A tutorial on machine learning in educational science

Łukasz Kidziński[1], Michail Giannakos[2], Demetrios G. Sampson[3], and Pierre Dillenbourg[1]

[1] École Polytechnique Fédérale de Lausanne, Switzerland
[2] Norwegian University of Science and Technology, Norway
[3] University of Piraeus, Greece

Abstract. Popularity of massive online open courses (MOOCs) allowed educational researchers to address problems which were not accessible few years ago. Although classical statistical techniques still apply, large datasets allow us to discover deeper patterns and to provide more accurate predictions of student's behaviors and outcomes. The goal of this tutorial is to disseminate knowledge on elementary data analysis tools as well as facilitating simple practical data-analysis activities with the purpose of stimulate reflection on the great potential of large datasets. In particular, during the tutorial we introduce elementary tools for using machine learning models in education. Although, the methodology presented here applies in any programming environment, we choose R and CARET package due to simplicity and access to the most recent machine learning methods.

Keywords. MOOCs, educational data mining, learning analytics

1 Introduction

Continuous advancement in data collection and storage techniques changed many industries and research areas. Internet is taking the role of libraries, twitter brings information to public faster than any newspaper and stock markets are run by high-frequency trading algorithms. In education, still substantial part happens in classroom, however, we also experience new, global initiatives, exemplified by massive online open courses (MOOCs).

One of the key challenges of MOOC research is closing the gap between educational science and online education [2] [3]. Increasing number of computer scientists and data scientists are trying to solve educational problems without contextual knowledge, whereas educational scientist are often not familiar with modern modelling techniques.

Most of educational experiments were run on small groups of students, often from the same school, sharing similar background. Online education allows us not only to see a bigger picture, with millions of students from all over the world, but also gives us opportunity to approach each of these students individually.

New data streams require new methodology. In classical approach, with, say, 50 students in each condition, we could just apply t-test or ANOVA. Since the datasets were small, only the large effects were detectable, so the notion of significance implicitly implied relevance. Conversely, when the number of students is large we can easily end up in rejecting the null hypothesis and detecting an effect irrelevant in practice. Moreover, in the massive context predictive models can be more accurate if only associated with large number of valuable variables.

During this tutorial we present methodology for forming and testing hypothesis in this new setup. We also present practical guidance for building data-driven predictive models with the state-of-the-art machine learning methods.

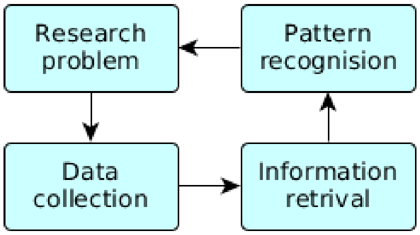

Fig. 1. The flow of data-driven educational research is now altered by information retrieval step, where we find an adequate representation of a vast dataset.

2 Dataset

We use the data from the Introduction to C++ and Introduction to Java offered by the EPFL in the fall 2013. We had 13787 students in the C++ course and 17716 students in the Java course. Both courses had very similar structure in terms of number of weeks, assignments and the abstract object-oriented content.

3 Hypothesis

The main step of any analysis is the formation of good questions. Classical experiment design still holds and large datasets allow us to analyze deeper patterns, like, for example: *to what extent perceived video difficulty is reflected by video interactions (pauses, speed ups, etc.)?* [5] or *Does forum activities and in-video interactions reflect decreasing engagement over time?* [9]. In this tutorial we predict the students grade based on their temporal behavior. We hypothesize that the model is independent of the course, since both courses have similar structure and they were supposed to deliver object-oriented programing paradigms. We expect students to behave similarly.

4 Data collection

As soon as we have formulated the hypothesis, we start gathering the data to support it. In the online context we can still use the classical tools (e.g. questionnaires), but also new sources of data are available. We can record, among others: clickstream (sequence of sites clicked), mouse moves, keyboard writing pattern, video interactions (pauses, forwards, etc.), scroll depth and growing amount of other information provided by web browsers.

In addition, in an experimental setup, we can ask users for access to their cameras or microphones. Increasing popularity of social networks may provide additional information about student's background and social context. Interesting research may arise just from the analysis of these streams of data. We can, for example, assess student's attention from the camera images, leveraging the small sample research [7].

In this work we analyze student's time series. We will look on the activities of the student over time, we extract information supporting the hypothesis and we build a predictive model. For each student we have a timestamps of events of following types: *Forum View, Forum Subscription, Thread View, Lecture Re-View, Thread Subscription, Post on Thread, Quiz Submission (Video), Quiz Re-Submission (Video), Assignment Re-Submission, Thread Launch, Quiz Re-Submission (Quiz), Forum Upvote, Quiz Submission (Quiz), Forum Downvote, Lecture Download, Lecture Re-Download, Comment on Post, Quiz Submission (Survey), Quiz Re-Submission (Survey), Lecture View, Assignment Submission, Registration.*

5 Information retrieval

After gathering the relevant data we extract features important for the research question. First, we extract simple characteristics like the number of: videos watched, posts written, posts read, etc. Next, we add more sophisticated constructs. To that end we use the existing domain knowledge and we explore the dataset.

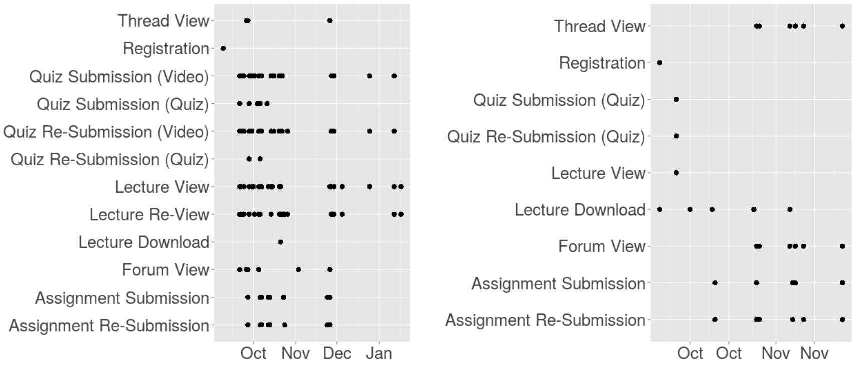

Fig. 2. Visualization of time series of two students, one who succeeded with 87% points (left) and one who failed with 66% points (right). Although both of them completed the assignments, the left one was clearly more engaged. Observations from visual assessment can help us to engineer variables informative in the given context.

In our context, visualization of a students' time series may give us insights about how to extract variables as illustrated in Figure 2. We can also look for well-establish constructs, defining, for example, *Procrastination* as the number of times a student submitted the assignment just before the deadline, *Persistence* as the number of retries of assignment submission or *Regularity* as the variance of difference between two watching sessions.

As an output of this step we have a large structured table, with one row per each student and extracted variables. In the next step we use this table for training machine learning models.

6 Pattern recognition

We identify two main branches of machine learning: supervised learning and unsupervised learning. The goal of supervised learning is to identify patterns within independent variables to explain a dependent variable. The key example here is the linear regression and logistic regression, known from classical statistics. Recent techniques like Support Vector Machines [1], Random Forests [6], Generalised Boosted Regression [8] and many others are gaining popularity due to their robustness, computational feasibility and effectiveness. The unsupervised learning whenever there is no dependent variable and we want to investigate patterns in the data, most commonly clusters of *similar* observations.

For our example, we use supervised learning to predict grades of students. To this end, we represent each student as a vector of his characteristics as described in Section 5. To fit the model to the known instances from the training set we need an accuracy measure. In our example we use the Root Mean Square Error, which, intuitively, expresses the mean distance of the prediction to the observed value.

```r
library(caret)
# Build the model
control <- trainControl(method="repeatedcv", number=10, repeats=3)
model.svm <- train(Grade ~ ., data=students.tr, method="svmRadial", trControl=control, tuneLength=5)
# Predict the grades
grades = predict(model, students.ts)
```

Listing 1.1. Building an SVM model using the CARET package in R.

Listing 1 presents a process of model building using the dataset with features described in Section 5. We use a very convenient R framework CARET [4] for application of machine learning methods. In particular, to choice of the underlying supervised learning technique is govern by `method` and with `method="rf"` we apply Random Forests instead of SVM. This allows us to quickly prototype and compare models.

Since over 90% of students dropout out before finishing any assignment, prediction of their score equal to 0 is easy, and therefore we focus only on students who achieved at least 10% points from the assignments.

To asses the quality of various models we look on the estimated RMSEs. The simple commands to compute and plot these values are presented in Listing 2, where we assume that models `model.svm`, `model.rf` and `model.gbm` were build as described above.

```
results <- resamples(list(svm = model.svm, rf = model.rf, gbm = model.gbm))
# boxplots of results
bwplot(results)
```

Listing 1.2. Comparison of performance of different models

The best model, Generalized Boosted Regression, achieves RMSE around 13 as presented in Figure 3. We consider this result satisfactory, taking into account simplistic, illustrative approach.

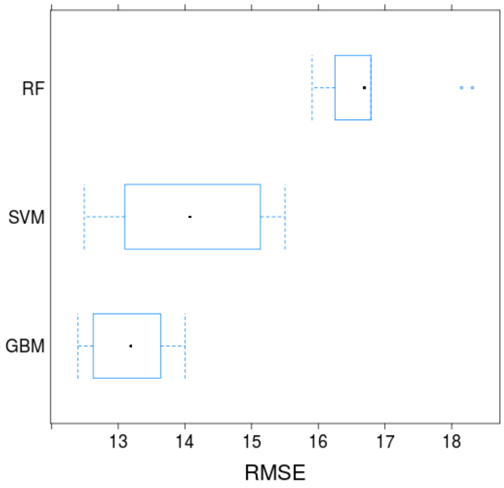

Fig. 3. Errors of each model in terms of the RMSE.

Exploratory Data Analysis [10] can be useful for finding an appropriate technique, for adequate data transformation, for outlier detections, etc. Moreover, this exploration brings new insights and hypothesis and eventually closes the cycle in Figure 1.

7 Discussion

The analysis of educational data in the massive context requires new techniques and methodologies. The goal of this tutorial was to shed light on usage of machine learning and the process of the analysis, in the context of on-line education. Since it is not possible to introduce advanced techniques in details during a short tutorial, we focused on illustrating the simplicity of application of state-of-the-art machine learning using the R package CARET.

References

1. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2(2), 121–167 (1998)
2. Dillenbourg, P.: Orchestration Graphs: Modeling Scalable Education. EPFL Press, Switzerland (2015)
3. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5-6), 304–317 (2012)
4. Kuhn, M.: Contributions from Jed Wing and Steve Weston and Andre Williams and Chris Keefer and Allan Engelhardt and Tony Cooper and Zachary Mayer and Brenton Kenkel and the R Core Team and Michael Benesty and Reynald Lescarbeau and Andrew Ziem and Luca Scrucca., caret: Classification and Regression Training (2015), http://CRAN.R-project.org/package=caret, r package version 6.0-47
5. Li, N., Kidziński, Ł., Jermann, P., Dillenbourg, P.: How do in-video interactions reflect perceived video difficulty? In: Proceedings of the European MOOCs Stakeholder Summit 2015. pp. 112–121. No. EPFL-CONF-207968, PAU Education (2015)
6. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News 2(3), 18–22 (2002), http://CRAN.R-project.org/doc/Rnews/
7. Raca, M., Kidziński, Ł., Dillenbourg, P.: Translating head motion into attention-towards processing of students body-language. In: Proceedings of the 8th International Conference on Educational Data Mining. No. EPFL-CONF-207803 (2015)
8. Ridgeway, G.: Generalized boosted models: A guide to the gbm package. Update 1(1), 2007
9. Sinha, T., Li, N., Jermann, P., Dillenbourg, P.: Capturing "attrition intensifying" structural traits from didactic interaction sequences of mooc learners. arXiv preprint arXiv:1409.5887 (2014)
10. Tukey, J.W.: Exploratory data analysis. Reading, Mass. (1977)