Features related to Students' Academic Performance: a Prediction about the Grade of Portuguese Language Course

Kewen Ding, Rui Zhu

1. Introduction

Education is closely related to everyone, and grade is a factor that cannot be ignored when it comes to education quality and student's performance. How a student's behavior and the environment he or she is in influence a student's grade? In our project, 30 factors that may explicitly or implicitly play roles in a student's grade are taken into consideration by applying different algorithms to how and to what extent it has an impact on the grade. In addition to that, predictions about the grade and if a student can pass the course will be made, each followed by an accuracy assessment.

The input to our algorithm is {school, higher pursuit, studyTime, gender, etc...}, then {linear regression, neural network, and SVM} three algorithms will be applied to output a student's Portuguese grade and if a student can pass this course. Besides, features are further selected during the process of linear regression to gain better performance.

2. Related work

What affects students' grades? Studies consistently have found a negative bivariate association between alcohol use and academic performance among college students.[1] From another research, alcohol had indirect effects on sleepiness and GPA, primarily through its effect on a sleep schedule.[2] However, many other factors also play a role in a student's grade. Is it better to include as many features as possible when predicting the students' performances? One study compares the performances of different feature sets and found there is a significant performance difference of feature selection algorithms using the datasets with different numbers of features; shows 10 to 20 percent difference in accuracy percentages. The performance of the filter feature selection techniques reduces as the number of features increases.[3] So it is important to choose features wisely.

Nowadays many techniques have been used in this field. Linear Regression, Decision Tree and Artificial Neural Network are common methods to build the classification model. [4] [5] In another experiment, K mean clustering algorithm [6]was applied. All these algorithms perform well, some of which reached an accuracy of 97%. However, these algorithms consider only 20 attributes and are very advanced for beginners. How will the basic and relatively primitive algorithm perform? Will the accuracy increase when more features are added? This is the main point that will be focused on in this project.

3. Dataset and Features

There are 649 training examples from the dataset. According to the rate of 60%, 20%, 20%, the corresponding numbers of training, cross validation and test set examples are 390, 130 and 129. Data were preprocessed by normalization to ensure that the gradient descent moves smoothly towards the

[1] Singleton, "Collegiate Alcohol Consumption and Academic Performance."
[2] Singleton and Wolfson, "Alcohol Consumption, Sleep, and Academic Performance Among College Students."
[3] Zaffar et al., "A Study of Feature Selection Algorithms for Predicting Students Academic Performance."
[4] Alsalman et al., "Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance."
[5] Ibrahim, "PREDICTING STUDENTS' ACADEMIC PERFORMANCE: COMPARING ARTIFICIAL NEURAL NETWORK, DECISION TREE AND LINEAR REGRESSION."
[6] Oyelade, Oladipupo, and Obagbuwa, "Application of k Means Clustering Algorithm for Prediction of Students Academic Performance."

minima and that the steps for gradient descent are updated at the same rate for all the features.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The dataset is obtained from the website https://www.kaggle.com/uciml/student-alcohol-consumption It contains 30 basic attributes of a student as the input and his or her first period, second period and final grades for Portuguese language courses as the output. The basic attributes include school, sex, age family size, parents' education, etc. All these attributes can be numerical or boolean, by classifying them into values of one to five or summarizing into five groups for example. For the output, we take the mean value of the three grades to obtain the average grade, representing the overall academic performance.

4. Results and Discussion of different Methods

Linear Regression:

In statistics, **linear regression** is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables.[7] Suppose vector X is the input and scalar y is the output, linear regression is trying to find a vector θ to predict y from X by the equation:

$$h_\theta(x) = \theta_T x = \theta_0 + \theta_1 x_1$$

$h_\theta(x)$ is the prediction value of output y. To fit y better, different methods can be applied to train the θ.

Gradient Descent and Normal Equation will be mainly focused on as the first step. We will go through all features to analyze if a feature can fit the data appropriately. One characteristic of our dataset is that nearly half of the features are non-numeric types. To transfer them, categorical() is used to transfer two-value types like 'yes', 'no', 'female', and 'male' to 0 and 1s. features with more than two values are transferred by grp2idx() by randomly assigning different values to different words. These two methods are used respectively to determine the weight (θ) for each feature. After that, all the weights will be sorted from the largest to the smallest. As a result, a general idea of which features are of great importance can be obtained as shown below:

Table 1. Theta value for different attributes

feature	theta	feature	theta
constant	11.625	Fjob	0.030
school	0.585	traveltime	0.001
higher	0.489	famsup	-0.021
studytime	0.302	goout	-0.034
sex	0.259	nursery	-0.082
Medu	0.185	freetime	-0.087
Fedu	0.179	paid	-0.097
famsize	0.137	guardian	-0.100
age	0.136	Walc	-0.122
activities	0.114	romantic	-0.153
address	0.099	health	-0.171
reason	0.099	Dalc	-0.186
famrel	0.089	absence	-0.237
internet	0.079	schoolsup	-0.374
Mjob	0.067	failures	-0.694
Pstatus	0.056		

The absolute value of 0.1 is used as a threshold to select strongly-related features. As a result, the first 11 and last 9 features are chosen and formed a new dataset. From what can be seen from Table 1, school, higher pursuit and study time with weights equal to 0.59, 0.49 and 0.3 play vital roles in students' grades, followed by gender and parents' education. On the contrary, the number of failures, absences and workday alcohol consumption influence grades in a negative way with -0.69, -0.24 and -0.18. Remarkably, 'schoolsUp' which means extra educational support weighs -0.37.

In order to improve the linear regression, not only X should be considered, X^2, X^3 ...should also be taken into account. In that case, overfitting can occur. To prevent overfitting, lambda is included to penalize high order terms. Here, a list of lambdas and a set of models are selected. By iterating in these two loops, a collection of thetas and costs are acquired so that we can find the most promising lambda and model as shown below:

[7] "Linear Regression."

Table 2 . Error values of linear regression with different lambdas and different high order terms

lambda\p	p=1	2	3	4	5	6
0	3.19	3.03	3.01	3.25	3.27	3.26
0.001	3.19	3.03	3.01	3.24	3.27	3.27
0.003	3.19	3.03	3.01	3.27	3.26	3.27
0.01	3.19	3.03	3.01	3.25	3.26	3.28
0.03	3.19	3.03	3.01	3.25	3.26	3.27
0.1	3.19	3.03	3.01	3.25	3.25	3.25
0.3	3.19	3.03	3.01	3.23	3.23	3.24
1	3.19	3.03	3.00	3.18	3.19	3.20
3	3.18	3.03	2.98	3.12	3.13	3.15
10	3.15	3.02	2.97	3.03	3.05	3.08
30	3.08	3.01	2.97	2.99	3.00	3.02

The table indicates that the cubic model with lambda = 10 pair gives the smallest cost. Consequently, this setting will be used in the final test. Applying the chosen lambda and model (p) pair, the error value of the test set is 3.25, which is bigger than that of the same setting in the table, but still acceptable compared with other data. Thus a capable combo has been obtained for linear regression.

Neural network:

To improve the precision of linear regression, more features such as polynomial features can be added, but that can be very expensive to train. The neural network will be used here, because of its competence in representing complex models that form non-linear hypotheses.

For the input layer, all 30 features are used as input units. One hidden layer is used as a default. For the number of hidden units, at first, 30 units, which means one time of the input units is chosen but then it shows from the accuracy test that a satisfying prediction can not be obtained. But two times of the input units can take too much time in the calculation. Therefore, 50 units have been applied to our hidden layer. Since the Neural Network can only take care of the logistic problem, this linear regression problem has been transferred to a classification one. The prediction of student's grades now has 20 outputs indicated from one to twenty (zero grade does not exist in this dataset). Then the whole design of our neural network system has been finished.

During the whole progress of coding and testing, it becomes obvious that if we only count the prediction which is exactly the same as the true value, the accuracy from the cross validation test can be very low. That's where we start to consider a range of grades: The result will be composed of three parts except the training set: The first one is where the range is 0, which means the prediction is supposed to be the same as y. The second is from minus one to plus one (e.g. if ten is the true value, nine and eleven are also right), and the last one is from minus two to plus two. Accordingly, a more tolerant testing system is formed and thus we can have a better overview of its performance.

The result from our coding is as followed:

Table 3 . Accuracy of training set with 500 and 1000 number of iterations

lambda	500(nrIter)	1000(nrIter)
0	83.5%	100%
1	63.5%	80%
2	47.4%	52.1%
4	29.2%	31.3%
8	23.6%	23.3%
16	21.8%	21%

Table 4 . Accuracy of validation set with 500 and 1000 number of iterations

lambda	500(nrIter)	1000(nrIter)
0	15.4%	18.5%
1	16.2%	18.5%
2	14.6%	16.9%
4	19.2%	17.7%
8	20.0%	20.0%
16	14.6%	15.4%

Table 5 . Accuracy of validation set with 500 and 1000 number of iterations. (predicted value will be considered correct if its difference with the true value is less or equal to one)

lambda	500(nrIter)	1000(nrIter)
0	36.1%	40%
1	39.2%	35.4%
2	41.5%	43.1%
4	46.9%	49.2%
8	51.5%	52.3%
16	48.4%	48.5%

Table 6 . Accuracy of validation set with 500 and 1000 number of iterations. (predicted value will be considered correct if its difference with the true value is less or equal to two)

lambda	500(nrIter)	1000(nrIter)
0	57.7%	54.6%
1	62.3%	63.1%
2	63.8%	65.4%
4	70.0%	70.8%
8	71.5%	71.5%
16	71.5%	73.1%

The result can be analyzed from three dimensions:

To start with, it can be seen from the horizontal comparison that in most cases, 1000 iterations will produce a more accurate prediction than that of 500 times. That is because more training means a smaller cost. But on the other hand, when the time of training is too many, this prediction function tends to overfit especially when lambda is not big enough like in the comparison when lambda is four and the range is zero.

Following that, the vertical comparison reveals that the accuracy for the training set decreases dramatically as lambda increases, which is as expected since lambda is responsible for the regularization. Lastly, the range added in this function can be regarded as a way of generalization as can be seen from this table, the accuracy 'improves' a lot with the broadening of the range. What is remarkable is that with lambda equals to 16 and 1000 times of iterations, the accuracy with the biggest range increases up to 73.1%. Lambda equals eight with 1000 times iteration as the most promising pair in range ±1, from which the test set accuracy is 45.7%. This result is lower than that of the cross validation set but still acceptable, just like what we have in the linear regression part.

It is shown in the table that if no tolerance is given in the value range, the highest accuracy that can be attained is 20%. This is not surprising considering that the students' Portuguese grades with these 30 general features are not that strongly related. This is also the reason why we are using an additional range in the neural network.

SVM:

Support-vector machines (SVMs), are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.[8]

All information about students is considered as input. The students' grades are divided into two classes: pass or fail, corresponding to one and zero respectively. In the SVM algorithm, in the case of a nonlinear boundary, the Gaussian Kernel was chosen instead of Linear Kernel. During the test the Gaussian Kernel performs better than the linear kernel when comparing the errors of the predicted value.

$$Kgaussion(x^{(i)}, x^{(j)}) = exp(-\frac{|x^{(i)}-x^{(j)}|^2}{2\sigma^2})$$

It is important to choose a pair of C and sigma or the gaussian kernel. To find the most promising pair, different value pairs can first be applied to model training with the training set, then compute the error value of each model with the cross validation set. From all the errors obtained above, the smallest error value and its corresponding C and sigma pair can be found.

[8] "Support Vector Machine."

In this experiment, C and sigma was chosen from [0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 90]. Below are the error values obtained from different (C, sigma) combo. When C is 0.03 and Sigma is 0.15, the error becomes the smallest as 0.15. Based on the model, the accuracy is 81.39% from the test set.

Table 7 . Average error values of different SVM models, with different C and Sigma combo

C\Sigma	0.03	0.1	0.3	1	3	10	30	90
0.03	0.30	0.32	0.27	0.27	0.20	0.18	0.25	0.15
0.1	0.30	0.32	0.27	0.27	0.20	0.18	0.22	0.29
0.3	0.30	0.32	0.27	0.27	0.20	0.18	0.18	0.70
1	0.30	0.32	0.27	0.27	0.19	0.18	0.18	0.70
3	0.30	0.32	0.27	0.27	0.21	0.18	0.18	0.70
10	0.30	0.32	0.27	0.27	0.20	0.23	0.18	0.70
30	0.30	0.32	0.27	0.27	0.20	0.28	0.18	0.70
90	0.30	0.32	0.27	0.27	0.20	0.29	0.18	0.70

5. Conclusion and Future Work

By applying three algorithms to the dataset, a general idea of how every feature weighs in the prediction of the students' grade. Since the SVM part is used for predicting if a student can pass the exam, only the first two algorithms are considered. From what we have obtained of the test set accuracy, the neural network has a better performance with 45.7% for range ±1 compared with that of linear regression with 33.3% for the same range. That is understandable since the neural networks perform better when nonlinearities are involved, which is exactly the case in our dataset since all these features are too general for Portugeses course thus can hardly be linear.

For the future work, now that only Portuguese grade is focused on, we will try to make different prediction systems to predict different kinds of courses based on what we have now. That requires more specific features for each course, which takes more time to be looked into.

6. Contributions:

Most of the work was done together, like choosing the topic, training the algorithm and writing and finalizing the report and representation. For the training part, different team members focus on different parts. Kewen worked on linear regression, and Rui worked on the neutral network and SVM algorithms.

7. Reference:

Alsalman, Y. S., N. Khamees Abu Halemah, E. S. AlNagi, and W. Salameh. "Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance." In *2019 10th International Conference on Information and Communication Systems (ICICS)*, 104–9, 2019. https://doi.org/10.1109/IACS.2019.8809106.

Ibrahim, Zaidah. "PREDICTING STUDENTS' ACADEMIC PERFORMANCE: COMPARING ARTIFICIAL NEURAL NETWORK, DECISION TREE AND LINEAR REGRESSION," 2007, 7.

"Linear Regression." In *Wikipedia*, November 29, 2020. https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=991230675.

Oyelade, O. J., O. O. Oladipupo, and I. C. Obagbuwa. "Application of k Means Clustering Algorithm for Prediction of Students Academic Performance." *ArXiv:1002.2425 [Cs]*, February 11, 2010. http://arxiv.org/abs/1002.2425.

Singleton, Royce A. "Collegiate Alcohol Consumption and Academic Performance." *Journal of Studies on Alcohol and Drugs* 68, no. 4 (July 1, 2007): 548–55. https://doi.org/10.15288/jsad.2007.68.548.

Singleton, Royce A., and Amy R. Wolfson. "Alcohol Consumption, Sleep, and Academic Performance Among College Students." *Journal of Studies on Alcohol and Drugs* 70, no. 3 (May 1, 2009): 355–63. https://doi.org/10.15288/jsad.2009.70.355.

"Support Vector Machine." In *Wikipedia*, November 28, 2020. https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=991142258.

Zaffar, Maryam, Manzoor Ahmed, K.S. Savita, and Syed Sajjad. "A Study of Feature Selection Algorithms for Predicting Students Academic Performance." *International Journal of Advanced Computer Science and Applications* 9, no. 5 (2018). https://doi.org/10.14569/IJACSA.2018.090569.