PREDICTING STOPOUT IN MOOCS
Mining Behavioral Data

Kalyan Veeramachaneni
Colin Taylor
and
Una-May O’Reilly
unamay@csail.mit.edu

STU EPS Course: co-teacher Erik Hemberg

Evolutionary Processes and Systems

Shantou University,
China
Week 1:
• In person
• Active learning
  – Evolution of cooperation
  – Matting game to teach Sexual selection

Massive Online Open Courses
The Emergence of MOOCs

- internet being around us and available all the time
- Scalable platforms running on the cloud
  - Elastic, inexpensive
- the high cost of college education in the US
- the recognition of many in the workforce that they need to become life long learners
- MIT has a mission to contribute to global learning

Online/MOOC education offers a new way to teach and learn

The observations of learner behavior we gather differs between online/MOOC and brick-n-mortar courses

Online learning platforms allow us to track "every" student. This data is at a very fine grained level: every key stroke and mouse click with a timestamp and student ID, collected as platform runs
Big Data and MOOCs

<table>
<thead>
<tr>
<th>Data Category</th>
<th>Description</th>
<th>Data Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Aggregators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consolidated Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global Community</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Analysts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database experts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scripts archive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy experts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course DB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reformat Sql</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sql scripts</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evolutionary Processes and Systems

Week 1:
- In person
- Active learning
Weeks 2-10
- Private, online learning
Week 10:
- In person
- Project-based

Not for focus today but
- Active learning online?
- Blended learning

MoocDB Data Model

data organization to support many eyes on the data

Our Experience

- I learned first hand what everyone else was talking about
- MOOCs have lots of challenges ahead of them in design to become effective

For example,
- isolation
- Issue with persistence
  - ⅓ students stopped taking the online content,
  - but returned when we got back
- The question was arising all over
Most stopout occurs in first few weeks

Explanations

- window-shopping
  - never even tried to submit some sort of quiz answer or exercise answer.
- remaining set of students left, some didn’t intend to finish but they would go forward to some degree
- others really intended to finish but didn’t.

How can stopout be explained from the data?

- the answer is important:
  - Intellectual: it reflects the challenge of understanding how students learn
  - Practical: if predictors of stop-out could be determined,
    - we could intervene, personalize learning,
  - we could design courses better!

Agenda

- Background and motivation
- Intro to ALFA
  - Machine learning
  - MOOCDB data science commons project
- Solving Stopout Prediction
  - Define problem
- Some findings

Machine Learning

Training, Testing, Deployment
Scalable Machine Learning

- Probabilistic models: DBNs, HMMs, Copulas (EM, MCMC)
- Functional models: NNs, GP
- Time series models: Un/Supervised, Non/Parametric
- Matrix Methods
- Large Scale SVD
- Classification Algs: DTs, SVM, NN...

ALFA Platforms: Delphi, EC-Star, FlexGP, FCUBE

Data Handling
- Approximation techniques
- Sampling techniques
- Ensemble techniques

Cloud
GPU
HPC
VCN

AnyScale Learning for All

- Energy
- Medicine
- Education

Big Data Knowledge discovery
- Scalable Machine Learning

ALFA Group Introduction

MOOCDB data science commons

- organize
- interpret
- extract
- model
- visualize

Stopout Prediction

MOOCDB
FEATUREFACTORY
LABELME
FORUMS
MOOCVIZ

Started and being directed by ALFA co-lead Kalyan Veeramachaneni, PhD
The Prediction Problem

We can use students data during these weeks

Machine Learning often tacitly starts from the Matrix

Think and propose

Extract features

A feature is an attribute (or variable) that is useful or meaningful to your problem

But, the matrix must be formulated!

Feature engineering

Machine Learning

Curation of raw data

More than a year

A week
The Machine Learning Workflow

Organize

Extract & Aggregate

Interpret

Measure

Model

This is end to end notion of machine learning

ALFA Weekly Variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>dropout</td>
<td>Whether the student has dropped out or not</td>
</tr>
<tr>
<td>total errors</td>
<td>Total errors per error</td>
</tr>
<tr>
<td>number forum</td>
<td>Number of forum posts</td>
</tr>
<tr>
<td>number wild posts</td>
<td>Number of wild posts</td>
</tr>
<tr>
<td>average length forum</td>
<td>Average length of forum posts</td>
</tr>
<tr>
<td>number distinct prob</td>
<td>Number of distinct problems attempted</td>
</tr>
<tr>
<td>number orphaned</td>
<td>Number of orphaned</td>
</tr>
<tr>
<td>n2</td>
<td>Number of distinct correct problems</td>
</tr>
<tr>
<td>average number sub</td>
<td>Average number of submissions per problem (p / n)</td>
</tr>
<tr>
<td>observed event date</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
<tr>
<td>observed event time</td>
<td>Ratio of total time spent to number of distinct correct problems (p / n)</td>
</tr>
</tbody>
</table>

“Crowd”-sourced weekly variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>f_1</td>
<td>Number of forum responses</td>
</tr>
<tr>
<td>f_2</td>
<td>Average number of submissions (feature 9) as compared with other students that week as a percentile of the maximum average number of submissions that week</td>
</tr>
<tr>
<td>f_3</td>
<td>A student’s average number of submissions (feature 9) as a percent of the maximum average number of submissions that week</td>
</tr>
<tr>
<td>f_4</td>
<td>Number of the week’s homework problems answered correctly / number of that week’s homework problems</td>
</tr>
<tr>
<td>f_5</td>
<td>Difference in grade between current past grade and average of student’s past past grade</td>
</tr>
<tr>
<td>f_6</td>
<td>Number of the week’s lab problems answered correctly / number of that week’s lab problems</td>
</tr>
<tr>
<td>f_7</td>
<td>Difference in grade between current lab grade and average of student’s past lab grade</td>
</tr>
<tr>
<td>f_8</td>
<td>Number of correct submissions</td>
</tr>
<tr>
<td>f_9</td>
<td>Percentage of the total submissions that were correct (p / n)</td>
</tr>
<tr>
<td>f_10</td>
<td>Average problem submission time for each problem that was solved correctly / number of that week’s homework problems</td>
</tr>
</tbody>
</table>
Cohort Definitions

Active students are grouped
• Use forums: discussion generator
• Use wiki: content generator
• Use both: fully collaborative
• Use neither: passive collaborative

The Big Data Analytics Numbers

• 154,763 students registered in 6.002x Spring 2012
• 200+ Million events
  – 60 GB of raw click stream data
• 52000+ students in our study
  – 130 Million events
• 44,526 never used forum or wiki
• Models use 27 predictors with weekly values
  – 351 dimensions at max
• Predictors reference clickstream to consider
  – Time, performance on assessment components
    › homeworks, quizzes, lecture exercises
  – Time, use of resources
    › videos, tutorials, labs, etexts, ...
• 91 prediction problems for each of 4 cohorts

Predictions

Passive Learners Cohort

Question

Q. After only the first week, how accurately can we predict who will drop out by the final week?

Answer: Pretty well! Predictions for the end of the course, using just the first week, were robust: We attained AUC accuracies of ~0.7
Q. After only the first week, how accurately can we predict who will drop out by the final week?

Q2

Q. Across different cohorts of students what is the single most important predictor of dropout?

Answer: A predictor that appears among the most influential 5 in all 4 cohorts is the “pre-deadline submission time”. It is the duration between when the student starts to work on the problems and their deadline. Perhaps it is indicative of how busy the learner is otherwise.

Q 3

Q. What predicts a student successfully staying in the course through the final week?

Answer: A student’s average number of weekly “submissions” (attempts on problems) *relative* to other students', e.g. a percentile variable, is highly predictive. It appears that relative and trending predictors drive accurate predictions. A student's lab grade each week is more predictive than a count of his/her problem submissions. ($\chi^2$ value)
Q6 Great Predictions

Q. When can we predict well?
Answer: It is easy to predict accurately 1 week in advance. In general, when predicting one week in advance, our models averaged ~0.88 AUC accuracy. Our week 7 model predicting week 8 dropout has an AUC = 0.95 for the fully collaborative student cohort. Its high accuracy is likely due to the data on mid-term participation during week 8. For the forum contributor cohort, our week 8 models predicting week 9 and week 10 dropout were also highly accurate. Here the data supported just slightly lower AUC (0.87).

Predicting Next Week

More info

- **Edx blog post**
  - Who is likely to drop out and why?
    - https://www.edx.org/blog/who-likely-drop-out-why


- arXiv #1408.3382
  - **Likely to stop? Predicting Stopout in Massive Open Online Courses**, Colin Taylor, Kalyan Veeramachaneni, Una-May O’Reilly

- arXiv #1407.5238
  - **Towards Feature Engineering at Scale for Data from Massive Open Online Courses**, Kalyan Veeramachaneni, Una-May O’Reilly, Colin Taylor.

Future Work

Is there a universal predictor?
Model transfer learning

- Forthcoming paper
  - Transfer Learning for Predictive Models in Massive Open Online Courses, Sebastien Boyer and Kalyan Veeramachaneni,
    » 17th International Conference in Artificial Intelligence in Education, 2015
  - One course, successive offerings

- MoocDB Data Science Commons
- Interventions and designs
Thank you!

- Questions?