Development of a Divorce prediction model

Author: Joan Prat Sicart

Date: 15th of December of 2019

Contact: joan.prat.sicart@est.fib.upc.edu

This document supposes the deliverable of the project of the subject Data Analysis and Knowledge Discovery corresponding to the Autumn semester of the MIRI master degree during the course 2019/20. **This report is divided according the different steps done to obtain the final model of the project**(see a further explanation in section 3). Moreover, should be accompanied by the corresponding resources containing the scripts required to reproduce the output of this work.

Contents

1 Executive summary

Analysis

In this project it's going to be analyzed the data of a poll fulfilled by 175 different people in a phycological research study done in United States, the poll attempt to understand whether the individual is divorce.

Problem

The problem to be solved consists in predict whether the people will remain married or not in function with the answer they give in this poll.

Solution

The solution is a model capable to predict if there will be divorce or not with the best accuracy possible, to do so, in the project will be studied two methods: Support Vector Machines and Generalized Linear Model.

According to the results, SVM is the **less effective** model with an accuracy of 0.941

Generalized Linear Model is the **best model** with an accuracy of 0.988

2 Introduction

This document explain the different procedures taken into account to create a model capable to predict whether a person will be divorced or not. Moreover, for each methodology the accuracy will be calculated in order to distinguish which is better.

The techniques used in the project are two: Support Vector Machines, since the problem can be understood as a classification problem, and the Generalized Linear Models, because the response variable is binary and therefore we can't use the normal Linear Model.

Furthermore, the dataset used to create the model has 25 attributes, 24 of them represents a numerical answers of questions and the last one, the one named "Class" is binary and specifies if the person was divorce or not, the details of this dataset will be deeply explained in the Data section.

3 Project Flowchart

The different steps done during the project in order to obtain a Final Model as accurate as possible are depicted in the picture 1.

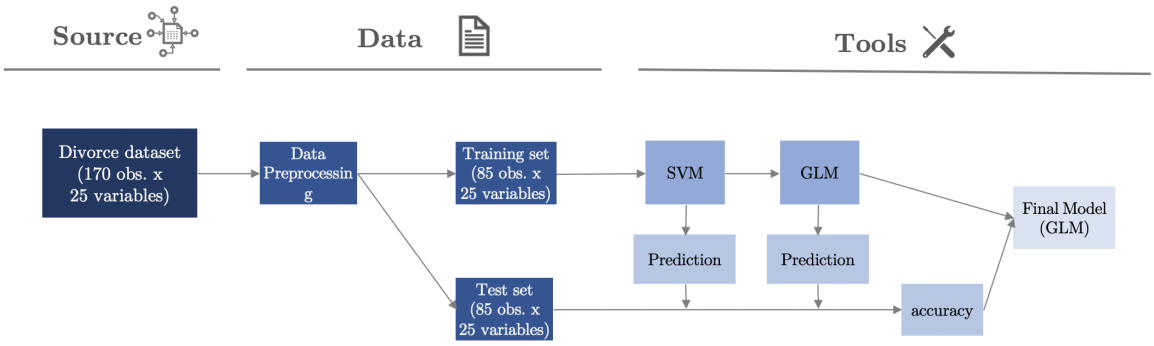

Figure 1: Project Flowchart

1. **Explanatory data analysis,**
   to understand the data it's going to be used during the project.
2. **Data preprocessing,**
   to make sure all the values are correct, and for instance there aren't Null values.
3. **Split the dataset randomly,**
   into the trainning dataset and the test dataset, with the first one we are going to build the model and with the second one we are going to test how the model works.
4. **Build the model,**
   using both methods SVM and GLM.
5. **make prediction,**
   implement both models with the test dataset to see how they perform.
6. **Choose the best model,**
   Once the predicction have been done the last step is compare the accuracy of both models and work with the one with a higher accuracy.

The project report will follow the same flow as the project and consequently will describe first the data, and the operations done on it, then the different methods used to find a usefull model, afterwards the accuracy metrics to evaluate which of the models is better and finally get the conclusions over the results.

4 Data

The dataset used contains the numerical answers of the following questions rated between 0 and 4:

1. If one of us apologizes when our discussion deteriorates, the discussion ends.
2. I know we can ignore our differences, even if things get hard sometimes.
3. When we need it, we can take our discussions with my spouse from the beginning and correct it.
4. When I discuss with my spouse, to contact him will eventually work.
5. The time I spent with my wife is special for us.
6. We don't have time at home as partners.
7. We are like two strangers who share the same environment at home rather than family.
8. I enjoy our holidays with my wife.
9. I enjoy traveling with my wife.
10. Most of our goals are common to my spouse.
11. I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
12. My spouse and I have similar values in terms of personal freedom.
13. My spouse and I have similar sense of entertainment.
14. Most of our goals for people (children, friends, etc.) are the same.
15. Our dreams with my spouse are similar and harmonious.
16. We're compatible with my spouse about what love should be.
17. We share the same views about being happy in our life with my spouse
18. My spouse and I have similar ideas about how marriage should be
19. My spouse and I have similar ideas about how roles should be in marriage
20. My spouse and I have similar values in trust.
21. I know exactly what my wife likes.
22. I know how my spouse wants to be taken care of when she/he sick.
23. I know my spouse's favorite food.
24. I can tell you what kind of stress my spouse is facing in her/his life.

Therefore, for instance if for the column "Atr23" there's a value 4, means that this person in particular, completely agree at "I know my spouse's favorite food" Apart form this 24 questions the data set also contains the column class in where it's specified if they are divorce and no, where 0 will mean "yes" and 1 "no". Summarizing, the dataset has 25 columns, 24 of them referring to the pull's questions and the last one whether they are divorce and no, besides, the dataset will contain also 170 rows each one of them representing the answers of a person.

4.1 Training and Test datasets

In order to build the model and then observe how well perfoms, the dataset will be splitted into two different datsets:

1. **Training dataset,**
   This dataset will have 85 rows randomly selected, and will be used to build the model.
2. **Test dataset,**
   This dataset will contain the remaining 85 values not picked for the training dataset, and will be used to predict with the model obtained previously and then with accuracy metrics asses whether the model is usefull or not.

4.2 Explanatory data analysis

In the very beginning when working with the data the first step is to perform an Explanatory data analysis in order to check that the data doesn't contain anomalies or have null values and perform an initial investigation about the data that it's going to be treated. Therefore to obtain all this information parametrics statistics are going to be applied, in particular it's going to be used the basicStats function available in the R software. And the results obtained are satisfactory:

For all the attributes the Nan parameter is 0.000, consequently there are not null values in the dataset, furthermore, for all the attributes the minimum and the maximum were between 0 and 4 except for the last one, the attribute class that was between 0 and 1, therefore all the values are in the expected range therefore apparently there can't be appreciated anomalies.

Moreover, to get a first idea of the data and check that all values are integers and not float numbers, for each explanatory variable there is going to be a plot between this explanatory variable and the response variable, consequently the conclusions obtained for the plots are:

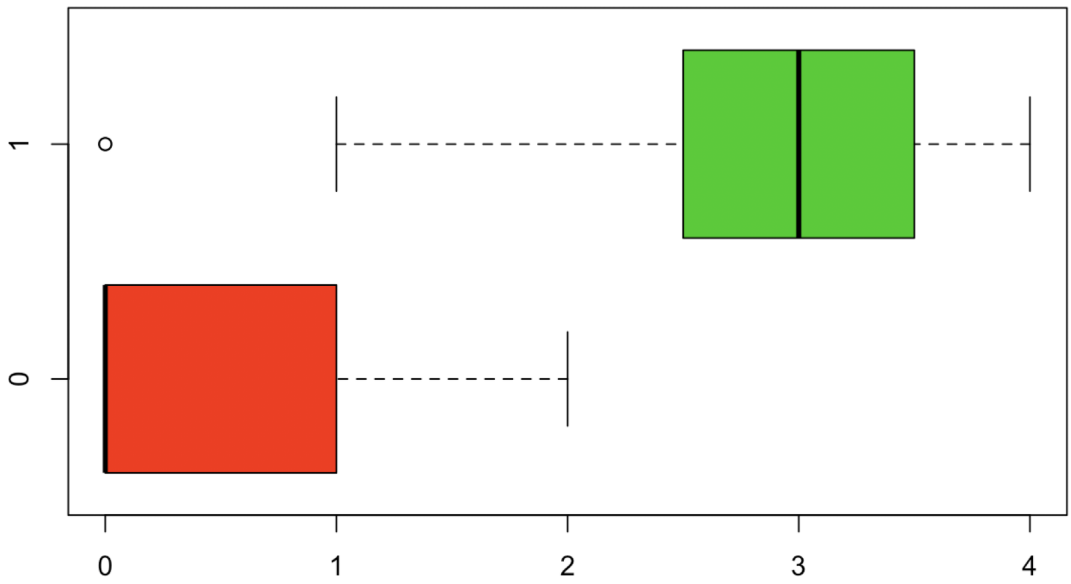

Figure 2: Plot of the response variable "Class" in function with the explanatory variable "Atr12"

All the values are integer values and in the plots for some explanatory variables it can be observed that for the values of the explanatory variable there aren't almost overlaps with the response variable, for instance, as it can be appreciated in the figure 2 there are only overlaps in the "Class" response variable when the variable "Atr12" is 1 and 2 for the other values we can already distinguish which value will have the response, therefore it's possible to deduct that the model will have a high accuracy and besides, the model will contain few parameters.

4.3 Data independence

To guarantee the independence of the data there is not a infallible method which can always say if the data it's independent or not, however one of the most trustable method is the Durbin Watson Test and applying it on the divorcedataset the p-value obtained whas lower than 0,05 so it has to refuse the null hypothesis, and therefore the data is independent.

5 Models

Once the data has been processed and validated, the next step is build the model and as mentioned before the two methods chose to predict the data are the Support Vector Machine and Generalized Linear Models:

5.1 Support Vector Machines

Since the problem can be seen as a classification problem, Support Vector Machine methods work specially well when the objective is to classify data in function of some parameters.

To obtain the optimal classification have been considered two different approaches, the dual and the primal and in both the parameters obtained to classify the data should be equal if it has been solved correctly.

Three articles has been considered to solve the problems: for the primal [1], for the dual [2] and finally one last article to see an application of this methods in real live, in this case in breast cancer [3].

Since to solve the problem with the SVM techniques can be seen as an Unconstrained Optimization Linear problem, the software used to compute the final solution is the AMPL, and the objectives function and constraints are the followings:

Primal formulation

In the primal formulation the function to minimize and the constraints are depicted in the figure below 3:

$$\begin{aligned} \min_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} \quad & \frac{1}{2}w^\top w + \nu e^\top s \\ \text{s. to} \quad & Y(Aw+\gamma e) + s \geq e \\ & s \geq 0 \end{aligned}$$

Figure 3: primal formulation

Dual formulation

In the other hand we have the dual formulation, in which the constraints and the objective function are the followings:

$$\begin{aligned} \max_{\lambda} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i y_i \lambda_j y_j K_{ij} \\ & \sum_{i=1}^{m} \lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq \nu \quad i = 1, \ldots, m \end{aligned}$$

Figure 4: dual formulation

And then, to get the W and the intercept from the variable λ already obtained it's going to be used the followings equations 5 6:

$$w = \sum_{i=1}^{m} \lambda_i y_i \phi(x_i)$$

Figure 5: obtain W from λ

$$\gamma = \frac{1}{y_i} - w^\top \phi(x_i).$$

Figure 6: obtain intercept from λ

Results

Once applied both formulations with the divorce dataset the w and *intercept* values are:

```
PRIMAL for the divorce dataset:
-------------------------------------------------------
CPLEX 12.9.0.0: optimal solution; objective 0.7843129204
13 separable QP barrier iterations
No basis.
w [*] :=
 1  0.0411462      7  1.19311e-10   13 -0.197341      19 -0.0697364
 2  0.104831       8  0.0737828     14  0.125989      20  0.736134
 3  0.391656       9  0.0737828     15  0.202978      21 -0.142525
 4 -0.0937877     10 -0.186417      16  0.174615      22  2.7756e-10
 5  0.199193      11  0.188815      17  0.366118      23  2.91166e-10
 6  0.224804      12  0.00582145    18  0.449524      24 -0.438774
;
gamma = 1.793


DUAL for the divorce dataset:
-------------------------------------------------------

CPLEX 12.9.0.0: optimal solution     objective 0.9780947034
15 QP barrier iterations
```

```
No basis.
w [*] :=
 1  0.0411462      7  1.19311e-10   13 -0.197341      19 -0.0697364
 2  0.104831       8  0.0737828     14  0.125989      20  0.736134
 3  0.391656       9  0.0737828     15  0.202978      21 -0.142525
 4 -0.0937877     10 -0.186417      16  0.174615      22  2.7756e-10
 5  0.199193      11  0.188815      17  0.366118      23  2.91166e-10
 6  0.224804      12  0.00582145    18  0.449524      24 -0.438774
;
gamma = 1.793
```

Table 1: AMPL Solution for both formulations

As it can be appreciated for both formulations the result obtained are exactly the same, which it's a clear indicator that the results are correctly obtained, moreover, observing the values of the W we can say that the explanatory variables that are more significant are the 20, 18 and -24.

5.2 Generalized Linear Models

Apart form the SVM the other technique used wull be the Linear Model, however, due to the response variable must be binary, it can not be used the ordinary linear model, in fact, it should be used a Generalized linear model with a binomial family. The articles used to undestand this techique are [4] for the GLM and in particular, for the binomial GLM [5].

Multicolinearity

When modelling with Generalized Linear Methods, most of the times it's essential avoid multicollinearity among the variables, since if the collinearity between variables exists can difficult the interpretation of the model, the variance of the parameters increments and the matrix X may not be singular which makes impossible to compute the parameter estimations. To observe the multicollinearity normally a scatter-plot between two variables is performed, or a Principal Components Analysis or calculate the Variance Inflator Factor, however, few experts in statistics argue that when dealing with explanatory variables with numerical values and besides the dataset isn't very large the collinearity among variables can be dismiss, there are few articles talking about it, one of them is [6]. Moreover, it's important remark that there could be good models with multicollinearity between the variables.

GLM Model

As it has been said previously, because of the response variable it has to be binary it's necessary use the Generalized Linear Models with a binomial distribution. Moreover, the procedure to determine which parameters are significant and which not is the AIC [**?**], taking as $K = log(num_rows)$ the algorthm iteratively will add the more significant parameters until the AIC score increments when adding one parameter, besides the variables that will take into account the algorithm are all the parameters plus all the interactions between them plus all parameters squared.

Results

The final model obtained with the AIC procedure [7], predicts the variable "Class" with the variables "Atr18" and "Atr20",in particular, the function will remain as:

$$Class = -61.30756 + 61.30756 * Atr20 + 40.58777 * Atr18.$$

In order to understand deeper the model obtained the following plots have been done 7 8:

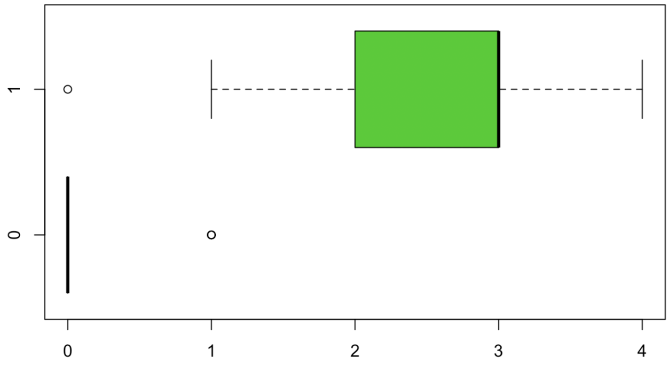

Figure 7: plot of variable "Class" versus variable "Atr18"

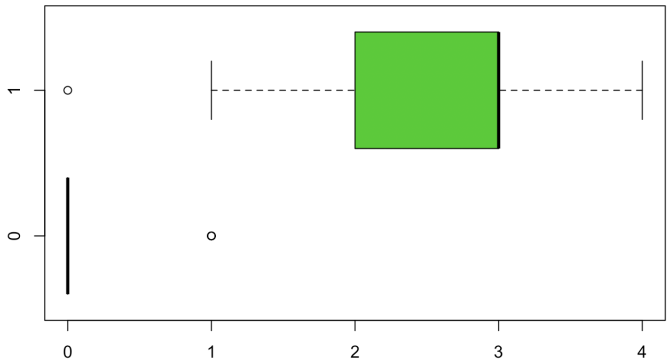

Figure 8: plot of variable "Class" versus variable "Atr20"

Even thought they apparently have the same quartiles and means, the data is different for "Atr18" and "Atr20".
Moreover, looking at the plots it's possible to deduce why the model only used two variables to determine the response output, because as it can be appreacited for both variables there's only a little bite of uncertainity when the variables have as a value 0 or 1, and therefore are really good estimators for the "Class" response variable

5.3 Models Accuracy

First of all, to calculate the SVM model accuracy, first it is going to be necessary to predict the response variable values with the test dataset, to do so, since as it can be appreciated in the picture below 9 the classification of the data is done by means of the variables W and *intercept* it's going to be used an Excel spreadsheet to compute them:

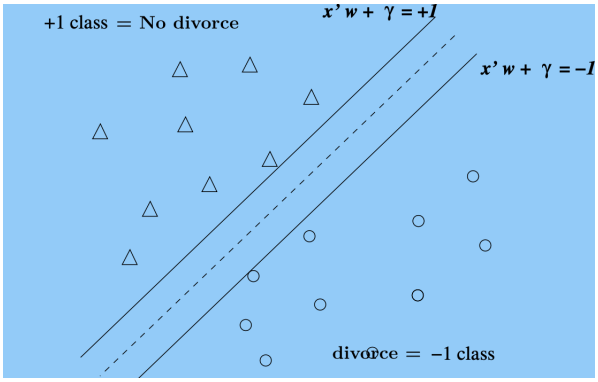

Figure 9: SVM classification

finally once the prediction has been done, the accuracy it's calculated with the formula 5.3:

$$Accuracy = \frac{1}{n_{rows}} \sum_{i:n_{rows}} y_i - \hat{y}_i$$

And the accuracy result for SVM is: 94,1 %

Furthermore, to obtain the accuracy for the GLM the prediction has been done with the GLM models on the test dataset in R, and the metric to calculate the accuracy was the 5.3 as well, and the result obtained was: 98,8 %, therefore even though both models obtained a high accuracy the GLM made the prediction better.

6 Conclusions

The purpose of this report was to understand which factors contribute in the divorce decision, and obtain a model capable to predict whether they will divorce or not given some answers to particulars questions 4. Therefore, to understand why the divorce occurs, and consequently which factors contribute more, in this project it has been implemented two different techniques and the models obtained from them differ in almost all the parameters, for instance, in the SVM 19 variables have been considered significant (Weight associated with the variable greater than 0,001) and in the GLM only 2 parameters. So since the both methods don't converge in the same solution, the method refused is SVM because GLM has obtained a more simplex model and which is more relevant, has lager accuracy, therefore the conclusions extracted from the SVM soltuion will be dissmiss.
Consequently, the final model to predict if there is going to be divorce or not is:

$$Class = -61.30756 + 61.30756 * Atr20 + 40.58777 * Atr18$$

.

Then, observing the function the following conclusion can be extracted:

1. **Atr20,**
   The explanatory variable that has a larger impact on the response variable is the "Atr20", then can be deduced, that for the question 20: "My spouse and I have similar values in trust", the

probability of doesn't divorce will increment if the answering person agrees with the sentence and consequently punctuate with a high numerical value.

2. **Atr18,**
   The second most relevant and the last significant explanatory variable is the "Atr18", that corresponds to the question "My spouse and I have similar ideas about how marriage should be", and as the "Atr20" the larger the numercal value of this question the higher will be the probability of doesn't divorce.

3. **Model simplicity,**
   Even though the common sense may tend to think that the bigger number of explanatory variables the higher will be the accuracy when predicted, this is not like this, since as it can be seen with the output of this project, the GLM model with 2 significant variables obtained a higher efficiency than the SVM model with 19 significant variables, furthermore, since the better model only used two variables, the pull that have to answer the people can be reduced from 24 questions to a 2 questions, in particular questions 18 ans 20, and thus, save a large amount of resources and time.

To conclude, since the model obtained had a high accuracy, in particular around 98,8%, and besides, the model had only two variables and therefore simplified significantly the problem the results obtained are considered satisfactory.

References

[1] Olivier Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, may 2007.

[2] O L Mangasarian and David R Musicant. Active Support Vector Machine Classification. Technical report.

[3] Leonardo de Oliveira Martins, Geraldo Braz Junior, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, and Marcelo Gattass. Detection of Masses in Digital Mammograms using K-means and Support Vector Machine. Technical Report 2, 2009.

[4] J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370, 1972.

[5] A. R. GILMOUR, R. D. ANDERSON, and A. L. RAE. The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72(3):593–599, 1985.

[6] Role of Categorical Variables in Multicollinearity in the Linear Regression Model.

[7] Hirotugu Akaike. A Bayesian Analysis of the Minimum AIC Procedure. pages 275–280. 1998.