Relationship Between Students' Study Environments and Academic Results

Gary, Chon Kit Kuok

College of Education, University of Illinois at Urbana-Champaign

Department of Computer and Information Science, Faculty of Science and Technology, University of Macau

INTRODUCTION

This research is intended to investigate the university students' preferences on different study environments and how it affects their academic accomplishments. Intelligence is not the exclusive element which determines the academic achievement of a student. By analyzing the results, it is suggested that if students want to achieve better academic performance, they should study in places which they wish to, with several other factors' contribution that they can prepare. Most importantly is their exceptional perseverance.

The research accumulated data from various academic year's students. After that the data is interpreted with statistical models and machine learning algorithms. Both quantitative and qualitative methods are involved in the research.

Outcomes recommend that students should study in quiet places such as but not limited to libraries, in addition to the help of quietness and isolation, one can also eat some snacks during the study. Students who are not taking too many credits can avail their harvest, too. By doing so, they can focus on manifold courses more. Besides, having enough sleep is a significant factor that is dedicated to improving on academic consequences as well.

METHODS

In quantitative research progress, statistical models and machine learning algorithms are presented as below:

- **Classification** by **Logistic Regression** with **S**tochastic **G**radient **D**escent

Acquired data have nine parameters, each is regressed to GPA, these nine parameters are: Aloneness, Comfortableness, Concentration, EatDrink, Improvement, Isolation, Listen, Proximity and Quietness.

We define

$$y = \begin{cases} 1, if\ \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0, otherwise \end{cases}$$

Here β is the best fit parameter that is derived by an iterative search process, ε is the error derivation.

$$h(\theta) = \sum_{j=0}^{n} \theta_j x_j$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - h_{\theta(x^i)}\right)^2$$

Gradient descent is to minimize risk function and loss function in order to obtain a more sophisticated result. $h(\theta)$ is a fitting function, $J(\theta)$ is a loss function, n is the number of parameters, m is the number of training count, j is a parameter and i is a record.

We then obtain the loss function of each sample

$$cost\left(\theta, (x^i y^j)\right) = \frac{1}{2}\left(y^i - h_{\theta(x^i)}\right)^2$$

Take the partial derivative with respect to θ and update it

$$\theta_j' = \theta_j + \left(y^i - h_{\theta(x^i)}\right)x_j^i$$

Assume the length of descending step is α, we have

METHODS (cont.)

$$\theta_j' = \theta_j + \alpha\left(y^i - h_{\theta(x^i)}\right)x_j^i$$

Prior to get the final result, $J(\theta)$ needs to be regularized

$$J(\theta) = -[\frac{1}{m}\sum_{i=1}^{m} y^i log h_{\theta(x^i)} + (1 - yi)\log(1 - h_{\theta(x^i)})]$$

An advantage of using logistic regression other than linear regression for classifying is all results are simplified into 0 and 1, such output is expressed as 'possibility', so that results are convincible as well as eliminating possible outliers in the data.

Since each response is considered as independent, we can measure the accuracy of generalized model by mean squared error and mean absolute error to evaluate goodness of fit

$$MSE = \frac{\sum_{i=1}^{n}(y_i - mean(y_i))^2}{n} \qquad MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

Additionally, it's worth to compute the **Correlation** between these 10 variables to reflect correlation

$$Corr = \frac{\sum_{i=1}^{n}[(x_i - mean(x))(y_i - mean(y))]}{\sqrt{\sum_{i=1}^{n}(x_i - mean(x))^2 \sum_{i=1}^{n}(y_i - mean(y))^2}}$$

Next, we used the **Linear Regression** model to anticipate each candidate's GPA scale by ordinary least squares formula

$$min_w ||Xw - y||_2^2$$

Here the model fits with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between observed GPA scale in dataset, the responses predicted by linear approximation.

Apply Frobenius Norm for all values

$$||A||_F = \sqrt{\left[\sum_{i,j} abs(a_{i,j})^2\right]}$$

Last but not least, we used **k-means Clustering Algorithm** to cluster students into different groups by their sum of index of the nine variables in unsupervised learning setting. The formula is divided into two steps:

Given randomly chosen initial set of m centroids,

Assignment step:

$$S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \le \left\|x_p - m_j^{(t)}\right\|^2 \ \forall j, 1 \le j \le k\right\},$$

where each x_p can only take exactly one $S^{(t)}$

Update step:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|}\sum_{x_j \in S_i^{(t)}} x_j$$

||(args)|| notation is to calculate least squared Euclidean distance for each point. We keep iterating *m* until the centroids are converged when the result in assignment step no longer changes.

In the qualitative research process, a few interviewees were chosen to ask deeper questions regarding their survey responses. For example, they were asked to clarify their optimal study environments and their study habits. The coding schemas generated from the interview procedures are utilized as a proof for decoded data, as the same time providing complementary information to retrieve compelling elements that have important development of their collegiate results other than only based on study environments.

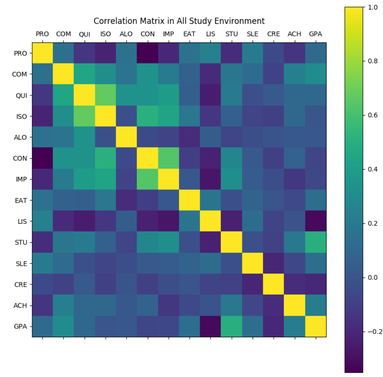
RESULTS

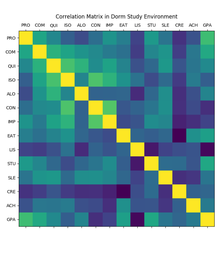
The survey collected 47 effective responses and 4 successful interviews. The form below shows classified results errors for all responses with 9 originally scaled parameters:

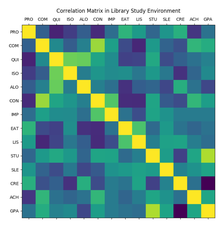
Classified Results Errors with SGD on Different Variables

	Proximity	Comfortable	Quiet	Isolation	Alone	Concentration	Improvement	EatDrink	Listen
Explained Variance of Training Set	0.1357	0.0263	0.0678	0.1563	0.0193	0.0674	0.0597	0.2230	0.4375
Mean Absolute Error of Training Set	0.2105	0.1622	0.2439	0.3243	0.1389	0.2188	0.2353	0.3056	0.4375
Mean Squraed Error of Training Set	0.2105	0.1622	0.2439	0.3243	0.1389	0.2188	0.2353	0.3056	0.4375
Explained Variance of Test Set	0.1111	0.0900	0.0278	0.2000	0.1570	0.1644	0.0769	0.2810	0.8000
Mean Absolute Error of Test Set	0.1111	0.3000	0.1667	0.4000	0.4545	0.4667	0.2308	0.7273	0.8000
Mean Squraed Error of Test Set	0.1111	0.3000	0.1667	0.4000	0.4545	0.4667	0.2308	0.7273	0.8000

The correlation matrix below proposed that in fact among all study environments, as long as places offer good isolation, quietness that facilitate them concentrating, their GPA is thus improved. Students who spend much time to study every week can presumably obtain excellent performance. They tend to stay in places they are familiar with, so there's no much difference between library and dorm environments.

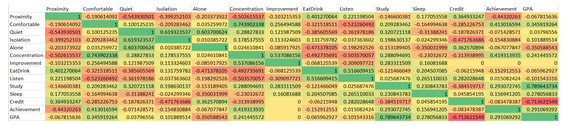






However, in these two separate matrices in dorm and library environments, we can clearly see students study in libraries have high correlation with many variables. Such results suggest that for dorm environment students, their improvement is mainly from proximity and time spend on study. In contrast, students study in library expect this place provide them comfortableness, isolation, concentration and thus boosted their GPA if they study alone. They also spend many hours on study (correlation is as high as to 0.7896!) has positive influences on their success, too.

Numerical Representation for Library Matrix

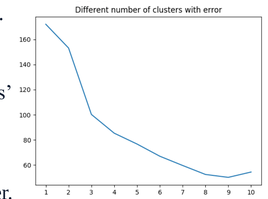


In library matrix, the more green color the block is, the higher correlation it is between two variables. Some red blocks reveal that they even have negative correlation, especially between credit and GPA. This points out that students take many credits in a semester may perhaps perform poorer.

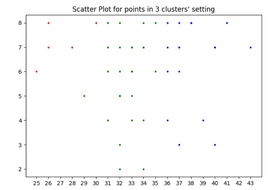
RESULTS (cont.)

	ALL	Place	Proximity	Comfortable	Quiet	Alone	Concentration
MSE	0.664094208	2.077220113	2.297730674	1.996855631	2.097579943	1.907254364	2.567456902
MAE	0.109883757	1.177919415	1.204407894	1.085352507	1.139719376	1.032742337	1.258458369
	Improvement	EatDrink	Listen	Study	Sleep	Credit	Achievement
MSE	2.358388758	2.056337962	2.978716929	3.317262612	2.043725305	1.337237017	2.577729155
MAE	1.19248325	0.969849944	1.030555842	1.297421431	1.136422088	0.719832168	1.208857244

This is the linear regression results of test set within 13 variables so as to estimate students' GPA scope. With all variables the model gives pretty good prediction. Place, proximity, quiet and improvement also supply favorable forecast and consistency on data.



The two graphs at right are clustered results. Trained model signifies a big elbow at 3 clusters' setting, which affords the best result. The bottom one is sum of scaled nine variables and their GPA section: the higher the better. Red dots seems be a legitimate answer to illustrate even students under unfavorable environments, they can still receive at least a scale of 5 GPA.



In four individual interviews, we identified several key factors that promote prominent GPA (from high to low): brightness, quietness, big tables, source of concentration (motivation), atmosphere, comfortableness, proximity.

CONCLUSIONS

Consequently, the purpose of this study is to inspect carefully whether study environments influence GPA or not. The research insinuates that study environments don't alter students' overall performance substantially, it's more related to own study habits and conformable practices. Assembled results project justifiability in the dataset, nevertheless, because of inadequate size of responses, this may lead to certain deviation. However, it's yet appropriate for students studying in library and gain concentration from personal determination, one should achieve better and better achievements as time goes by. Furthermore, with the assist of latest machine learning algorithms, in such novel question, it's capable of discussing it in a brand new intensity.

REFERENCES

Influence of Learning Environment on Students' Academic Achievement in Mathematics: A Case Study of Some Selected Secondary Schools in Yobe State – Nigeria, 2015

The Quality of Learning Environment and Academic Performance from a Student's Perception, 2009

Apache Spark Machine Learning Library, pandas Python Data Analysis Library, scikit-learn Machine Learning Library

ILLINOIS