APA- Practical Work 2017-2018

Albert Ribes Kerstin Winter

November 16, 2017

Contents

1 Introduction **2**
1.1 Desciption of the work and its goals . . . 2
1.2 Desciption of available data . . . 2

2 Related Previous Work **3**

3 Data exploration process **3**
3.1 Pre-processing . . . 3
3.1.1 Treatment of missing values . . . 3
3.1.2 Treatment of anomalous values . . . 3
3.1.3 Treatment of incoherent values . . . 3
3.1.4 Coding of non-continuous or non-ordered variables . . . 3
3.1.5 Possible elimination of irrelevant variables . . . 3
3.1.6 Creation of new useful variables (Feature extraction) . . . 4
3.1.7 Normalization of the variables . . . 4
3.1.8 Transformation of the variables . . . 4
3.2 Clustering . . . 4
3.3 Visualization . . . 4

4 Resampling protocol **4**

5 Results obtained using linear/quadratic methods **4**
5.1 Naive Bayes . . . 4
5.2 KNN . . . 5
5.3 LDA . . . 5

6 Results obtained using non-linear methods **5**

7 Desciption and justification of the final model chosen **5**
7.1 Estimation of the generalization error . . . 5

8 Self-assessment of successes, failures and doubts **5**

9 Scientific and personal conclusions **5**

10 Possible extensions and known limitations **5**

Todo list

Redactar que FEV1 está mal, referenciar algún artículo que hable sobre el tema y para justificar que está mal, decidir qué haremos con esos pacientes (eliminarlos, o inferir sus valores de FEV1 en función de sus vecinos) y si los inferimos poner el proceso como lo hemos hecho 3
Redactar bien esta parte, indicando que puesto que tenemos pocos datos, podemos permitirnos hacer varias ejecuciones, y que probaremos cada combinación de eliminar y no eliminar cada una de estas variables y veremos cual da mejores resultados con la cross-validation . . 3
Poner los nuevos valores 4
Buscar la biblioteca que lo calcula y aplicarlo a nuestros datos 4

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 false 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#` [**zieba2013boosted**]

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.

3.1.2 Treatment of anomalous values

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

3.1.3 Treatment of incoherent values

The variable FEV1 which is the Forced Expiratory Volume in 1 second, shows a few anomalously high values. Depending on factors lika age and sex of a patient the average value of the FEV1 is around 3-6 litres, whereas the dataset shows values up to 86. As most of the values are within 0 and 10 we assume that the dataset contains the FEV1 in litres. To decide which values are to be determined as outliers we calculate the FEV1/FVC ratio which gives the percentage of the lung volume exhaled in the first second over the whole exhaled volume. All patients having a unrealistic ratio higher than 100%, which are 22 patients, are determined to be outliers and eliminated. We chose not to apply any stricter constraints because the dataset does not include the sex of the patients which influences the normal values of the FEV1 much.

Source of knowledge about FEV1 and FVC: https://www.nuvoair.com/blogs/blog/do-you-know-how-to-interpret-the-results-of-your-spirometry-test

Redactar que FEV1 está mal, referenciar algún artículo que hable sobre el tema y para justificar que está mal, decidir qué haremos con esos pacientes (eliminarlos, o inferir sus valores de FEV1 en función de sus vecinos) y si los inferimos poner el proceso como lo hemos hecho

El FEV1 tiene valores incoherentes. La mayoría están sobre 3, pero algunos están sobre 60

3.1.4 Coding of non-continuous or non-ordered variables

3.1.5 Possible elimination of irrelevant variables

Some of the variables of our dataset are not well represented. In particular:

DGN	There is just one patient with DGN = 1 and just 8 have DGN = 8
PAD	Just 8 patients have PAD = True
ASHTMA	Just 2 patients have ASHTMA = True

Redactar bien esta parte, indicando que puesto que tenemos pocos datos, podemos permitirnos hacer varias ejecuciones, y que probaremos cada combinación de eliminar y no eliminar cada una de estas variables y veremos cual da mejores resultados con la cross-validation

3.1.6 Creation of new useful variables (Feature extraction)

Entender cómo funciona MCA, y ver si podemos sacar una variable nueva Quizá es interesante añadir la variable FEV/FEV1

3.1.7 Normalization of the variables

We need to normalize only our numeric variables, which are the AGE, FEV and FV1.

Solo se pueden normalizar FVC, FEV1 y AGE. Miraremos cuales dan mejores resultados

3.1.8 Transformation of the variables

Acording to the paper we found the accetable range for skewness in a numeric variable is $(-2, +2)$. The skewness of our original variables AGE, FVC and FEV are:

AGE	-0.1899413
FVC	0.5417132
FEV1	5.597584

But we have to take into account that we've eliminated 22 patients, so the new values are:

Poner los nuevos valores

As the three variables are in the specified range, there is no need of transforming them.

Referenciar (y leer un poco...) el paper

3.2 Clustering

hacer varios k-means con distintos valores de k (2,3,4,5,6) para ver si descubrimos algún cluster que nos permita crear una variable nueva

3.3 Visualization

Hacer MCA

4 Resampling protocol

5 Results obtained using linear/quadratic methods

5.1 Naive Bayes

Buscar la biblioteca que lo calcula y aplicarlo a nuestros datos

5.2 KNN

5.3 LDA

Suponiendo que las varianzas de cada una de las clases son la misma, se usa este algoritmo, (que simplifica QLA) para ver la probabilidad de pertenencia a una clase

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

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

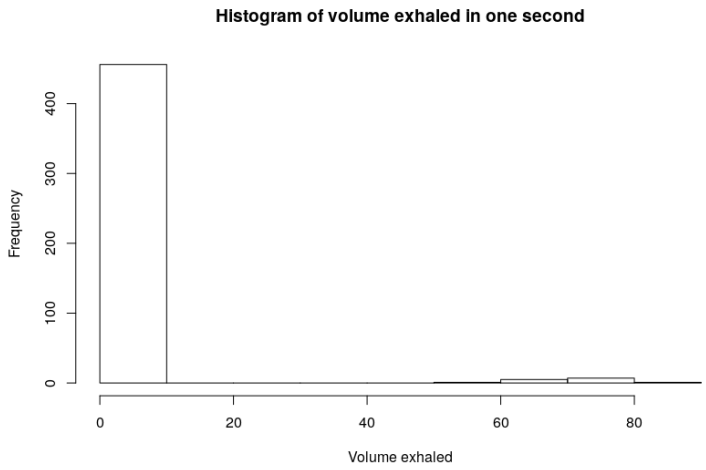

Figure 1: Show the ammount of people having each value