APA- Practical Work 2017-2018

Albert Ribes Kerstin Winter

November 8, 2017

Contents

Todo list

1 Introduction

1.1 Desciption of the work and its goals

The goal of this project is to build a classification model to predict whether a lung cancer patient will die within one year after surgery or not. To do so we will study a dataset with real lung cancer patients.

As this is very sensitive information, our priority will be to minimize the amount of falses negatives, i. e, avoid predicting a patient will not die within one year when it certainly does.

The data is taken from `https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data#` [1]

1.2 Desciption of available data

The data we are working with is about patients who underwent major lung resections for primary lung cancer in the years from 2007 to 2011. For each patient we are given information about his diagnosis and effects produced by the cancer.

The dataset is very limited in the number of instances available: it only has 470. In addition, the distribution of the predicted class isn't quite balanced, since only 70 of the patients died in one year period. This may become a problem in some of the prediction models due to the fact that the results will be biased towards the biggest class. However, we can suppose that the data has been collected uniformly and that this proportion is similar to the real one.

For each patient we have 16 different atributes. 3 of them are numerical, and the rest are categorical. From those, 10 are binary. The response atribute is also binary.

2 Related Previous Work

3 Data exploration process

3.1 Pre-processing

3.1.1 Treatment of missing values

Our dataset do not have missing values, so there is no need to treat them.

3.2 Treatment of anomalous values

Quizá hay que quitar algunas personas por ser demasiado jóvenes comparadas con el resto

3.2.1 Treatment of incoherent values

Mirar si tenemos valores incoherentes

3.2.2 Coding of non-continuous or non-ordered variables

3.2.3 Posible elimination of irrelevant variables

3.2.4 Creation of new useful variables (Feature extraction)

3.2.5 Normalization of the variables

3.2.6 Transformation of the variables

3.3 Feature extraction/selection

3.4 Clustering

3.5 Visualization

4 Resampling protocol

5 Results obtained using linear/quadratic methods

5.1 LDA

Usar LDA para tener 2 centroides, que son los de cada una de las clases. Cuando evaluamos un dato nuevo, le aplicamos la transformación y miramos si queda más cerca de uno o de otro. Así podemos ver la probabilidad de que pertenezca a cada una de las clases.

Mirar el vecino más cercano para precedir

Si suponemos que las variables son independientes: -Haces naive Bayes para ver la probabilidad de que pertenezca a cada una de las clases (habría que estudiar si las variables son independientes) - Logistic regression

LDA (Fisher) para reducir la dimensión y hacer los otros

6 Results obtained using non-linear methods

7 Desciption and justification of the final model chosen

7.1 Estimation of the generalization error

8 Self-assessment of successes, failures and doubts

9 Scientific and personal conclusions

10 Possible extensions and known limitations

References

[1] Maciej Zikeba et al. "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients". In: *Applied Soft Computing* (2013). DOI: [WebLink].