Identifying At-Risk Students in Massive Open Online Courses

Yi Han

May 16, 2019

Abstract

Massive Open Online Courses (MOOCs) have received widespread attention for their potential to scale higher education, with multiple platforms such as Coursera, edX and Udacity recently appearing.

Major Problem: low completion rates

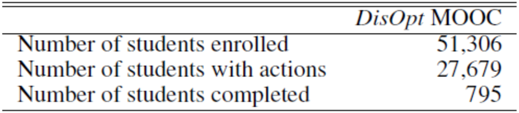
We explore the accurate early identification of students who are at risk of not completing courses.

Background

MOOCs aim to make higher education accessible to the world, by offering online courses from universities for free, and have attracted a diverse population of students from a variety of age groups, educational backgrounds and nationalities.

Major Problem

MOOCs face a major problem: low completion rates.



	DisOpt MOOC
Number of students enrolled	51,306
Number of students with actions	27,679
Number of students completed	795

Figure: Low Completion of MOOCs

Of 51,306 students enrolled, only 795 students completed: a completion rate of 1.5%. And only 27,679 (about 54%) students ever engaged in lectures and quizzes/assignments; even restricted to this group, the completion rate was a mere 2.9%.

Contributions

1. The first exploration of early and accurate prediction of students at risk of not completing a MOOC, with evaluation on multiple offerings, under potentially non-stationary data
2. An intervention that presents marginal students with meaningful failure probabilities: to the best of our knowledge a novel approach to completion rates
3. Two transfer learning logistic regression algorithms which would be practical for deployment in MOOCs, for balancing accuracy & inter-week smoothness
4. Experiments on two offerings of a Coursera MOOC that establish the effectiveness of our algorithms in terms of accuracy, inter-week smoothness and calibration

Problem Statement

Problem Definition

We explore the accurate and early prediction of students who are at risk of failing, which we cast as a supervised binary classification task where possible class labels are whether or not a student will fail a course.

4 Requirements

- Early & accurate predictions
- Well-calibrated probabilities
- Smoothed probabilities
- Interpretable models

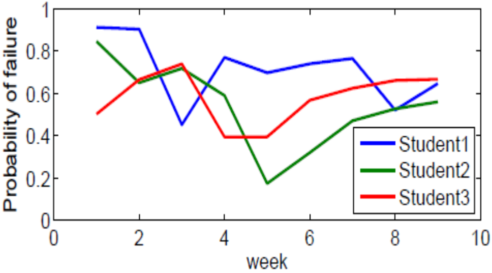


Figure: Failure-probability trajectories for three students across nine weeks produced by logistic regression with cross-validation performed weekly on DisOpt launched in 2014.

Alogrithms

Logistic regression predicts label y (fail for $y = 1$ and pass for $y = -1$) for input vector x_{ij} (a student) according to

Basic Logistic Regression

$$\begin{aligned} p(y|\boldsymbol{x}_{ij}, \boldsymbol{w}_i) &= \sigma(\boldsymbol{w}_i^T \boldsymbol{x}_{ij}) \\ &= \frac{1}{1 + exp(-y\boldsymbol{w}_i^T \boldsymbol{x}_{ij})} \\ \mathcal{L}(\boldsymbol{w}_i) &= \sum_{j=1}^{n_i} log(1 + exp(-y_{ij}\boldsymbol{w}_i^T \boldsymbol{x}_{ij})) + \frac{\lambda}{2}\|\boldsymbol{w}_i\|^2 \end{aligned}$$

where $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \ldots, w_{id_i}]^T$ is the weight vector to be learned.

Alogrithms cont.

In order to smooth probabilities across weeks, we propose a transfer learning algorithm, Sequentially Smoothed Logistic Regression (LR-SEQ).

LR-SEQ

$$\mathcal{L}(\boldsymbol{w}_i) = \sum_{j=1}^{n_i} log(1 + exp(-y_j \boldsymbol{w}_i^T \boldsymbol{x}_{ij})) + \frac{\lambda_1}{2} \|\boldsymbol{w}_i\|^2 + \lambda_2 \sum_{j=1}^{n_{i,i-1}} \|\boldsymbol{w}_i^T \boldsymbol{x}_{ij}^{(i,i-1)} - \boldsymbol{w}_{i-1}^T \boldsymbol{x}_{i-1j}^{(i,i-1)}\|^2$$

LR-SEQ seeks to minimize the difference between$\boldsymbol{w}_i \boldsymbol{x}_i^{(i-1,i)}$ and $\boldsymbol{w}_{i-1} \boldsymbol{x}_{i-1}^{(i-1,i)}$ where $\boldsymbol{x}_i^{(i-1,i)}$ denotes the set of students in week i that also exist in week $i-1$.

Alogrithms cont.

The drawback of LR-SEQ is that early inaccurate predictions cannot benefit from the knowledge learned in later weeks. To combat this effect, we propose Simultaneously Smoothed Logistic Regression (LR-SIM) that simultaneously learns models for all weeks. LR-SIM allows early and later prediction to be correlated and to influence each other.

Alogrithms cont.

LR-SIM

$$\begin{bmatrix} x_{1j} & [0,\ldots,0] & \ldots & [0,\ldots,0] \\ [0,\ldots,0] & x_{2j} & \ldots & [0,\ldots,0] \\ \vdots & \vdots & \ddots & \vdots \\ [0,\ldots,0] & [0,\ldots,0] & \ldots & x_{nj} \end{bmatrix}$$

LR-SIM first extends the feature space for each student $\boldsymbol{x}_{ij}$ to a new space with n components. The student $\boldsymbol{x}^{*}_{ij}$ with new feature space has $\sum_{i=1}^{n} d_i$ dimensions, with the ith component having di features corresponding to the features in the original feature space by the end of week i, and others zero.

Alogrithms cont.

Based on the extended $\boldsymbol{x}^*_{ij}$ and $\boldsymbol{w}$, we can minimize the difference of probabilities predicted for week i and week $i-1$ for $i(i \geq 2)$ together.

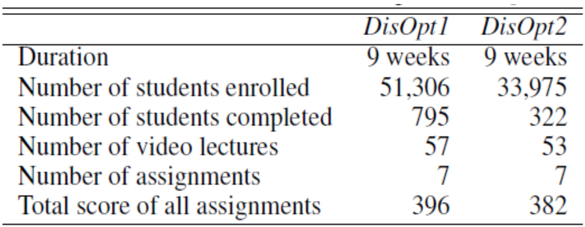
LR-SIM

$$\mathcal{L}(w) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} log(1 + exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_{*ij})) + \frac{\lambda_1}{2}\|\boldsymbol{w}\|^2$$

$$+ \lambda_2 \sum_{i=2}^{n}\sum_{j=1}^{n_{i,i-1}} \|\boldsymbol{w}^T \boldsymbol{x}^{*(i,i-1)}_{ij} - \boldsymbol{w}^T \boldsymbol{x}^{*(i,i-1)}_{i-1j}\|^2$$

Experimental Results

Dataset Preparation



	DisOpt1	*DisOpt2*
Duration	9 weeks	9 weeks
Number of students enrolled	51,306	33,975
Number of students completed	795	322
Number of video lectures	57	53
Number of assignments	7	7
Total score of all assignments	396	382

Figure: Overview on two offerings for DisOpt

Experimental Results cont.

Dataset Preparation

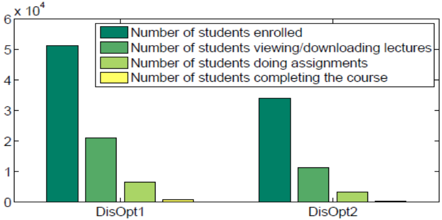
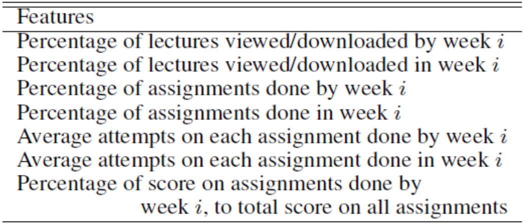


Figure: Student participation in the first and second offering of Discrete Optimization

In DisOpt1, among all the students enrolled, only around 41%, 13% and 2% of the students watch/download videos, do assignments and complete the course respectively. The same thing happens in DisOpt2 with low completion rate, and DisOpt2 had fewer students enrolled.

Experimental Results cont.

Features Used



Features
Percentage of lectures viewed/downloaded by week i
Percentage of lectures viewed/downloaded in week i
Percentage of assignments done by week i
Percentage of assignments done in week i
Average attempts on each assignment done by week i
Average attempts on each assignment done in week i
Percentage of score on assignments done by week i, to total score on all assignments

Figure: Features for each week i for DisOpt

We extract features from student engagement with video lectures and assignments, and performance on assignments by the end of each week to predict their performance at the end of the course.

Experimental Results cont.

Smoothness Measure

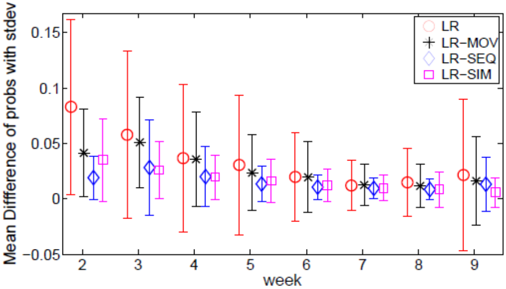


Figure: Comparison of LR, LR-MOV, LR-SEQ and LRSIM on smoothness across weeks. Mean difference of probabilities across students plus/minus standard deviation. Closer to zero difference is better.

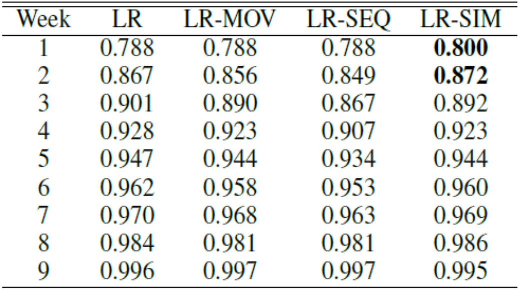
Experimental Results cont.

Smoothness Measure

LR-SEQ and LR-SIM achieve better smoothness (average difference) and low standard deviation, especially in the first five weeks where early intervention is most critical. LR attains smooth probabilities in the last few weeks, but with high standard deviation, when intervention is less impactful. LR-MOV achieves the same smoothness as LR with reduced standard deviation, demonstrating the need for performing some kind of smoothing.

Experimental Results cont.

AUC Measure



Week	LR	LR-MOV	LR-SEQ	LR-SIM
1	0.788	0.788	0.788	**0.800**
2	0.867	0.856	0.849	**0.872**
3	0.901	0.890	0.867	0.892
4	0.928	0.923	0.907	0.923
5	0.947	0.944	0.934	0.944
6	0.962	0.958	0.953	0.960
7	0.970	0.968	0.963	0.969
8	0.984	0.981	0.981	0.986
9	0.996	0.997	0.997	0.995

Figure: Comparison of LR, LR-MOV, LR-SEQ and LR-SIM on AUC across weeks.

LR-SIM maintains or even improves on LR's AUC in early weeks, while LR-SEQ suffers slightly inferior AUC in the first few weeks, and is comparable to LR in the last few weeks.

Experimental Results cont.

Parameter Analysis

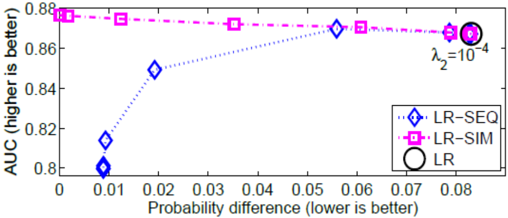


Figure: Smoothness versus AUC for LR, LR-SEQ and LR-SIM for week 2 when $\lambda_1 = 10$, and $\lambda_2 = 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}; 10^{0}; 10^{1}; 10^{2}; 10^{3}; 10^{4}$.

Experimental Results cont.

Calibration

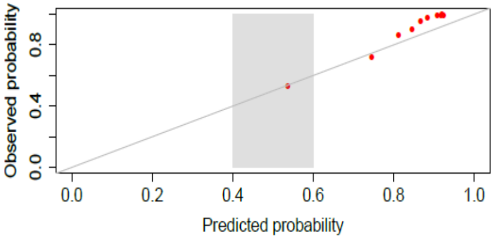


Figure: Reliability diagram for class fail using LR-SIM week 2. Grey area shows an intervention interval [0.4,0.6], which could be varied according to educational advice.

Figure shows the reliability diagram using LR-SIM for week 2. Our predicted probabilities agree closely with the observed probability in the gray region of marginal at-risk students for whom we wish to intervene.