Digital Learner Quantified
and
Towards MOOC Data Science Commons

Kalyan Veeramachaneni

Contact: kalyan@csail.mit.edu

Joint work with
Una-May O’Reilly, Colin Taylor, Elaine Han, Quentin Agren,
Franck Dernoncourt, Sherif Halawa, Sebastien Boyer, Max Kanter

Any Scale Learning for All Group
CSAIL, MIT
Suppose…

Given learners interactions up until a time point, we want to predict if s/he will dropout/stopout in the future?

- We must use *click stream, forums* as well *assessments*

We can use students data during these weeks

Note: By varying lead and lag we get 91 prediction problems
What can we do with that matrix?

- Cluster/segment
  - Lurkers, high achievers, interactive

- Predict an outcome
  - Who is likely to dropout?

- Analytics
  - Did this video help?
  - Correlation with performance
What can we do with that matrix?

Cluster/segment
- Lurkers, high achievers, interactive

Predict an outcome
- Who is likely to dropout?

Analytics
- Did this video help?
- Correlation with performance

Supervised learning machinery
- Neural networks, SVMs, Random Forests

Unsupervised learning machinery
- Gaussian mixture models, Bayesian clustering

Probabilistic modeling
- Graphical models, HMMs
But…. How did the matrix come about?

Think and propose

Extract
But….How did the matrix come about?

Curation of raw data

Feature engineering
But....How did the matrix come about?

Think and propose

Extract

Curation

Variable engineering

Machine learning
How do we shrink this?

Think and propose

Extract

Curation

Variable engineering

>6 months
How did the matrix come about?

Curation → Feature engineering → Machine learning

Think and propose → Extract

> 6 months → a week

Covariates

Entities

Time spent, Time before deadline, Number of correct answers, Number of forum responses, Time spent during weekends
The Overarching theme of my research

• How can we reduce time to process, analyze, and derive insights from the data?
How to shrink this time?

- Build fundamental building blocks for reuse
- Understand how folks in a certain domain interact with the data
  - make this interaction more efficient
- Increase the pool of folks who can work with the data
Organized data model

Digital learner quantified

Feature engineering

Multiple modeling pathways

MLBlocks

Extract, Interpret, Aggregate

Time series, Discretize, hmm train

hmm predict, Evaluate

Flat, Train classifier, Classifier predict

Classifier predict, Evaluate

ALFA

ANYSCALE LEARNING FOR ALL
What we would like to capture and store?

- Who, When, What Where?

Organize
What we would like to capture and store?

- **Who**, When, What Where?
What we would like to capture and store?

- **Who, When, What Where?**

Organize
What we would like to capture and store?

- **Who, When, What Where?**

Organize
What we would like to capture and store?

- **Who, When, What, Where?**

```
<table>
<thead>
<tr>
<th>observed_events</th>
<th>resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed_event_id</td>
<td>resource_id</td>
</tr>
<tr>
<td>user_id</td>
<td>resource_name</td>
</tr>
<tr>
<td>url_id</td>
<td>resource_uri</td>
</tr>
<tr>
<td>observed_event_timestamp</td>
<td>resource_type_id</td>
</tr>
<tr>
<td>observed_event_duration</td>
<td>resource_parent_id</td>
</tr>
<tr>
<td>observed_event_ip</td>
<td>resource_child_number</td>
</tr>
<tr>
<td>observed_event_os</td>
<td>resource_relevant_week</td>
</tr>
<tr>
<td>observed_event_agent</td>
<td>resource_release_timestamp</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>urls</th>
<th>resources_urls</th>
</tr>
</thead>
<tbody>
<tr>
<td>url_id</td>
<td>resources_urls_id</td>
</tr>
<tr>
<td>url</td>
<td>resource_id</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>resource_types</th>
</tr>
</thead>
<tbody>
<tr>
<td>resource_type_id</td>
</tr>
<tr>
<td>resource_type_content</td>
</tr>
<tr>
<td>resource_type_medium</td>
</tr>
</tbody>
</table>
```

**Context Medium Hierarchy**
Organize: Constructing deeper hierarchies
Organize: Contextualizing an event

**Event Details**

- **EventType**: `problem_check`
- **Page**: `http://a/b/c/`
- **Problem ID**: `i4x://problem/123`
- **Resource**: `http://a/b/c/problem/123`

**Diagram Details**

- Interaction event: `problem_check`
- URL: `http://a/b/c/`
- Module URI: `i4x://problem/123`
- Constructed URI: `http://a/b/c/problem/123`
Organize: Inheritance

Navigational Event

Interaction Event

Sequence 1

Panel 3

Sequence 1

Panel 3

Sequence 1

Panel 3

inherit
Organize: Inheritance

Event 1

Event 2

URL

? 

URL

URL A

inherit
Organize: preprocessing

```
<table>
<thead>
<tr>
<th>Resource</th>
<th>Time spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>R_0</td>
</tr>
<tr>
<td>b</td>
<td>R_1</td>
</tr>
<tr>
<td>c</td>
<td>R_{12}</td>
</tr>
<tr>
<td>d</td>
<td>R_{11}</td>
</tr>
<tr>
<td>e</td>
<td>R_{22}</td>
</tr>
</tbody>
</table>
```
Feature engineering
Primitive constructs

• Students activity falls into either of three
  – Spending time on resources
  – Submitting solutions to problems
  – Interacting with each other
  – Other (peer grading, content creation etc)

• Basic constructs
  – Number of events
  – Amount of time spent
  – Number of submissions, attempts
Feature engineering
Primitive constructs

![Diagram showing resource allocation and time spent](image)

<table>
<thead>
<tr>
<th>Resource</th>
<th>Time spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>(t_2 - t_1)</td>
</tr>
<tr>
<td>b</td>
<td>(t_3 - t_2)</td>
</tr>
<tr>
<td>c</td>
<td>(t_4 - t_3 + t_6 - t_5)</td>
</tr>
<tr>
<td>d</td>
<td>(t_5 - t_4)</td>
</tr>
<tr>
<td>e</td>
<td>(t_7 - t_6)</td>
</tr>
</tbody>
</table>
Feature engineering
Aggregates

- Aggregate by resource hierarchy
- Aggregate by resource type
  - Book, lecture, forums

<table>
<thead>
<tr>
<th>Resource</th>
<th>Time spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>$R_0$</td>
</tr>
<tr>
<td>b</td>
<td>$R_1$</td>
</tr>
<tr>
<td>c</td>
<td>$R_{12}$</td>
</tr>
<tr>
<td>d</td>
<td>$R_{11}$</td>
</tr>
<tr>
<td>e</td>
<td>$R_{22}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource</th>
<th>Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_0$</td>
<td>$a + b + c + d + e$</td>
</tr>
<tr>
<td>$R_1$</td>
<td>$b + c + d$</td>
</tr>
<tr>
<td>$R_2$</td>
<td>$e$</td>
</tr>
</tbody>
</table>
Feature Engineering: Primitive aggregates

Total time spent on the course
number of forum posts
number of wiki edits
number of distinct problems attempted
number of submissions (includes all attempts)
number of collaborations
number of correct submissions
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
Feature Engineering: Primitive constructs

<table>
<thead>
<tr>
<th>Learner</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Primitive

... | ... | ... | ... | ... |
Feature Engineering - Statistical interpretations

Percentiles, relative standing of a learner amongst his peers

Uni-variate explanation

<table>
<thead>
<tr>
<th>Learner</th>
<th>Feature value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verena</td>
<td>32</td>
</tr>
<tr>
<td>Dominique</td>
<td>61</td>
</tr>
<tr>
<td>Sabina</td>
<td>21</td>
</tr>
<tr>
<td>Kalyan</td>
<td>12</td>
</tr>
<tr>
<td>Fabian</td>
<td>32</td>
</tr>
<tr>
<td>John</td>
<td>33</td>
</tr>
<tr>
<td>Sheila</td>
<td>88</td>
</tr>
</tbody>
</table>

\[ \int = 73\% \]
Feature Engineering: Statistical interpretations

Percentiles, relative standing of a learner amongst his peers
Multivariate explanation

<table>
<thead>
<tr>
<th>Learner</th>
<th>Feature value 1</th>
<th>Feature value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verena</td>
<td>32</td>
<td>12.4</td>
</tr>
<tr>
<td>Dominique</td>
<td>61</td>
<td>2.3</td>
</tr>
<tr>
<td>Sabina</td>
<td>21</td>
<td>6.1</td>
</tr>
<tr>
<td>Kalyan</td>
<td>12</td>
<td>7.8</td>
</tr>
<tr>
<td>Fabian</td>
<td>32</td>
<td>12.4</td>
</tr>
<tr>
<td>John</td>
<td>33</td>
<td>12</td>
</tr>
<tr>
<td>Sheila</td>
<td>88</td>
<td>12.4</td>
</tr>
</tbody>
</table>

\[ \int = 68\% \]
Feature Engineering: Statistical interpretations

Trend of a particular variable over time
Rate of change of the variable

<table>
<thead>
<tr>
<th>John</th>
<th>Feature value</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>Feature value</td>
</tr>
<tr>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

![Graph showing trend and slope](image)
More complex
Learner’s topic distribution on a weekly basis

Only available for forum participants
Modeling the Learners time series using HMM

One learners matrix
HMM state probabilities as features

Features for a learner at the end of second week
More specifically

<table>
<thead>
<tr>
<th>State</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.109</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.23</td>
</tr>
<tr>
<td>2</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>0.112</td>
</tr>
<tr>
<td>4</td>
<td>0.12</td>
</tr>
<tr>
<td>5</td>
<td>0.5370</td>
</tr>
</tbody>
</table>

H

Week 1

3

4

H

Week 2

4

5

H

Features

\[ V_1 \ldots V_n \]

\[ t=3 \]

Week 14

\[ V_1 \ldots V_n \]
Feature Engineering
Digital learner quantified!

<table>
<thead>
<tr>
<th>Learner</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>Feature n-1</th>
<th>Feature n</th>
</tr>
</thead>
</table>

- **Primitive**
- **Statistical**
- **time series based (including hmm)**
## Feature Engineering

### Features/Variables

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td></td>
</tr>
<tr>
<td>Feature 2</td>
<td></td>
</tr>
<tr>
<td>Feature 3</td>
<td></td>
</tr>
<tr>
<td>Feature 4</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Feature n</td>
<td></td>
</tr>
</tbody>
</table>

### Statistical interpretations

- Frequency of pdf
- Feature value
- John

### Aggregation

<table>
<thead>
<tr>
<th>Resource</th>
<th>Time spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>R_0</td>
</tr>
<tr>
<td>b</td>
<td>R_1</td>
</tr>
<tr>
<td>c</td>
<td>4b + b - a</td>
</tr>
<tr>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td>e</td>
<td>b</td>
</tr>
</tbody>
</table>

### Primitive constructs

<table>
<thead>
<tr>
<th>Resource</th>
<th>Time spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>R_0</td>
</tr>
<tr>
<td>b</td>
<td>R_1</td>
</tr>
<tr>
<td>c</td>
<td>R_2</td>
</tr>
<tr>
<td>d</td>
<td>R_11</td>
</tr>
<tr>
<td>e</td>
<td>R_0</td>
</tr>
</tbody>
</table>

Fully automated
What we can’t automate?

- **Constructs that are based on our intuition**
  - average time to solve problem
  - observed event variance (regularity)
  - predeadline submission time (average)
  - Time spent on the course during weekend

- **Constructs that are contextual**
  - pset grade (approximate)
  - lab grade
  - Number of times the student goes to forums while attempting problems

- **Ratios**
  - time spent on the course per-correct-problem
  - attempts per correct problems

- **Constructs that are course related**
  - Performance on a specific problem/quiz
  - Time spent on a specific resource
Feature Factory
Crowd source variable discovery

Feature discovery is a challenging aspect of the data science and knowledge discovery. Creating an online interactive space where data scientists can benefit from each other’s ideas on various features can significantly simplify and expedite the process. Feature Factory is an online platform where ALFA@CSAIL will present a prediction problem for which features are sought. For the prediction problem, the group will provide downloadable mock data so users can test their scripts and submit. Feature Factory seeks three kinds of contributions: ideas of new features, feature extraction code and comments on existing ones.

Upon the submission of the feature extraction code, it will be validated on our online mock dataset and you will be notified of the result immediately. Upon validation, our team will execute the code on the real dataset to generate the features and insert the new feature into a number of machine learning models using discriminative (Decision trees, Neural networks, support vector Machines), generative (logistic regression, Gaussian process) and time series models. As a result, your features will be ranked against one another.

Current Focus Problem: Predict Student Stopouts on Massive Open Online Courses
In this problem, our goal is to predict when a student will stop engaging with the course. A student is assumed to have stopped out from a course when s/he stops to attempt problems/homeworks. We have data captured from students online behavior, which includes click stream data, their online forum interactions and their submissions for problems. We have a comprehensive data schema, called MOOCdb which captures the student activity data on a MOOC platform. The data schema is documented here. A small mock dataset that is in the form of the data schema can be downloaded in two formats: sql or csv.

Featurefactory.csail.mit.edu
### Existing ideas and scripts

- **Average time (in days) the student takes to react when a new resource is posted.** This pretends to... [read more]
  - by Josep Marc Mingot

- **Average time between problem submission time and problem due date**
  - by Rob Miller

- **Total time spent on each resource during the week**
  - by Franck

- **Number of forum posts**
  - by Franck

- **Number of Wiki edits by week**
  - by Franck
How does one participate?

featurefactory.csail.mit.edu

1. Think and propose
2. Comment
3. Help us extract by writing scripts
### Feature Engineering

#### Features/Variables

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Statistical Interpretations

- **Frequency**
  - Feature: value
  - Distribution: $\sim 73\%$

#### Aggregation

<table>
<thead>
<tr>
<th>Resource</th>
<th>Time spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>1-2</td>
</tr>
<tr>
<td>R2</td>
<td>3-4</td>
</tr>
<tr>
<td>R3</td>
<td>5-6</td>
</tr>
<tr>
<td>R4</td>
<td>7-8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource</th>
<th>Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>R1</td>
</tr>
<tr>
<td>R2</td>
<td>R2 + R3</td>
</tr>
<tr>
<td>R3</td>
<td>R4</td>
</tr>
<tr>
<td>R4</td>
<td>R4</td>
</tr>
</tbody>
</table>

#### Primitive Constructs

- **User Defined**

A new idea:

```python
function [feature_value]
    = newfeature (learner, time_interval)

% Calculate a feature on a per learner basis
do this ...
do this ...
return
```
What did we assemble as variables so far?

**Simple**
- Total time spent on the course
- Number of forum posts
- Number of wiki edits
- Average length of forum posts (words)
- Number of distinct problems attempted
- Number of submissions (includes all attempts)
- Average number of attempts
- Number of collaborations
- Max observed event duration
- Number of correct submissions

**Complex**
- Average time to solve problem
- Observed event variance (regularity)
- Total time spent on lecture
- Total time spent on book
- Total time spent on wiki
- Number of forum responses
- Predeadline submission time (average)

**Derived**
- Attempts percentile
- Pset grade (approximate)
- Pset grade over time
- Lab grade
- Lab grade over time
- Time spent on the course per-correct-problem
- Attempts per correct problems
- Percent submissions correct
What did we assemble as variables so far?

Simple
Total time spent on the course
number of forum posts
number of wiki edits
average length of forum posts (words)
number of distinct problems attempted
number of submissions (includes all attempts)
number of distinct problems correct
average number of attempts
number of collaborations
max observed event duration
number of correct submissions

Complex
average time to solve problem
observed event variance (regularity)
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
predeadline submission time (average)

Derived
attempts percentile
pset grade (approximate)
pset grade over time
lab grade
lab grade over time
time spent on the course per-correct-problem
attempts per correct problems
percent submissions correct

Note:
• Red were proposed by crowd
• For definitions of simple, complex and derived
  Please check out http://arxiv.org/abs/1407.5238
Dropout prediction problem

Given current student behavior if s/he will dropout in the future?

We can use students data during these weeks

Note: By varying lead and lag we get 91 prediction problems
The 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, …
- 5000+ models learned and tested
  - 91 prediction problems for each of 4 cohorts
  - 10 fold cross validation and once on entire training -> 11 models per problem
  - Extra modeling to examine influential features
  - Multi-algorithm modeling on problems with less accurate models
  - HMM modeling and 2-level HMM-LR modeling
Splitting into cohorts

- Passive collaborator, 44526
- Fully collaborative, 441
- Discussion generator, 7860
- Content generator, 112
- No Attempts, 52683
Multiple modeling pathways

Organized data model

Digital learner quantified

Feature engineering

MLBlocks

Extract
Interpret
Aggregate

Time series
Discretize
hmm train

hmm predict
Evaluate

Form hmm features
Train classifier
Classifier predict
Evaluate

Train classifier
Classifier predict
Evaluate

Flat
Models

- Logistic regression
- Hidden markov models
- Hidden markov models + LR
- Randomized logistic regression
  - For variable importance
Learner per-week variable matrix

\[
\begin{array}{cccc}
  x_1 & x_2 & \bigcirc & \bigcirc & x_m & S \\
  w_1 & & & & & \\
  w_2 & & & & & \\
  w_{13} & & & & & \\
  w_{14} & & & & & \\
\end{array}
\]
Data Representation
Flattening it out for Discriminatory Models

Lag 2 – Lead 11 prediction problem

<table>
<thead>
<tr>
<th>Weeks</th>
<th>X₁</th>
<th>X₂</th>
<th></th>
<th>Xₘ</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>W₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W₁₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W₁₄</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Week 1
X₁ X₂ ○ ○ Xₘ

Week 2
X₁ X₂ ○ ○ Xₘ
Hidden Markov Model as a Prediction Engine

Week 1 data, predict 2 weeks ahead
Hidden Markov Model as a Prediction Engine

Week 1 data, predict 3 weeks ahead
Hidden state probabilities as variables

Use 2 weeks data, predict 3 weeks ahead

Week 1
- State 1: 0.02
- State 2: 0.001
- State 3: 0.84
- State 4: 0.03
- State 5: 0.109

Week 2
- State 1: 0.23
- State 2: 0.001
- State 3: 0.112
- State 4: 0.12
- State 5: 0.5370

Week 3
- Variables: 0.23, 0.001, 0.112, 0.12, 0.5370

Week 4

Week 5

Class label 5

Lag=2 weeks

Lead=2 weeks

Use 2 weeks data, predict 3 weeks ahead
Hidden state probabilities

Logistic Regression

Number of hidden states - 27
Randomized Logistic Regression

- Iterate over data multiple times
  - Each time resampling from data
  - Identify feature weights
  - Aggregate over multiple trials
Randomized Logistic Regression

Counts

Complex

Average feature weight

Feature number

Crowd proposed
Influential Predictors

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 all problems include self-tests and homeworks for grade) \*relative\* to other students', e.g. a percentile variable, is highly predictive.

Relative and trending predictors drive accurate predictions. E.G. a student's lab grade in current week relative to average in prior weeks is more predictive than the grade alone.
Influential Predictors

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 average “pre-deadline submission time”. It is the average duration between when the student submits a homework solution and its deadline.
Interesting Predictors

Human: how regularly the student studies
  • X13 “observed event variance”
    – Variance of a student’s observed event timestamp

Human: Getting started early on pset
  • X210: average time between problem submission and pset deadline

Human: how rewarding the student’s progress feels
  • “I’m spending all this time, how many concepts am I acquiring?”
  • X10: Observed events duration / correct problems

Student: it’s a lot of work to master the concepts
  • Number of problems attempted vs number of correct answers
  • X11: submissions per correct problem

Instructor: how is this student faring vs others?
  • tally the average number of submission of each student,
  • student variable is his/her percentile (X202) or percentage of maximum of all students (X203)

Instructor: how is the student faring this week?
  • X204: pset grade
  • X205: pset grade trend: difference in pset grade in current week to student’s average pset grade in past weeks
Top 10 features/variables that mattered

- For an extremely hard prediction problem
- Week 1
  - Number of distinct problems correct
  - Predeadline submission time
  - Number of submissions correct
- Week 2
  - Lab grade
  - Attempts per correct problem
  - Predeadline submission time
  - Attempts percentile
  - Number of distinct problems correct
  - Number of submissions correct
  - Total time spent on lectures

We can use students data during these weeks.
Parameters throughout this process

• Choices we make during the calculations of primitive constructs
  – Cut-offs for duration calculation

• Aggregation parameters

• Parameters for models
  – Number of hidden states
  – Number of topics

• We would next like tune these parameters against a prediction goal
What else can we predict?

<table>
<thead>
<tr>
<th>Learner</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>Feature n-1</th>
<th>Feature n</th>
</tr>
</thead>
</table>

- **Primitive**
- **Statistical**
- **time series based (including hmm)**

We can reuse

We can change this.
The MOOCdb Project

Towards MOOC data science commons
The MOOCdb Project

Status in April 2014
Create translation and curation software
Our current resources

Translate

Software

Github

MOOCdb Docs

Wiki
Shared infrastructure

> 100 course offerings
  67 Coursera courses
  30 edX courses
  5 OpenedX courses

Total number of events processed
> 1 Billion

Total amount of data
~2TB

Finished

In progress
Apps on top of shared data model

- An end-to-end packaged software chain with several value additions:
  - Integrated curation and processing tools
  - Adequate and easy-to-use interfaces
  - Documentation (web based)
  - Use cases
  - Demonstration
Current Apps (in progress)

Coursera
Translate

edX
Translate

MY MOOCViz

LabelMe

MY FEATURE FACTORY

Online evolving communities social networks

Current apps circa 2015
Acknowledgements- Students

- Roy Wedge
- Kiarash Adl
- Kristin Asmus
- Sebastian Leon
- Franck Dernoncourt
- Elaine Han aka Han Fang
- Colin Taylor
- Sherwin Wu
- Kristin Asmus
- John O’Sullivan
- Will Grathwohl
- Josep Mingot
- Fernando Torija
- Max Kanter
- Jason Wu
Acknowledgments

Sponsor: Project QMULUS

PARTNERS
- Lori Breslow
- Jennifer Deboer
- Glenda Stump
- Sherif Halawa
- Andreas Paepcke
- Rene Kizilcec
- Emily Schneider
- Piotr Mitros
- James Tauber
- Chuong Do