

Sleep quality analysis

Arturo Laflor

2017-05-08

Contents

1	Prerequisites	5
2	Introduction	7
3	Data adquisition	9
4	Data pre-process	11
5	Feature selection	13
6	Evaluation of Efficiency	15
6.1	Neural Networks Results	16
6.2	Logistic Regression Results	17
6.3	Supprot Vector Machine Results	18
7	Applications	21
8	Placeholder	23
9	Final Words	25

Chapter 1

Prerequisites

Chapter 2

Introduction

Chapter 3

Data acquisition

Chapter 4

Data pre-process

Chapter 5

Feature selection

Chapter 6

Evaluation of Efficiency

Before testing the selected factors, models were trained using the 21 sleep hygiene factors to know the predictive efficiency that these models from various techniques of machine learning could achieve. The result was that both, the support vector machines (SVM) with linear kernel and logistic regression, were the two techniques with the best results. The SVM algorithm had an efficiency of 67% and the logistic regression reached an efficiency of 70%. With this background, the tests described below were made, taking into account only the four selected factors. If any of the techniques reaches an efficiency equal to or higher than the previous results, the selection of variables can be considered a successful process and these factors will be used for the prediction model of the study hereafter.

One of the steps in the development of the investigation project, includes the selection of a technique to train a predictive model on supervised automated learning. We did a review of the literature and we select three techniques under certain criterion based in the nature of the problem. The purpose is train the model with the available data and select the one given the best prediction. So, at the same time that the evaluation of efficiency of the selected factors was performed, the selection of the technique that will be used for the final training was done. The three techniques that meet the inclusion criteria, were: artificial neural networks, vector supported machines And logistic regression with regularization. As in feature selection, a Shiny application was developed to process the data and compare the outcomes for these three algorithms, training a model with total of the records and only the four features selected in the feature selection process as was explained in Section 5.

The evaluation was performed by the cross validation technique using an iteration process of training, validation, analysis and refinement as the figure 6.1 shows. In this process a sixty percent of the data was used to train the model, when training conclude, the cross validation is performed through the prediction of the target variable in the cross validation set, containing a twenty percent of the main dataset. The analysis is done at that time and depending on the results, the parameters are adjusted to make a new iteration or reach the stop point. If the stop point was reached, the model is proved in the test set to obtain the final efficiency of the model.

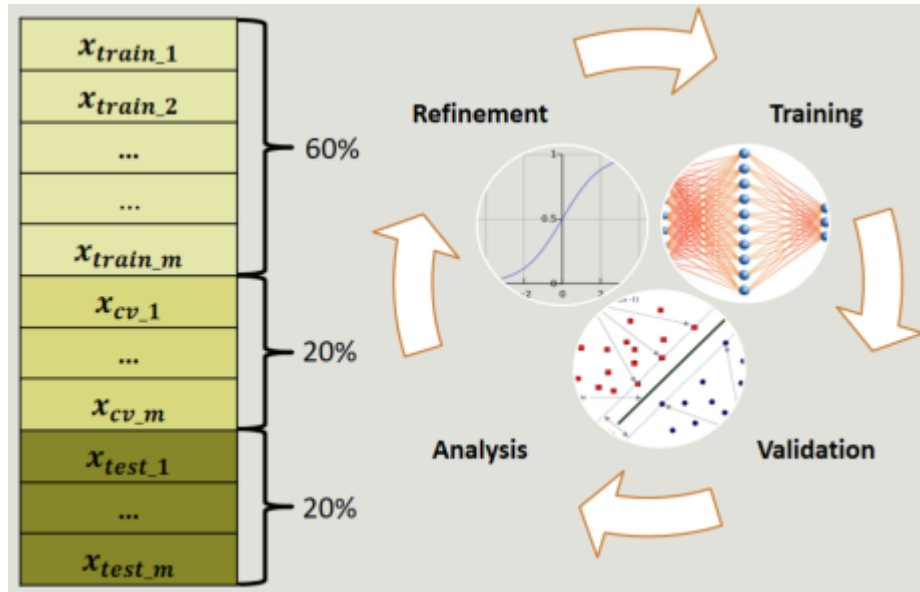


Figure 6.1: Cross Validation Process

6.1 Neural Networks Results

Two neural networks were trained and validated by cross validation process, both structures with a hidden layer. The first neural network had three neurons in the hidden layer and the second four neurons. The Fig. 6.2 shows the structure of the neural network with four neurons in the input layer, one neuron for each factor selected in the feature selection process. The second layer is the hidden layer with four neurons and the last layer contains one neuron for the result (good sleep quality/bad sleep quality). Additionally it is possible to observe the two activation neurons in the top of the figure.

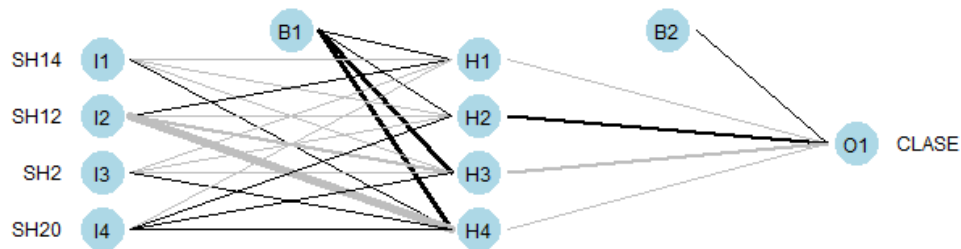


Figure 6.2: Structure of neural network with four neurons in the hidden layer

The results for the two neural networks and the appropriate comparison between them, are in the Fig. 6.3. The network with better efficiency of two networks is the network with four neurons. The table describes that in the three sets, the behavior was superior in terms of efficiency, while the plot represents the error per each set with three and four neurons. Clearly, the lines decrease in favor of the training and validation with four neurons, where the error of the prediction is smaller.

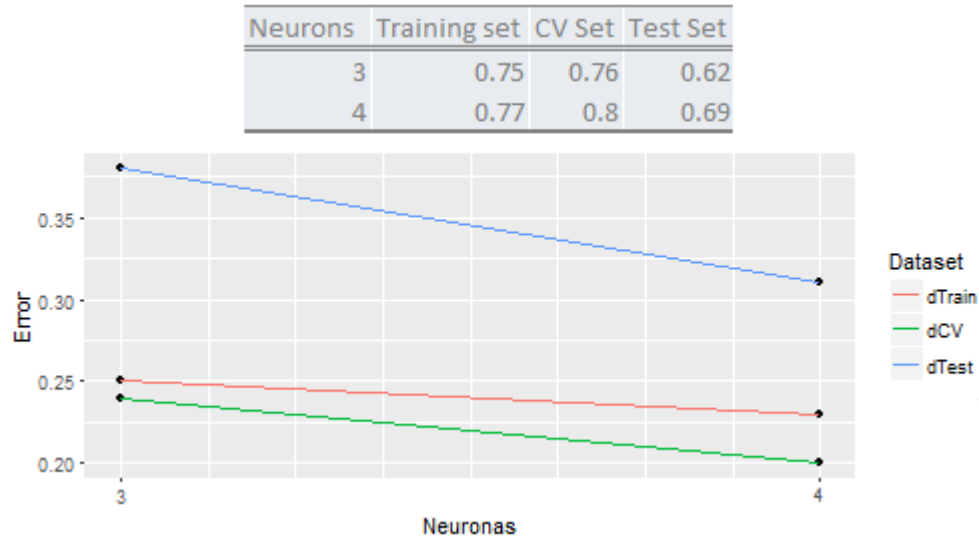


Figure 6.3: Comparison of the results for the two trained neural networks

The results of the neural network, satisfy the conditions sought, because although it is not greatly improved in efficiency when compared to what can be obtained by employing all the factors of sleep hygiene, we gain in the amount of factors that must be sensed to obtain input data. This fact has great relevance for the project because it greatly limits the design and infrastructure of the data acquisition module.

6.2 Logistic Regression Results

We train the model through logistic regression (LR) with regularization parameter and polynomials of degree one, two and three, in order to look for the optimal point between over fit and bias. The regularization parameter based on the norm l_2 takes the form of the equation (6.1), where λ took values from 0.1 to 0.6 with intervals of 0.03 to choose the optimal value.

$$reg = \frac{\lambda}{2m} \sum_{j=2}^n \theta_j^2 \quad (6.1)$$

The stop condition for the adjustment of the parameters of the regression is of the order of one hundred thousandths, that is to say, while the previous and the current cost function did not have a difference of 0.00003 between both, the regression continued to iterate.

The results of LR's are shown by the application in the format of the Fig. 6.4. This figure shows the original results for the LR with the polynomial of degree one, we obtained five coefficients including the intercept coefficient, the right table have the data of prediction, 70% of efficiency for the training set, 76% for the cross validation set and 69% in the test set. The plot in the top of figure, shows the behavior of the cost function through the iterations in the compute and refinement of the parameters.

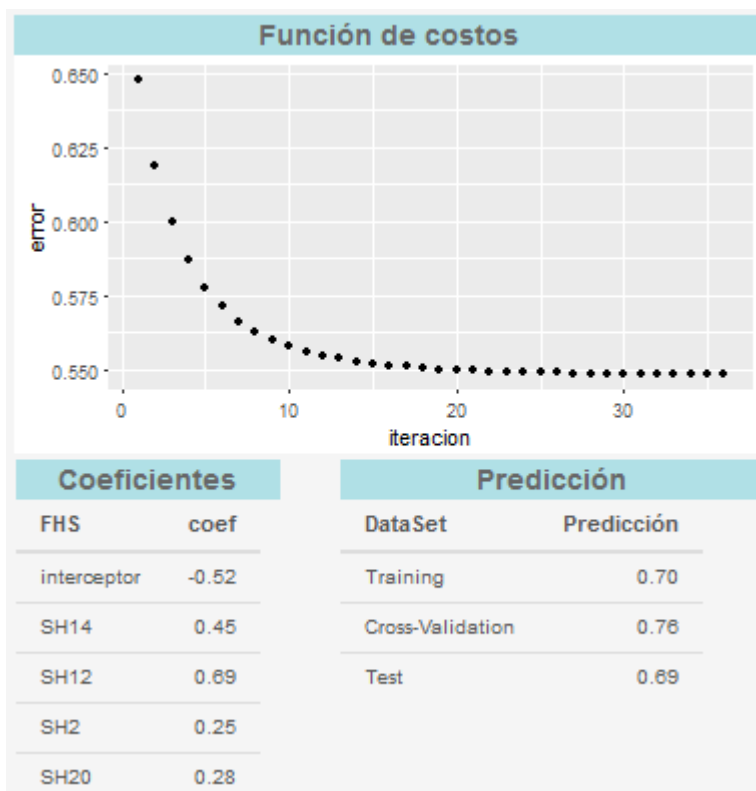


Figure 6.4: Results of the LR and polynomial grade one

For polynomials of degree two and three we have a similar figure, the difference is the number of coefficients that in the case of the polynomial of degree two are 16 and, in the polynomial of grade three are 35, including in both cases *dummy factors*. In the case of polynomial of degree two the cost function iterated 150 times and the predictions were 72% of efficiency for the training set, 78% for the cross validation set and 69% for the test set. The LR with the polynomial of degree three, did 446 iterations, having a precision in the prediction of 75% for the training set, 70% for the cross validation set, falling to 62% for the test set.

Table 6.1: Comparison of efficiency of LR with polynomials of degree one, two and three

	degree 1	degree 2	degree 3
training set	70 %	72 %	75 %
cross validation set	76 %	78 %	70 %
test set	69 %	69 %	62 %

Comparing the three results in the table 6.1, we conclude that the polynomial of degree one is the best choice for this study, because, is the algorithm that consumes the lower resources of the processor and memory and have similar predictions than the other two models of degree two and three. Results, also are satisfactory if they are compared with the results using the 21 input data.

6.3 Support Vector Machine Results

As in the previous algorithms, for support vector machines algorithm, a cross validation test was performed. In this case, were used four kernels, two lineal kernels with polynomials of degree one and two, one radial kernel and one sigmoide kernel. The Fig. 6.5 shows the results as they are presented in the Shiny application,

we can see in the left panel, the plot showing diferents values of C and Gamma parameters and how is the behavior of the error depending of these two parameters. In the right side we observe that the best values for C is 0.04 and the best value for Gamma is 0.5 to reach the best prediction for this kernel, 76% of prediction in the test set.

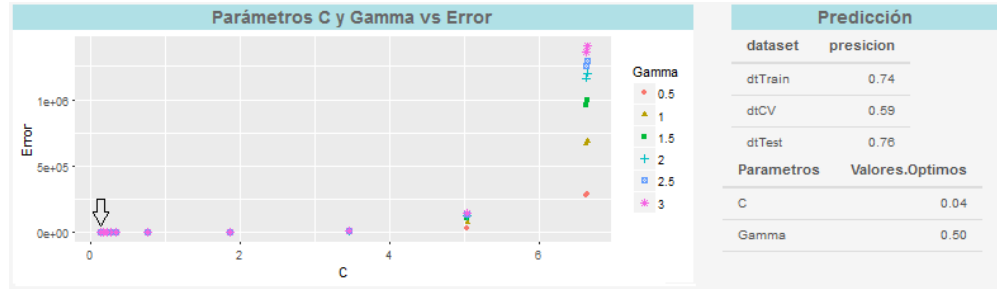


Figure 6.5: Results of SVM with sigmoide kernel

The Table 6.2 show a comparative framework of results of the four kernels that were tested. We can observe that the sigmoid and radial are the best evaluated with a slight advantage of 3 percentage points of the sigmoid over the radial. The linear kernel is not a bad choice if one thinks in terms of simplicity to program it and the little memory and processor that consumes.

Table 6.2: Results of efficiency in prediction with SVM algorithm

	Lineal degree one	Lineal degree two	Radial	Sigmoide
Training dataset	72 %	69 %	80 %	74 %
Cross Validation dataset	70 %	78 %	69 %	59 %
Test dataset	71 %	67 %	73 %	76 %
Parameter C	0.04	0.14	0.40	0.04
Parameter Gamma	0.5	0.5	0.5	0.5

Chapter 7

Applications

Chapter 8

Placeholder

Chapter 9

Final Words

Bibliography