Using clickers to predict students final courses grades, an artificial intelligence approach.

Fionn Delahunty and Alexander Schliep

March 19, 2019

Using clickers to predict students' final courses grades, an artificial intelligence approach.

Keywords; Artificial Intelligence, Data Science, Clickers, Pedagogy, Interactive Learning

ABSTRACT

Clickers are useful for improving interaction in teaching. Here, we demonstrate how artificial intelligence can improve this usefulness even further by predicting final exam scores of students early in the semester. We also investigate how the number of attempts a student makes effect the correctness of the answer.

Authors note: This presentation is a short summary of the overall project and focuses more on the practical outcomes and less on the technical aspects of the project. A full and more technical description of the project can be found at https://schlieplab.org/Static/Supplements/KUL2019/FullReport.pdf.

INTRODUCTION

The traditional one directional teaching method, instructors lecturing to students, has been superseded in recent years by the concept of flipped classrooms and blended learning. These approaches have become common place in many higher level educational institutions in the western world. The transition stems from the work in the field of Pedagogy, where studies have repeatedly demonstrated that students taking an active role in the teaching process improves their learning [1]. This can take on many forms, one of the most basic and easily implemented is the concept of clickers.

Clickers are small hand-held devices that are assigned to each student in a lecture. During the lecture, instructors will ask questions which students can answer in real time using the clickers. They have been shown to increase attendance by up to 50% [2], improve attention span during lectures [3] and improve overall student satisfaction for a class [2]. Clickers come in the form of infrared devices, or smart phone apps and are often low cost and require little technical training to either the students or the instructor [4].

One aspect of clicker usage that has received on-going debate within the pedagogy literature is how in-class clicker score relates to final exam performance [5]. Students have self-reported improved exam performance [6] but empirical studies have failed to sufficiently isolate the effects of clicker alone on final term exam performance.

In 2016, the CSE department of GU/Chalmers accepted its first cohort of students on the applied data science master's program. One of the courses on offer in this program is the ability to complete an individual research project with a supervisor sufficient to 7.5 ECTs. This report is a summary of the work completed in one such project.

In this short abstract, we report on the results of two hypotheses. We first hypothesised that clicker usage behaviour differs between students, while our second hypothesis was that clicker behaviour can be used to predict final exam performance through artificial intelligence. For instructors to be able to predict how students will perform on final exams during early periods would be a considerable advantage to supporting students in real time and providing a more personalized teaching schedule to all students.

EXPERIMENTAL SETTING

A. Data Collection

Our datasets were composed of two years' worth of clicker response to a required Computer Science course covering mathematical foundations in an American university. The first dataset contained 82 students while the second dataset contained 130 students. Since there was some difference in the teaching of the course over the two years, we mostly considered the datasets as separate. Responses were recorded for 24 lectures, along with scores from two midterm exams and a final term exam. The response system employed was iClicker.

HYPOTHESIS ONE

A. Introduction

Iclicker, and many other real time response systems provide instructors with information regarding how many clicker attempts a student has made before their final decision. Although question types vary, for all courses in our study students where provided with several multiple response answers and asked to choose one. Our measure of attempts was the number of different choices they made before the time elapsed. General experience might suggest to the instructor that students who make a lot of attempts are less confident in their answers, and in turn are probably less knowledgeable about the content being examined. This might lead to instructors using the numbers of attempts as a proxy measure of knowledge about a topic. However according to our research, no empirical investigation has investigated this connection. Our first hypothesis is to empirically investigate if the number of attempts does actually bare any relation to the correctness of an answer.

B. Data preparation

For this hypothesis, we combined both years' datasets for investigation. This presented us with a dataset which contained a total of 12,298 responses to in-class questions, of which 56% were correct. All rows which contained missing values were dropped.

C. Methodology

We investigated this hypothesis by comparing the mean number of attempts for all correct answers, and all incorrect answers.

D. Results

The mean number of attempts for all correct answers was 4.97, and 4.92 for incorrect answers. These two values are almost identical. We performed a Mann-Whitney *U* test to test if the two results are statistically significantly different, the test returned a non-significant result (*p*=0.29). Distribution plots for correct answers and incorrect answers are presented below in respective order. The two plots show a similar distribution. Following the results below, we accepted the null hypothesis that the number of responses to a question made no difference to whether the response was correct or false.

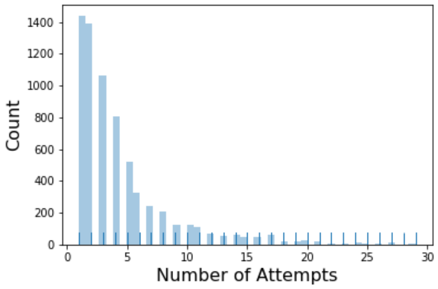

Figure 1. Distribution of correct answer attempts

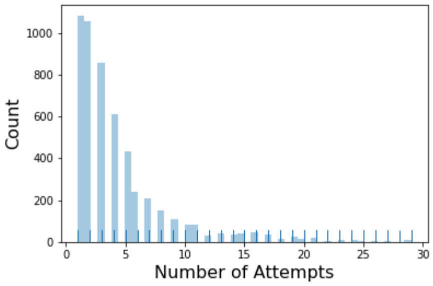

Figure 2. Distribution of incorrect answer attempts

E. Discussion and take-home messages

Using standard statistical tests, we demonstrate that the number of attempts a student make on a question, bare no relation to the chance that student will correctly answer that question. The take home message here is that instructors should not consider students who have a higher number of clicker attempts any less knowledgeable about a subject area, just because of their high number of attempts.

HYPOTHESIS TWO

A. Previous Research

Much of the research on predicting final exam score from students work early in the course period comes from within the discipline of computer science. The authors of [5] demonstrated that by analysing submitted assignments early in the semester, they can predict with up to 90% accuracy which grade band the student would fall into [5].

Work investigating the clicker score specifically has been undertaken by [4]. Their work does not employ artificial intelligence, but rather looks at correlations between clicker scores and final grades, finding a R value of .64 [4]. Their work is also limited to that of a single computer science course, and mostly focused on certain questions and how their respective topics predict final performance. Our work aims to build upon these two studies, by exploring how more advanced artificial intelligence methods may be better able to predict final performance from clicker scores.

B. Data Preparation

Following the work in [4], we grouped individual classes into periods of three. This was to prevent individual variance that may occur within a class [4]. For each student, we calculated the total number of correct answers out of all possible answers. This value was normalized by removing the mean and scaling to unit variance. A value of one would indicate all answers across those classes where correct, while a value of zero would indicate the opposite. Midterm and final exam scores were normalized using the same process.

C. Exploratory Data Analysis

We began by grouping all students in the course into four bands, based on their result in the final exam. We then average their inclass clicker score across the course period. The results are presented in Figure 3. Each trend line is a group of students. From this graph, we are presented with in-depth view into a student's learning across the course period. We see that students who fall in the highest final grade band (above 75% on the final exam) are consistently in that period across the course, and the same for all other bands.

Figure 4 presents us with the same information, but in a more fine-grained formatted. Interestingly, we see that some students who failed the final exam (>25%) scored quite highly in the in-class clicker scores. Which goes against the conventional wisdom that if you do well in class, you will also do well in the exam. We talk more about the take home message of this section in section E.

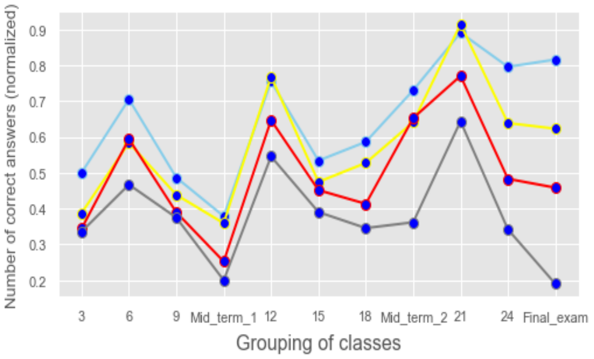

Figure 3. Students grouped into quartiles based on final exam score. First dataset.

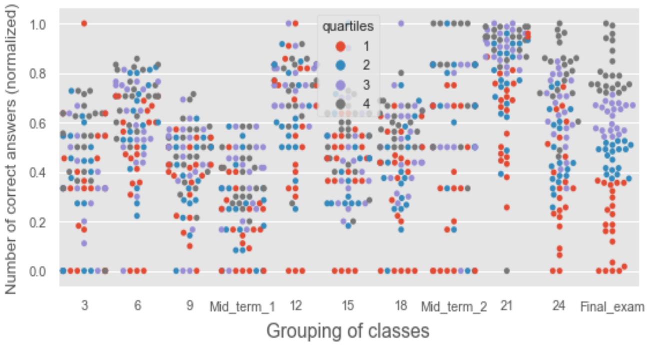

Figure 5. Swarm plot of students grouped in quartiles based on final exam score. First dataset

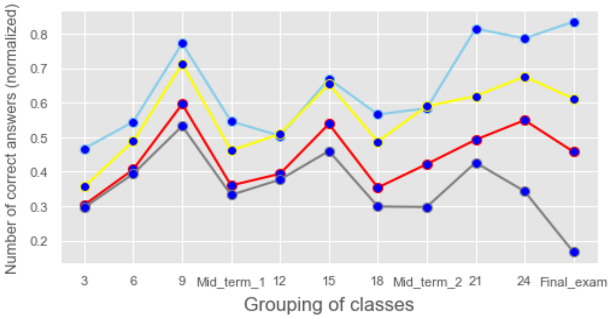

Figure 4. Students grouped into quartiles based on final exam score, mean scores of groups across classes. Second dataset

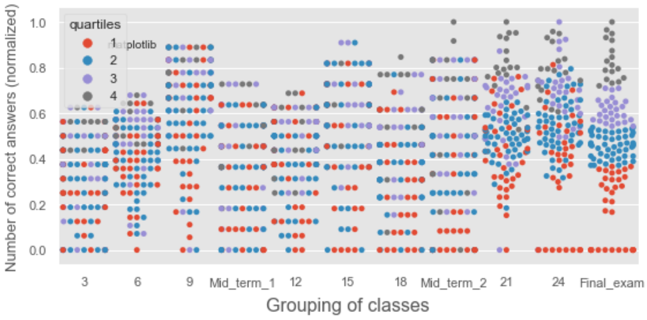

Figure 6. Swarm plot of students grouped into quartiles based on final exam scores. Second dataset

D. Artificial Intelligence Classifiers

The main aspect of this report was to investigate if we could develop an artificial intelligent classifier to predict final course grades during the course period. To do this, we draw on the field of machine learning. Our first step is to select a suitable algorithm for the data problem we have at hand. The way this algorithm works is by looking at the clicker data scores we provide it, and it tries to predict a relationship between these scores and the final exam grades.

For example, an incredibly basic relationship it might see is if all students who score 100% correct answers in the first two lectures, also score 100% in the exam. Than we show it some new data without exam scores, and it sees some students with 100% correct answers in the first two lectures. It would predict 100% in the exam for them. However, the relationships are often considerably more complex than this.

Fundamentally, the AI algorithm has three steps. Step one, it starts by looking at the clicker data and tries to see some relationship inside the data for each student. Step two, based on this relationship it predicts an exam score for each student. Step three, it finds out whether the prediction was correct or incorrect, if it was incorrect it tries step one and two again looking for a new relationship each time (see figure 7). This process repeats several hundred times until it develops a statistical model of how clicker scores relate to final grades.

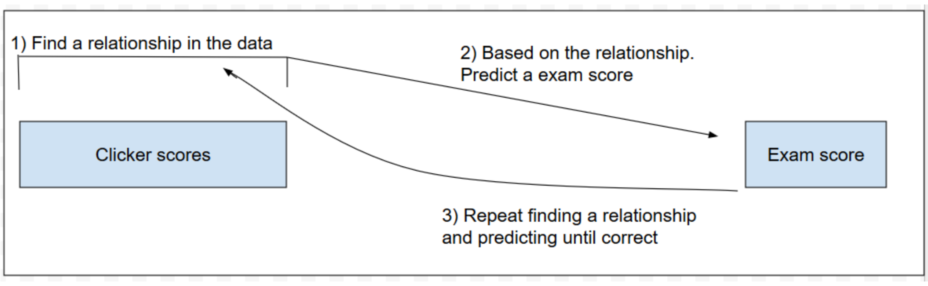
1) Find a relationship in the data
2) Based on the relationship. Predict a exam score
Clicker scores
Exam score
3) Repeat finding a relationship and predicting until correct

Figure 7. Visual diagram of how a machine learning algorithm works.

One method we use to measure the perform of this algorithm is by a metric called accuracy. Once our algorithm has learnt its relationship successfully, we ask it to predict scores on a new dataset which it hasn't seen before. For this dataset we know the final exam scores already, therefore we can compare the number of correct predictions from the algorithm.

$$Accuracy = \frac{The\ number\ of\ correct\ predictions\ from\ the\ algorithm.}{All\ predictions.} * 100$$

Our algorithm achieves an accuracy score of 46%. Which means if we show it clicker scores for a student, in 46% of cases it correctly predicts the band which the final exam score will fall into. Since we are trying to predict one of four classes, our baseline prediction is 25% (If you labelled all students as one class, 25% would be correct) while our actual prediction is 46%. Giving us 21% above baseline prediction power.

E. Conclusion and take-home messages

Firstly, in our exploratory data analysis section we see that how students perform in clicker questions often roughly matches how they perform in the final exam. This alone provides a useful take home message for instructors to monitor and support learning during the period. However, we did note some exceptions to this, specifically there was a cohort of students who despite performing very well in the clickers, scored very low in the exam. One proposal for this observation is possibly exam stress, and it might be interesting to see if this reoccurs in the context of take-home exams which are often less stressful.

In terms of the practicality of our algorithm, at this stage making use of it requires a medium degree of programming knowledge to set up the system. It was not within the scope of the project to develop an end to end application that instructors could use. Furthermore, the development of such a system would require a more in-depth study on the topic. However, as far as we are aware, no university in the world makes use of a system to predict student final course grades based on clicker scores during the term. This proof of concept study might suggest a future direction of research for some pedagogy or data science researchers in Chalmers or within Sweden. The successful implementation of such a system would possibly allow instructors to offer support to students "falling behind" or who were predicted to fail early in the course period before it is to late.

There are some limitations that should be noted however, the courses we investigated where 24 weeks in length, twice that of normal 7.5 ECTS Swedish courses. A 12-week period would arguably give less time to find a relationship between clicker scores and grades but wouldn't be impossible. The algorithm we developed would require retuning, but most of the fundamental work would still be useful. We believe that the fact this course was an American course should not make a different in terms of algorithm effectiveness. Factors that would affect its effeteness are the number of times the clicker is used in a class, whether course credit is offered for clicker responses etc. These are factors that we do not believe would majority differ between Sweden and America.

REFERENCES

[1] Jacqueline O'Flaherty and Craig Phillips. "The use of flipped classrooms in higher education: A scoping review". In: *The internet and higher education* 25 (2015), pp. 85–95.
[2] Kathy Kenwright. "Clickers in the classroom". In: *Tech Trends* 53.1 (2009), pp. 74–77.
[3] Kathleen Hoag, Janet Lillie, and Ruth Hoppe. "Piloting case-based instruction in a didactic clinical immunology course". In: *Clinical Laboratory Scienc*e 18.4 (2005), p. 213
[4] Leo Porter, Daniel Zingaro, and Raymond Lister. "Predicting student success using fine grain clicker data".In: *Proceedings of the tenth annual conference on International computing education research.* ACM. 2014,pp. 51–58.
[5] Alireza Ahadi et al. "Exploring machine learning methods to automatically identify students in need of assistance". In: *Proceedings of the eleventh annual International Conference on International Computing Education Research*. ACM. 2015, pp. 121–130.
[6] Gregory A DeBourgh. "Use of classroom "clickers" to promote acquisition of advanced reasoning skills". In: *Nurse Education in Practice 8*.2 (2008), pp. 76–87