K-Means	0.194	0.293	0.355	0.128	0.179	0.192
K-Means+HMM	0.076	0.105	0.117	0.027	0.015	0.043
SVM	0.191	0.341	0.389	0.243	0.318	**0.357**
Vanilla RNN	0.235	0.308	0.057	0.036	0.000	0.000
L^2S	**0.263**	**0.369**	**0.418**	**0.246**	**0.320**	0.341

• Course satisfaction prediction. In our dataset, we have recorded students' responses of the post-course surveys in both courses, and some questions asked about the students' satisfaction and overall expectation of the course. We selected three typical questions to categorize students into multiple classes based on their responses. Specifically, the first question we chose is "Q1.1 How good or bad was your overall experience with the course?"; the second is "Q1.4 How satisfied are you with the amount you learned in the course?"; and the third is "Q1.7 How much did you learn in the course relative to your expectations?" The provided options for these questions are 1) extremely bad; 2) very bad; 3) mostly bad; 4) neutral; 5) mostly Good; 6) very good; 7) extremely good. Because the survey was only collected after a student finished the course, we only evaluated the algorithms on this set of students and we used their behavior sequence in the whole course span for clustering.

We categorize options 1, 2 and 3 as "negative", option 4 as "neutral", and the rest as "positive". In this case, students who have responded the survey will be classified into one of the three classes. As we mentioned in Section 3, the dropout rate of course Stat is as

Table 3: Survey-based clustering performance comparison.

Question	K-Means		K-Means+HMM		L^2S	
	F1	NMI	F1	NMI	F1	NMI
Q1.1	0.667	0.013	0.689	0.012	**0.878**	**0.025**
Q1.4	0.642	0.013	0.661	0.006	**0.692**	**0.023**
Q1.7	0.437	**0.025**	0.453	0.014	**0.527**	0.011

Table 4: Information gain comparison between three clustering algorithms over different demographic attributes.

Models	CS101			Stat		
	Gender	Age	Edu	Gender	Age	Edu
K-Means	0.019	0.049	0.064	0.046	0.033	**0.039**
K-Means+HMM	0.007	0.019	0.017	0.005	0.003	0.005
L^2S	**0.032**	**0.058**	**0.066**	**0.051**	**0.095**	0.025

high as 90%, and we found less than 100 out of 8090 students had finished post-course surveys, while for the course CS101, 711 out of 3078 students answered the post-course survey, so we performed this experiment only on the CS101 course, and compared with K-Means and K-Means+HMM algorithms. Results shown in Table 3 demonstrate that our L^2S model can better differentiate students with various expectations of the course or themselves purely from their behavior history in the course.

•**Demographics-based clustering.** Our datasets also include demographic information of the enrolled students. Due to the open nature of MOOCs, students' demographics are extremely diverse. We select gender, age and education level to analyze the clustering results form L^2S model, and compared with the results from K-Means and K-Means+HMM baselines. The metric of information gain is used for evaluating the expected reduction in entropy caused by the clustering results from an algorithm versus the prior distribution under one of three demographic attributes. Table 4 presents the results on two datasets, and our L^2S model achieved much better information gain than the second best method, K-Means, on all three demographic attributes, except the education level in the Stat course. This further verifies the utility of the latent user groups identified by our L^2S model, which infers students' background purely from their learning activity patterns.

6 CONCLUSION

In this work, we build a generative model to model students' sequential learning activities. The concept of latent students group is introduced to group students with similar underlying behavior models, which reflect the homogeneity of their learning activities and intents. Compared with several popularly used analytic models for students' learning behaviors, the identified latent student groups in our proposed model can better differentiate students with and without course certificates, students who have higher risk to dropout in a short term, and predict students' satisfaction about the course.

In our current solution, students and courses are independently modeled. It is beneficial to consider students' behaviors across multiple courses, and students' influence on each other. For example, students who interact with each other more often in the course content might behave more similarly to each other. Explicit modeling of such observations will help our model better capture the dynamics of students' behaviors during the learning process. In addition, the available supervised signals, such as their dropout labels and quiz grades, can also be leveraged to enhance our model's modeling quality of their latent learning behavior and intents.

ACKNOWLEDGMENTS

This paper is partially supported by the National Science Foundation under grant IIS-1553568, National Natural Science Foundation of China under grant No.61232002, and the Ministry of Science and Technology of China, National Key Research and Development Program (Project Number: 2016YFB1000700).

REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. 2014. Engaging with massive online courses. In *WWW'14*. 687–698.
[2] L. Barnard, W. Y. Lan, Y. M. To, V. O. Paton, and S. Lai. 2009. Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education* 12, 1 (2009), 1–6.
[3] C. A. Coleman, D. T. Seaton, and I. Chuang. 2015. Probabilistic use cases: discovering behavioral pattern for predicting certification. In *L@S'15*. 141–148.
[4] R. S. Dunn, K. J. Dunn, and G. E. Price. 1989. *Learning style inventory (LSI)*. Price Systems, Incorporated (PO Box 1818, Lawrence 66044).
[5] T. S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics* (1973), 209–230.
[6] D. Glance, M. Forsey, and M. Riley. 2013. The pedagogical foundations of massive open online courses. *First Monday* 18, 5 (2013).
[7] R. Gnanadesikan, R. S. Pinkham, and L. P. Hughes. 1967. Maximum likelihood Estimation of the Parameters of the Beta Distribution from Smallest Order Statistics. *Technometrics* 9, 4 (1967), 607–620.
[8] P. J. Guo and K. Reinecke. 2014. Demographic differences in how students navigate through MOOCs. In *L@S'14*. 21–30.
[9] Z. Jiang, Y. Zhang, and X. Li. 2015. Learing behavior analysis and prediction based on mooc data. *Journal of Computer Research and Development* (2015).
[10] K. Jordan. 2014. Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning* 15, 1 (2014).
[11] R. F. Kizilcec, C. Piech, and E. Schneider. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *LAK'13*. 170–179.
[12] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. 2014. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In *EMNLP 2014*. 60–65.
[13] R. Kop, F. Hélène, and J. S. F. Mak. 2011. A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses. *The International Review Of Research In Open And Distributed Learning* 12, 7 (2011), 74–93.
[14] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. 2011. Identifying task-based sessions in search engine query logs. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 277–286.
[15] F. G. Martin. 2012. Will massive open online courses change how we teach? *Commun. ACM* 55, 8 (2012), 26–28.
[16] F. Mi and D. Yeung. 2015. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. In *ICDM Workshops 2015*. 256–263.
[17] R. M. Neal. 2000. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphic statistics* 9, 2 (2000), 249–265.
[18] H. Pashler, M. McDaniel, D. Rohrer, and R. Bjork. 2008. Learning styles concepts and evidence. *Psychological science in the public interest* 9, 3 (2008), 105–119.
[19] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue. 2016. Modeling and predicting learning behavior in MOOCs. In *WSDM'16*. 93–102.
[20] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. 2014. Learning latent engagement patterns of students in online courses. In *AAAI'14*. 1272–1278.
[21] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard. 2014. Who does what in a massive open online courses? *Commun. ACM* 57, 4 (2014), 58–65.
[22] M. D. Svinicki and N. M. Dixon. 1987. The Kolb Model Modified for Classroom Activities. *College Teaching* 35, 2 (1987), 142–146.
[23] F. Wang and L. Chen. 2016. A Nonlinear State Space Model for Identifying At-Risk Students in Open Online Courses. In *EDM'16*. 527–532.
[24] M. Wen and C. P. Rose. 2014. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *CIKM'14*. 1983–1986.
[25] J. Wilkowski, A. Deutsch, and D. M. Russell. 2014. Student skill and goal achievement in the mapping with google MOOC. In *L@S'14*. 3–10.
[26] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*.