Intention Inference via MixHMMs in Clickstream Data

Mason Victors, Nino Paul Batanay, Mary Ann Lim, Gerardo "Rarry" Abatol

February 6, 2015

1 Introduction

In this grant project, we have studied the problem of determining intention of visitors to the ToysRUs website, using Mixture Hidden Markov Models (MixHMMs). In section 1, we provide a brief introduction to the models in order to develop some intuition surrounding them and their applicable uses. In section 2, we detail the training undergone in the beginning of this grant project. Section 3 includes a discussion of the software development involved in this project. In section 5, we give a summary of the data available for use, along with details on the final datasets used. We show our findings and results in section 5.

2 MixHMMs

A Hidden Markov Model (HMM) describes a process by which time series data is generated. When using an HMM, there are two time series involved: a latent time series (called the 'hidden states') and an observed time series (called the observation sequence). The generative process is as follows: an observation o_t is generated from a probability distribution dependent on the hidden state s_t at time t, and the process then transitions to a hidden state s_{t+1} according to a categorical distribution dependent only on the hidden state s_t. Then this process repeats, creating the observation sequence time series. The parameters defining an HMM include an initial state distribution (a categorical distribution π over the hidden states), a state transition matrix (a row-stochastic matrix A, where a_{ij} denotes the probability of transitioning from hidden state i to hidden state j) and a set of emission distributions (one for each state, used to generate each observation o_t).

A Mixture HMM (MixHMM) is yet another model describing a process by which time series data is generated, in which there are K different HMMs. An observation sequence is generated by first selecting one of the K HMMs from a categorical distribution, and then using the selected HMM to generate a sequence of data.

MixHMMs appear to be rarely used, as there is little by the way of literature describing them. When working with clickstream data for a website, they have a natural application. As a user 'clicks' to navigate through a website, he is naturally creating a time-series of

observations. The purpose of each *click* might be described by a hidden state, and the intention of each *visit* might be described by the transition through the hidden states.

3 Training

At the start of this grant project, there was much training required. We began by a deep dive into the Expectation Maximization (EM) algorithm, along with its underlying theory. This was necessary to build up the understanding of the chief method of searching for a maximum likelihood estimator when working in the presence of latent (hidden) variables, as is the case with HMMs and MixHMMs.

We then studied the application of EM to standard HMMs (those emitting categorical observations), and together developed extensions of this algorithm for alternative emission distributions (Poisson, exponential, and normal).

Finally, we thoroughly studied the algorithms pertaining to MixHMMs, as code for this was not publicly available at the time of this project.

4 Development

As mentioned above, at the time of this project, there was no readily available public code for working with MixHMMs. As a result, we developed this ourselves, extending a publicly available HMM Python package known as *hmmlearn* (see https://github.com/hmmlearn/hmmlearn). Our extension included many improvements:

- New observation distributions
  - Poisson distribution
  - Exponential distribution
  - Multinomial/Exponential distribution (multivariate)
- MixHMM code
- Parallelization for both HMM and MixHMM code
- Options for memory safe usage of HMM and MixHMM training code

This code is publicly available at https://github.com/mvictor212/hmmlearn.

5 Data

The data we had available to use on this project was ToysRUs clickstream data. The data includes the following data fields (among others):

- Visitor ID
- Visit ID
- Click ID
- Click URL
- Click Pagetype
- Click Page Tier
- Click Timestamp

From this data available in our AWS Redshift databases, we could represent each visit as a sequence of clicks. There were many options for choosing what each click's observation would be. We ultimately settled on the following four representations:

1. Pagetype
    - Assumed to be drawn from a categorical distribution
2. Pagetype-Tier
    - Concatenation of Click Pagetype and Click Page Tier
    - Assumed to be drawn from a categorical distribution
3. Elapsed Time
    - Derived from the ordered values of Click Timestamp
    - Assumed to be drawn from an exponential distribution
4. (Pagetype, Elapsed Time)
    - Multivariate observation
    - Assumed to be drawn from a categorical distribution and an exponential distribution

We defined 18 distinct Pagetypes, and 125 distinct Pagetype-Tiers. The data available consisted of over 2 billion clicks and more than 200 million visits. Due to time and resource constraints, we subsampled this data and trained our models on 500,000 visits, which included over 7 million clicks.

A few page types require some explanation. The 'Product Detail' is a page displaying information about a product. This includes images, descriptions, reviews, pricing details, etc. The 'Null' page type often indicates an on-page click, where the user is not actually navigated to a new URL. These include things like enlarging a product view, reading the

product description/specifications, or navigating the reviews of a product. The 'Family' page type displays a list of products, with links directly to the products themselves. The 'Category' page type displays a list of families, with direct links to URLs having the 'Family' page type. Thus, a 'Category' is coarser than a 'Family'.

6 Results

For each possible formulation of the observation sequences, we trained a MixHMM for each pair of general parameters:

- 2 - 4 HMM components
- 2 - 8 hidden states

resulting in 21 MixHMMs for each observation sequence formulation. Each model took anywhere between 2 to 12 hours to train, due to limited computational resources and the massive size of the data (in a similar study, the training data only included 10,091 visits, and 126,348 clicks).

Below, we provide several examples of trained MixHMMs, as well as their interpretation. For sequences defined as (Pagetype, Elapsed Time), we provide an example of a 2-Component, 6-State MixHMM. Because our MixHMM has 2 components, we have a parameter $\vec{\omega}$ denoting the mixture weights over the 2 components. With our 6 states, we have a rates vector $\vec{\lambda}$, where each entry corresponds to the rate parameter of an exponential distribution unique to each hidden state. As the reciprocal of the rate parameter is the expected value of the exponential random variable (here referring to the elapsed time between clicks in seconds), we display this instead for ease of interpretation. Each HMM has an initial state distribution $\vec{\pi}$ denoting the probability of beginning a sequence in each state, as well as a transition matrix A, where row i denotes the transition distribution from state i to each of the other hidden states.

$\vec{\omega}$	Component 1	Component 2
	0.307	0.693

$\frac{1}{\vec{\lambda}}$	State 1	State 2	State 3	State 4	State 5	State 6
	6.24	32.92	380.08	30.83	30.62	30.55

B

Pagetype	State 1	State 2	State 3	State 4	State 5	State 6
Product Detail	0.043	0.138	0.266	0.298	0.002	0.182
Family	0.004	0.012	0.168	0.03	0	0.423
Registry	0	0.705	0.125	0.002	0	0.004
Null	0.641	0.005	0.117	0.14	0.037	0.018
Home Page	0	0.015	0.055	0.039	0.446	0.036
Category	0	0.002	0.021	0.006	0.006	0.178
Successful Search Results	0	0.01	0.055	0.217	0	0.003
Parametric Refinement	0	0.004	0.024	0	0	0.099
Checkout	0.002	0	0.021	0	0.275	0
Shopping Bag	0.002	0.021	0.03	0.034	0.157	0.001
In Store Pick Up	0.171	0	0.011	0.001	0	0
Store Locator	0.096	0	0.016	0.002	0.033	0
Failed Search Results	0	0.001	0.014	0.053	0	0.006
Micro Site	0.025	0.015	0.013	0	0.012	0.017
Shop	0	0.005	0.014	0.023	0.014	0.004
My Account	0	0.04	0.007	0	0.001	0
Wish List	0	0	0.004	0.002	0.001	0.014
Other	0.014	0.026	0.041	0.153	0.014	0.014

$\vec{\pi}_1$

State 1	State 2	State 3	State 4	State 5	State 6
0.052	0.107	0.063	0.319	0.225	0.236

$\vec{\pi}_2$

State 1	State 2	State 3	State 4	State 5	State 6
0.012	0.064	0.063	0.013	0.742	0.107

A_1

States	State 1	State 2	State 3	State 4	State 5	State 6
State 1	0.821	0.004	0.07	0.032	0.059	0.014
State 2	0.027	0.787	0.095	0.006	0.08	0.005
State 3	0.189	0.058	0.153	0.294	0.21	0.097
State 4	0.059	0.002	0.079	0.776	0.043	0.041
State 5	0.036	0.038	0.04	0.023	0.833	0.03
State 6	0.023	0.001	0.036	0.067	0.006	0.866

A_2

States	State 1	State 2	State 3	State 4	State 5	State 6
State 1	0.943	0.002	0.015	0.019	0.005	0.016
State 2	0.001	0.908	0.042	0.014	0	0.036
State 3	0.014	0.089	0.294	0.225	0.105	0.275
State 4	0.012	0.017	0.058	0.855	0	0.059
State 5	0.015	0.2	0.046	0.273	0.024	0.442
State 6	0.004	0.012	0.042	0.032	0	0.91

Interpreting a MixHMM can be difficult, but this model has a fairly clear story to tell. The first thing we can notice is that approximately one-third of all visits to ToysRUs follow the behavior of the first HMM component, and the remaining two-thirds of visits follow the second HMM component.

From $\vec{\lambda}$, we can see that state 1 generally has a short mean duration (approximately 6 seconds), state 3 has a long mean duration (approximately 6 minutes), and the remaining 4 states have mean durations around 30 seconds.

From B, we get a much clearer picture of the meaning behind each hidden state. State 1 is dominated by the 'Null' page type, with reasonable support also from 'In Store Pick Up' and 'Store Locator'. Together, these three observations explain 90.8% of all clicks from state 1. Clicks with a 'Null' page type are generally quick to load and can be quick to digest, resulting in the short duration time explained by the state's exponential distribution. State 2 is dominated by 'Registry' page types, with some small support of 'Product Detail' as well. State 3 is highly focused on 'Product Detail', 'Family', and 'Registry'. Considering that it has a mean duration of over 6 minutes, we might describe this as an in-depth view of a product or family of products. State 4 is also dominated by 'Product Detail', followed closely by 'Successful Search Results'. This has a mean duration of 31 seconds, so this can be thought of as both navigating to products via the search toolbar, as well as short views of product pages. State 5 is explained primarily by 'Home Page', 'Checkout', and 'Shopping Bag', which are all self explanatory. State 6 is composed primarily of 'Family', 'Product Detail', 'Category', and 'Parametric Refinement', suggesting that this hidden state refers to navigation to products using the site menus and refining of parameters (such as brand, age recommendations, ratings, etc.) as well as short views of products themselves. It's interesting that the model has been able to identify that visitors can either search the site using the search toolbar (state 4) or by using drop-down menus and parametric refinement (state 6).

Now that we have a basic grasp on the 6 uncovered hidden states, we should look at the two separate HMM components and their parameters. One of the most basic things we can look for is how visitors enter the site, and this is done by examining the initial state distributions ($\vec{\pi}_1$ and $\vec{\pi}_2$). We might expect that people generally enter the site via the home page (which is dominant in state 5), and this is certainly the case for the second HMM

component, where 74% of its visits start in state 5. But in the first HMM component, over 77% of visits are likely to begin in one of the other 5 states! This suggests that visitors who more closely follow the behavior of the first HMM component are more likely to enter directly to a product, family, or category page than via the home page, which generally happens when people search for a product or group of products using an external search engine such as *Google* or *Yahoo!*.

The transition matrices confirm this for us as well. In the second HMM component, we see that visits to a state 5 page rarely result in clicks to another state 5 page, but rather transfer into either state 6, 4 or 2. Thus, users who enter the site from the home page, rarely stay in state 5, but quickly transition elsewhere. Also, in component 2, state 6 is quite dominant, as users have a high probability (91%) of staying in state 6 once there, as well as high transition rates into state 6 from states 5 and 3. This can be confirmed by looking at the *steady-state* distributions of the underlying Markov chains, which we give below for both HMM components:

$\vec{s}_1$

State 1	State 2	State 3	State 4	State 5	State 6
0.22	0.069	0.065	0.202	0.251	0.193

$\vec{s}_2$

State 1	State 2	State 3	State 4	State 5	State 6
0.099	0.17	0.057	0.227	0.007	0.44

From this, we can conclude that for the second component, 44% of all clicks are within state 6, and state 4 has the second highest proportion of clicks, at 23%. Together, those two states make up 67% of all clicks for visits following the behavior of the second HMM component, whereas in the first HMM component, they only comprise less than 40% of all clicks, with a much larger share of clicks coming from states 1 and 5. Recalling that state 5 is dominated by pages labeled as either 'Home Page', 'Checkout', and 'Shopping Bag', and state 1 is dominated by short visits to 'Null' page types, we are now able to provide high-level descriptions of the two website visit behaviors uncovered by this MixHMM. The first HMM component describes the behavior of those visits which tend to begin by a search for a product from *outside* the site, then thoroughly research a product, and ultimately have a higher chance of making a purchase. The second component describes those visits beginning primarily at the home page, and spend their time browsing the various products, but are ultimately less likely to make a purchase during the visit.