The impact of social features upon academic performance. Multilayer Perceptron versus Support Vector Machines

Andra Condurache
Departament of Computer Science

Abstract

The aim of this paper is to predict marks for Mathematics and Portuguese Language exams in secondary school. This investigation will focus on evaluating the results obtained using two supervised learning algorithms, Multilayer Perceptron and Support Vector Machines, which are aiming to solve a regression problem. Both of the models use KFold Cross Validation and Grid Search as optimisation techniques. This study also includes an evaluation of these two different approaches focusing on R-squared values and learning curves.

1.Introduction

Academic performance has been heavily influenced by current global situation. Every country tried to implement their own methods to carry on academic activities and, more importantly, to organise final exams or to predict pupils' marks, taking into consideration previous academic performance. So much has changed for education in the previous year and we still can not find a solution to facilitate education.

This investigation is willing to offer an overview of the factors that influence the academic performance. The aim of this paper is to outline the importance of social interaction between pupils. We will investigate this problem using two algorithms with two different approaches: Multilayer Perceptron (MLP) and Support Vector Machines (SVM). Both of them are supervised learning algorithms.

There will also be two main questions which will lead this investigation:

- What are the grades most influenced by (social or academic features) and can they be predicted only by academic features?
- What error is produced when social features are not taken into consideration?

2.MLP

A Multilayer Perceptron is one of the most used types of Neural Networks where the data is structured into three types of layers: input, output and hidden. The model will consist of more layers. The input layer will receive the input values (social and academic features), while the output layer will consist of one neuron representing the grade received by a pupil. As the data is structured into more layers, the learning process will occur during the hidden layers which are not directly exposed to the input layers.

The model will require weights and biases correlated with the input values. During the training process the weights will be adjusted based on a specific criterion. Hidden layers perform linear transformation and the weights will be updated using the backpropagation error algorithm. The model trained will solve a regression problem and consequently the mean squared error (MSE) method will be implemented. We will evaluate the model using the root mean squared error (RMSE).

2.SVR

Support Vector Machines are supervised learning models that are usually used to solve classification problems. These models are trained to find a hyperplane in a multidimensional space that with the aim of separating classes. This algorithm consists of maximazing the margin between the data points and the hyperplane. We will use a different approach called Support Vector Regression (SVR) as we will solve a regression problem. The core concept is

similar to SVM but the purpose of the hyperplanes is to set the maximum error we will accept for finding an appropriate line that will fit the data.
SVM method uses less information for the calculations comparing it with MLP method.

2.Dataset

The dataset used contains information regarding academic performance, demographic and social features. The data was collected by using school reports and questionnaires. There are two datasets, first of them contains the results at Mathematics exams and the second one contains Portuguese language results. The process of gaining information was conducted by P. Cortez and A. Silva during 2005 and 2006 in two public schools from the Alentejo region of Portugal[1]. The Mathematics results dataset contains 395 rows and the Portuguese language results dataset contains 649 rows, both of them have 32 attributes.

We will split the data into two categories: academic features and social features. These two categories can be also spilt as subjective features and objective features. We will implement different models taking into consideration social features as the aim of this investigation is to raise the awareness of the human interaction. The splitting the attributes into these two categories purely results as a consequence of my personal analysis. Therefore, errors may occur.

2.1 Properties of Data and Data Analysis

The education system in Portugal is divided into three levels: pre-school education, basic education and secondary education. We will analyse the marks received during secondary education. This level consists of three years of schooling and pupils are usually 15 to 17 years old. The dataset contains the exam results for core courses (Mathematics and Portugueses Language) for every academic year listed asta G1, G2 and G3. The marks range from 0 to 20, where 20 is the highest mark and 10 is the passing mark.

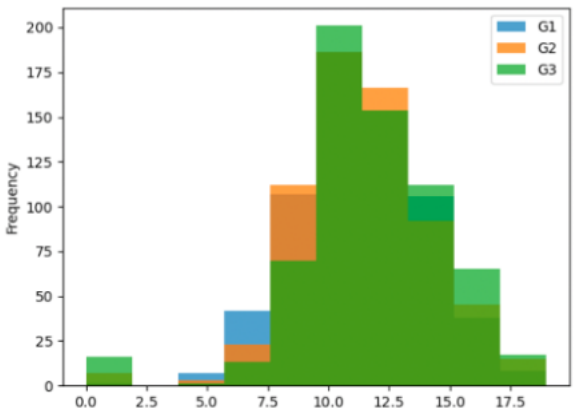

Fig 1 – Portugueses Language

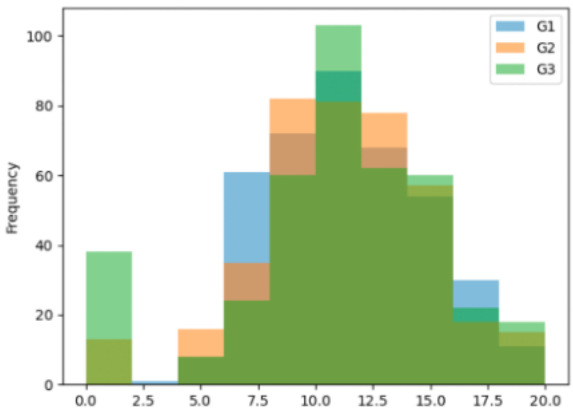

Fig 2 – Mathematics

The previous figures show a normal distribution for both of the courses. Fig 1 shows that the Portuguese language results are slightly right-skewed. Even though average grades for Portugueses Language exams are higher than Mathematics exams, figure 2 shows Mathematics results are more diverse. We can clearly see there are many perfect marks comparing it to the figure 1.

3. Methods

This section will focus on the computational methods used for the investigation. We will discuss different arhitectures and hyperparameters that were implemented and we will choose a suitable approach for every model.

3.1 Computational Approach

There are two datasets: one of them contains information about Mathematics exams and the other one contains information regarding Portuguese language exams. We will train and test the models for each of them. Also, for every dataset we will split the features and we will and test the models again to outline the differences between the results. Thus, there will be four configuration analysed using two algorithms.

We will split the data into training data (90%) and testing data (10%). This implementation is done using KFold cross-validation for better accuracy. The entire data will be split into 10 groups of samples using sklearn model_selection. The k value should not be too small or too high. A high k value may produce overfitting. We will select the best model after testing and training all the samples.[3]

All the features will be used for the first part of the investigation and we will test different arhitectures. For the second part of the investigation we will remove the social features and we will train the models using the best configurations found during the first phase. The aim of this comparison is to outline how the removed features influenced the output value even for the best model in the first phase.

3.2 Parameters for MLP

We want to find which weights and biases minimize a certain the cost function. The model started with random weights and biases, it was trained using the stochastic descent optimization algorithm and the weights were updated using the backpropagation error algorithm. I used sigmoid function and ReLU function as activation functions. The model consists of two hidden layers. Same accuracy in both of the models means we don't need more layers because the arhitectures performs as well as the less deeper version.

The output layer consists of only one neuron which represents the predict mark for G3 exam. This is specific to regression problems that predict only one value. Neural networks that solve classification problems have a different structure for the output layer. These have one neuron for every class.

The model solved a regression problem and consequently the mean squared error (MSE) method was implemented. Also, we calculated the root mean square error (RMSE) for a better understanding of how the residuals are spread out. MSE and RMSE tell us how close the predicted values are to the real values. As these values don't describe exactly the accuracy of the model, we will calculate the R-squared value for a fair comparison with the SVM algorithm.

MLP models require a lot of hyperparametres for a good prediction comparing to the SVR algorithms. The hyperparametres strictly realated to the network were number of hidden layers, number of neurons for every hidden layer and activation functions. The other hyperparameters were learning rate, momentum and number of epochs. To avoid the overfitting I varied the last three hyperparametres. I set the number of epochs to 30 to avoid premature convergence. A high learning rate would also lead to unrealistic results. Best results were obtained using low learning rates (0.001 and 0.005). It is important to vary momentum to avoid local minima.

3.3 Parameters for SVR

SVM models set decision boundaries to separate classes. A decision boundary maximizes the distance from the nearest data points of all classes. However, this method can be also applied to regression problems. The objective is to find hyperplanes and boundary lines.

SVR models use less parameters than the MLP models. The most important parameter is the kernel function which will map the data from a dimension into another dimension. We will use the hyperplanes to fit the marks. SVRs will help us to fit the error within the certain threshold.

We will vary the parametres for C, gamma and epsilon using Grid Search and we will calculate the R-squared value. [2]

4.1 Experimental Results

There are enormous differences between MLP and SVR ways of learning. MLP models start with random weights and this implies huge train loss during the first epochs. K-Fold Cross-validation played an important role in obtaining better results. MLPs would perfom very poor (up to 0.80) without implementing K-Fold Cross-validation method. However, some configurations may lead to negative values, which means our model learned nothing and the predicted values don't follow the real values.
SVR models are more sensible at parametres modifications. R-squared values can vary from 0.4 to 0.9 even when the same parametres are used. This comportament is seen during the K-fold cross-validation process which plays a great role in evaluation. Also, better results are obtained using grid search of SVRs. The next table shows the best result obtained for each training process. We will focus only on Mathematics results as the other marks follow a similar trend.

MLP – Mathematics (all features)			
Structure	Learning Rate	Momentu	R-squared
[27, 15, 7, 1]	0.001	0.9	0.884
[27, 15, 7, 1]	0.001	0.5	0.803
[27, 15, 7, 1]	0.001	0.1	0.783
[27, 15, 7, 1]	0.005	0.9	0.925
[27, 15, 7, 1]	0.005	0.5	0.896
[27, 15, 7, 1]	0.005	0.1	0.794
[27, 15, 7, 1]	0.01	0.9	0.883
[27, 15, 7, 1]	0.01	0.5	0.904
[27, 15, 7, 1]	0.01	0.1	0.902
[27, 20, 12, 1]	0.001	0.9	0.886
[27, 20, 12, 1]	0.001	0.5	0.806
[27, 20, 12, 1]	0.001	0.1	0.795
[27, 20, 12, 1]	0.005	0.9	0.770
[27, 20, 5, 1]	0.001	0.9	0.886
[27, 20, 5, 1]	0.005	0.9	0.960
[27, 20, 5, 1]	0.005	0.5	0.898
[27, 15, 7, 1]	0.1	0.1	-0.021
[27, 15, 7, 1]	0.1	0.9	-0.021
Mathematics (without social features)			
[19, 12, 6, 1]	0.005	0.9	0.56
[19, 10, 5, 1]	0.005	0.9	-0.22
[19, 15, 7, 1]	0.005	0.9	0.881

Table 1 - MLP results

SVR – Mathematics (all features)					
Kernel function	C	gamma	epsilon	Polynomial order	R-squared
rbf	1	0.1	0.1	-	0.263
rbf	10	0.1	0.1	-	0.527
rbf	100	0.1	0.1	-	0.527
Rbf	1000	0.1	0.1	-	0.527
rbf	1	0.01	0.1	-	0.804
rbf	10	0.01	0.1	-	0.843
rbf	100	0.01	0.1	-	0.812
rbf	1000	0.01	0.1	-	0.724
rbf	1	0.001	0.1	-	0.912
rbf	10	0.001	0.1	-	0.828
rbf	100	0.001	0.1	-	0.826
rbf	1000	0.001	0.1	-	0.823
Linear	1	0.1	0.1	-	0.811
linear	10	0.1	0.1	-	0.811
linear	100	0.1	0.1	-	0.811
linear	1	0.01	0.1	-	0.811
linear	10	0.01	0.1	-	0.811
linear	100	0.01	0.1	-	0.811
poly	10	0.01	0.1	3	-0.1
poly	10	0.1	0.1	3	-0.49
poly	1	0.1	0.1	4	-4.4
poly	10	0.1	0.1	4	-4.4

Table 2 - SVR results

4.2 Algorithm Comparison

Figure 3 and 4 show how the MSE values decrease during the epochs. This trend is observed for all MLP models the configurations. Training and testing a MLP model by simply splitting the dataset lead to good results but these can be improved. Implementing a KFold Cross Validation method helps the research. Using the first method the MSE values range from 120 to 14 but it never decreases to lower values, a KFold Cross Validation implementation helps the model to vary from 120 to 1.5-2. It is specific to MLP to have huge train loss at the beggining due to random its random weights, but these values improve rapidly. [4]

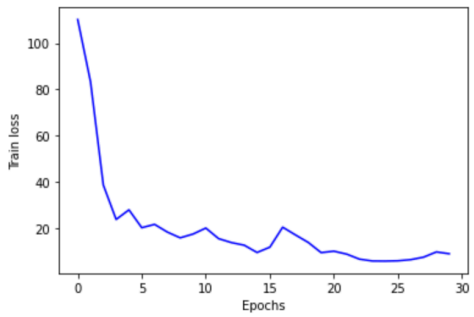

Fig 3 - Train Loss without Kfold CV

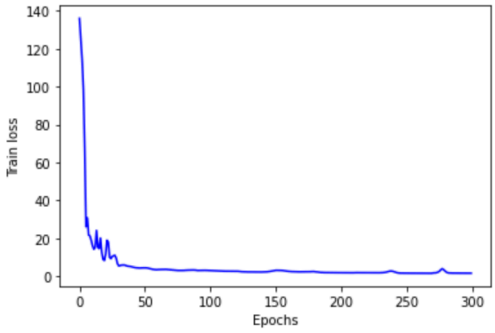

Fig 4 - Train Loss with KFold CV

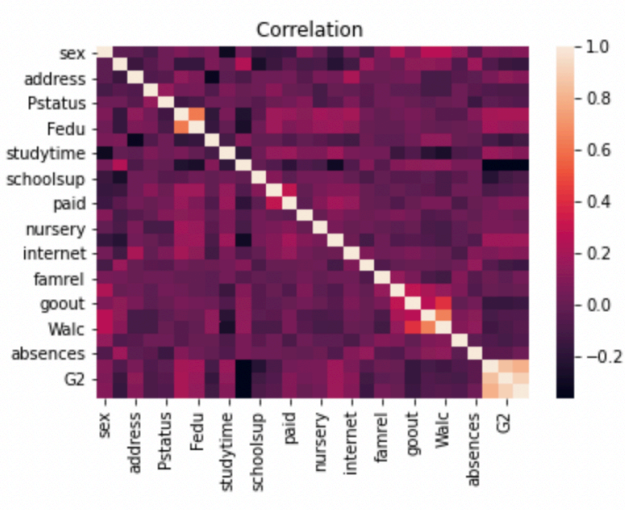

Fig 5 - Correlation Matrix

We have seen in the previous section 4.1 Experimental Results that removing the social features affect the predicted values. This shows shows the importance of them. Also, less features could lead to less accurate. Thus, we can not conclude the predictions were only influenced by this. However, figure 5 show there is a strong correlation between social features like going out, free time or alcohol consumption. [5]

Figure 6 shows some of chracteristics of the SVR model. This behaviour is similar for all the models which have tested during the investigation. Even though, MLP models have better results, SVR models still predict the marks with a high accuracy, having the R-squared values up to 0.9. We can see in the first graph that the cross validation scores increases from 0.6 to almost 0.85 during 300 epochs which is a good improvement. Also, the performance of the model shows an ascendent trend.

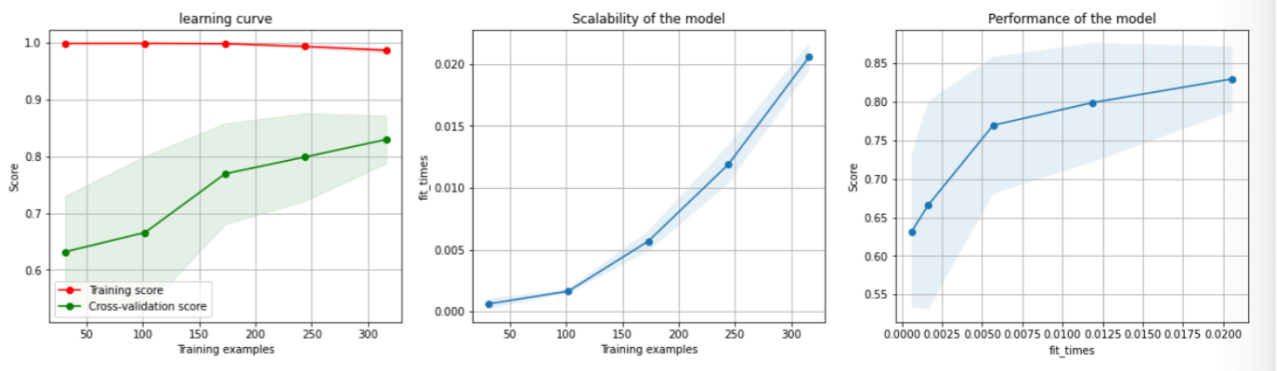

Fig 6 - Performance SVR(kernel='rbf', gamma=0.01, C=100, epsilon=0.1) code using sklearn[6]

5. Conclusion

This investigation has focused on solving a regression problem using two different algorithms. Both of them are supervised learning algorithm but they follow totally different approaches.

Both of MLP models and SVR models had good results (R-squared values ranging from 0.75 to 0.92 for best configurations) especially for human behavior predictions. It is still considered that predicting human behaviour is more difficult than predicting physical processes[7], especially predicting marks has always been a sensitive topic. The MLP approach showed better results and KFold Cross Validation played a huge role in its optimization process. As for the SVR model, Grid Search has been an important factor in getting better results.

We have already seen that these methods can predict marks but these can be improved. The data collected by P. Cortez and A. Silva is considered a relatively small amount of data. Also, this data might or might not be biased. Implementing a genetic algorithm is a good optimisation technique which would increase substantially the performance and would definitely help to avoid premature convergence and overfitting due to its heuristic properties relying on mutation, crossover and selection.

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[2] https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[3] Sanjay. M, Why and how to Cross Validate a Model? https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

[4] Useful Plots to Diagnose your Neural Network, https://towardsdatascience.com/useful-plots-to-diagnose-your-neural-network-521907fa2f45

[5] Correlation matrix, https://www.displayr.com/what-is-a-correlation-matrix/

[6] https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py

[7] How To Interpret R-squared in Regression Analysis, https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Appendix 1 – glossary

Sigmoid function - it decides what value to pass as output and map the data in (0,1) interval.

ReLU function - it decides what value to pass as output and pick max(0,x) having a better convergence than sigmoid

Grid Search - a searching technique aiming to tune the hyperparameters by searching exhaustively through a list of given parameters

KFold Cross Validation - a technique that randomly spilts the entire data into K number of samples for a better evaluation

Support Vector Regression - a method derived from Support Vector Machines that uses hyperplanes to predict values

Multilayer Perceptron - a type of neural network which has 3 types of layers (input, hidden and output) and it uses backpropagation algorithm

Backpropagation - an algorithm that adjust model's weights and biases

Learning Curve - uses an estimator to determine cross-validated training and test scores

Appendix 2 – Implementation details

Some of the configurations of the MLP models lead to negative results which means out models has learned nothing during the training process and its predictions don't fit the data. MLP models' structures played an important role, especially the activation functions. The models are very sensitive when it comes to activations functions.