Geo-Based Prediction of Academic Performance

Progress Report

Star Chen (sc7828), Scott Ye (my2152)

Problem Motivation

The outbreak of COVID-19 has had an impact on global education. To alleviate the drawbacks and challenges presented in the online learning environment, our study focuses on utilizing students' geolocation, a factor that is significant in EDM/LA but is seldom considered, to predict academic performance. Due to the lack of previous studies, our concentration is mainly on processing and analyzing the geolocation data to find the correlation between geolocation and students' academic performance. We divide our approach into three steps: derive geolocation from students' IP addresses, preprocess the IP geolocation data, and build a predictive model with appropriate machine learning techniques.

Current Status

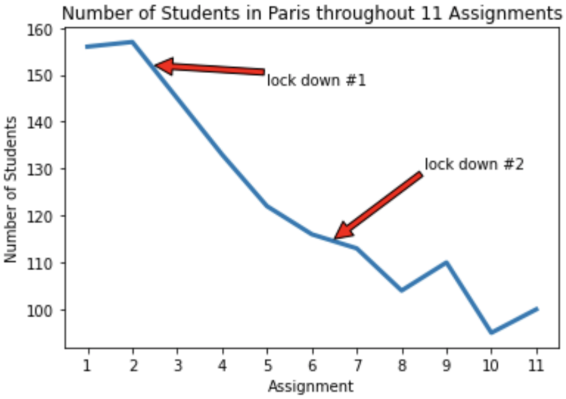
1. Preliminary Exploration
   Our dataset contains the assignments, labs, and exams information of 253 students studying at Sorbonne Université in Fall 2020. We first predicted the Midterm Score for F20 using the following linear regression.

   $$\text{Midterm Score} = \alpha \times \text{Average HW} + \beta \times \#\text{Labs Attended} + \epsilon \tag{1}$$

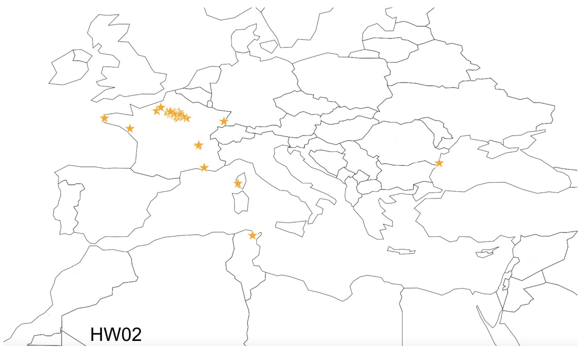
   The estimated coefficients are significant. Holding other variables constant, if lab attendance is increased by 1, midterm score will increase by 0.386. Similarly, if average HW score is increased by 1, midterm score will increase by 0.16. However, the r-squared is low for this prediction, meaning that the fluctuation of the students' grades are not adequately explained (10%) by homework grades and lab attendance. The result prompted us to study variations in geolocations as the disturbance term.

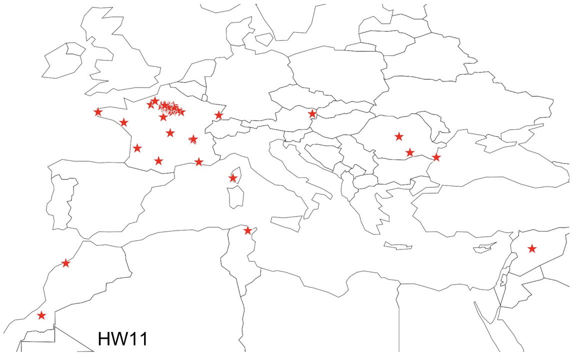
2. IP to Geolocation
   Using Abstract API, we created a dataframe that contains the corresponding information of each IP address, including longitude and latitude, city and country, and the use of VPN. There are 6646 unique IP addresses in total.

3. Data Processing
   From the geolocation data that we retrieved from students' IP addresses recorded by their online activities, we observed that geolocation can reflect much information on students' learning status.



The figure shows a decline in the number of students who submit their assignments in Paris where the campus is located. The red arrows indicate two lockdowns on the campus. Classes were forced to be taught online because of the lockdown. The decline reflects the influence of the lockdown on the students' decision of staying around the campus or returning home outside Paris. The following two graphs also illustrate the changes in students' geolocations before and after the lockdowns.





Work Plan

The previous results indicate that rather than the geolocation itself, the changes in students' geolocations may contain more information regarding their academic performance. Therefore, we will continue to work on the following two subtasks:

1. (Dec 2) Based on the geolocation data of the students, since most of the students live around Paris, we will divide the location into five categories, Paris, Île-de-France excluding Paris, France excluding Île-de-France, Europe excluding France, and out of Europe. Then, we will use cluster analysis to group the students according to their location changes among the five categories in assignments, and build predictive rules of academic performance with association rule mining.

2. (Dec 9) Derive the number and magnitude (distance) of students' change in geolocations. Depending on the prediction accuracy, we will consider including the time differences between each homework submission as another factor. We will adopt model selection and LASSO regularization to improve the regression models.