FIFTH STEP

Miguel Lobo Martin, Borja Ruiz Amantegui, Francesca Sallicati, Antonio Martínez Payá

February 4th 2018

Contents

1 Introduction

The objective of the fifth step is to perform supervised classification in our data set in order to correctly classify players on the position that better adjusts to their characteristics. For this purpose, we are going to apply three different techniques which are methods based on Bayes Theorem , KNN and Logistic Regression. In order to perform the bayes rule we will use different methods to approximate the density functions. The methods are Linear Discriminant Analysis, Quadratic Discrimant Analysis and Naive Bayes: each one of them relies on different assumptions.

In order to prevent any unbalanced problems with our data set we decided to create a new data frame containing 1000 observations of each class selected randomly. Nevertheless, in KNN and Logistic Regression we use the complete dataset since no issues were found.

For this project we decided to use train and test validation since the number of observations that our data set contains are enough to offset the possible drawbacks that the use of this method could have. The high number of observations would make crossvalidation very computational demanding and we believe train and test would provide enough accuracy.

2 Supervised Algorithms

2.1 Methods based on Bayes Theorem

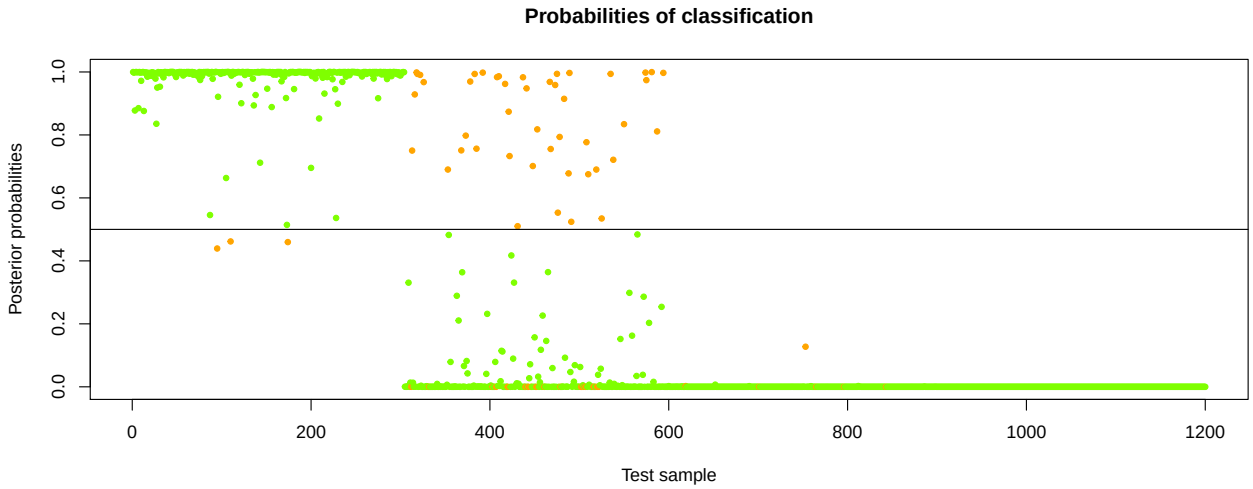
2.1.1 Linear Discriminant Analysis

	A	DF	GK	M
A	291	0	0	3
DF	0	276	0	12
GK	0	0	311	0
M	49	28	0	230

Here we can observe the contingency table with the classifications made by the algorithm. We can infer high accuracy from this technique since most of the observations have been correctly classified. However, there are some problems in the midfielder category.

The error rate that **Linear Discriminant Analysis** yields is equal to 0.07666667, number that corroborate our first impression.

In this graph we can see the correct classifications (plotted in green) and the wrong ones, which are colored in orange. Most of the orange dots correspond to the midfielder category, where most of the wrong classifications lay.

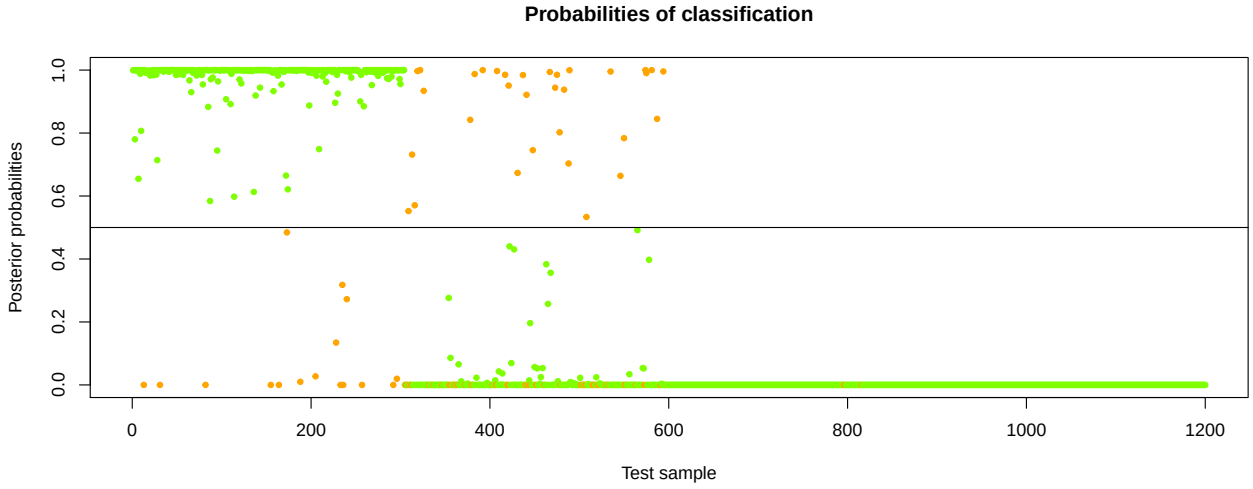


2.1.2 Quadratic Discriminant Analysis

	A	DF	GK	M
A	282	0	0	12
DF	0	280	0	8
GK	0	0	311	0
M	36	44	0	227

This method has the same difficulties as the **Linear Discriminant Analysis** when classifying the midfielders. We want to check the error rate in order to verify which one provides better results.

The error rate that **Quadratic Discriminant Analysis** yields is equal to 0.08333333, which is indicates a slighlty worst performance.

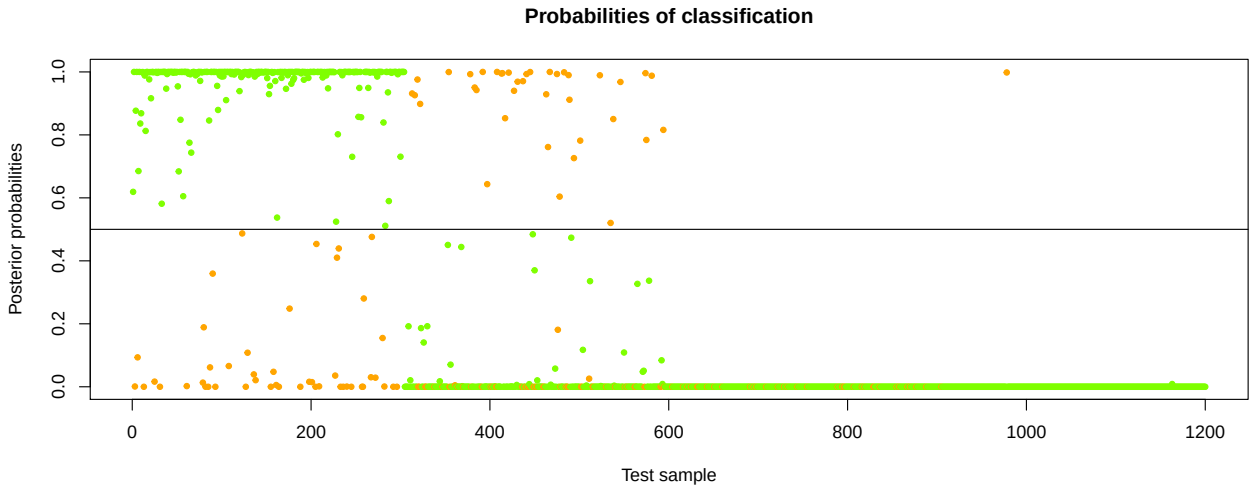


We have a similar situation as in **LDA**, but slightly worse since we can observe some extra wrong classifications that were not present in the first case.

Naive Bayes

	A	DF	GK	M
A	250	1	0	43
DF	0	246	0	42
GK	0	1	310	0
M	53	43	0	211

The error rate that **Naive Bayes** yields is equal to 0.1525, which is considerably higher than the previous two methods.



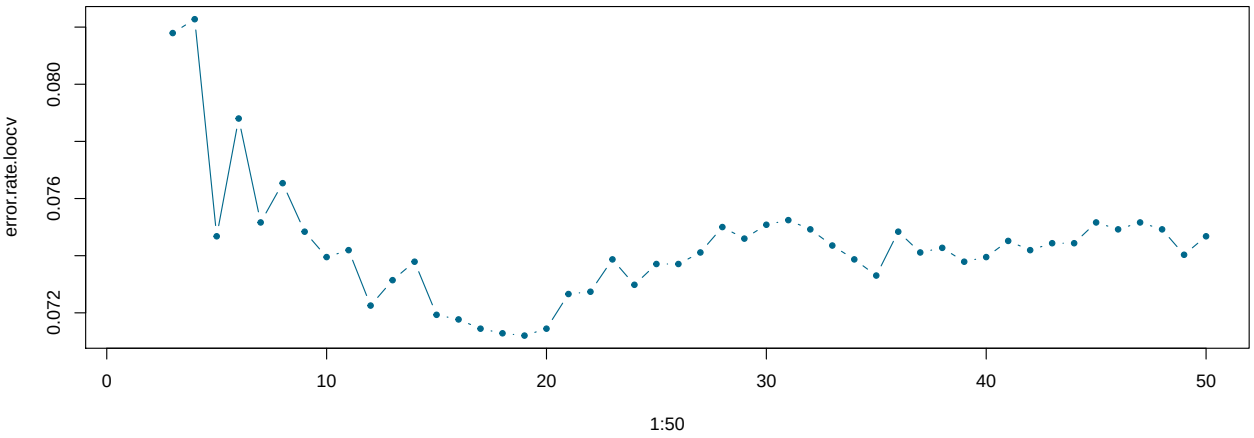
We can graphically appreciate the presence of more orange dots in this graph, which further verifies the error rate result.

2.2 KNN

Knn is one of the most popular and simpler methods for supervised classification in very general conditions. We tried our methods with and without balancing the dataset and since results were practically equal all results shown are with the full dataset.

Once our predictors have been standardized we then compute the error for K in a range between 3 and 50 and plot them against their errors.

We then compute the error for K in a range between 3 and 50.



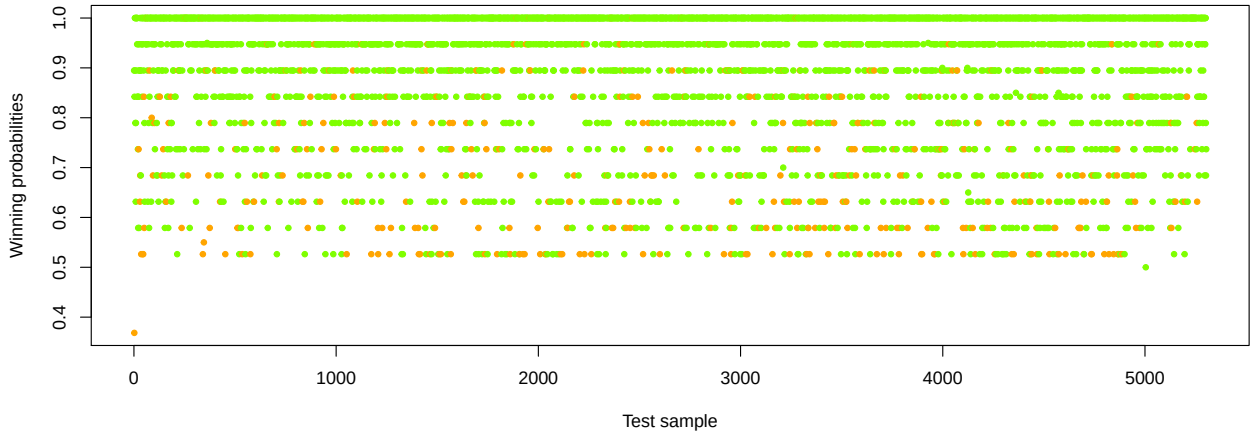
Clearly the least error is shown in with K=19 so it is the one we will use to classify our observations.

	A	DF	GK	M
A	205	0	0	158
DF	0	779	0	131
GK	0	0	600	1
M	45	38	0	3346

We can observe that though the KNN method classifies well, it fails some classifications for midfielders, this is a constant problem in our dataset and it is due to the diffuse features that separate midfielders from either attackers and defenders.

When computing the error we obtain: 0.07033754.

We plot the test observations over the winning probabilities and appreciate coherent results since more wrong classifications are made when the winning probabilities are low.

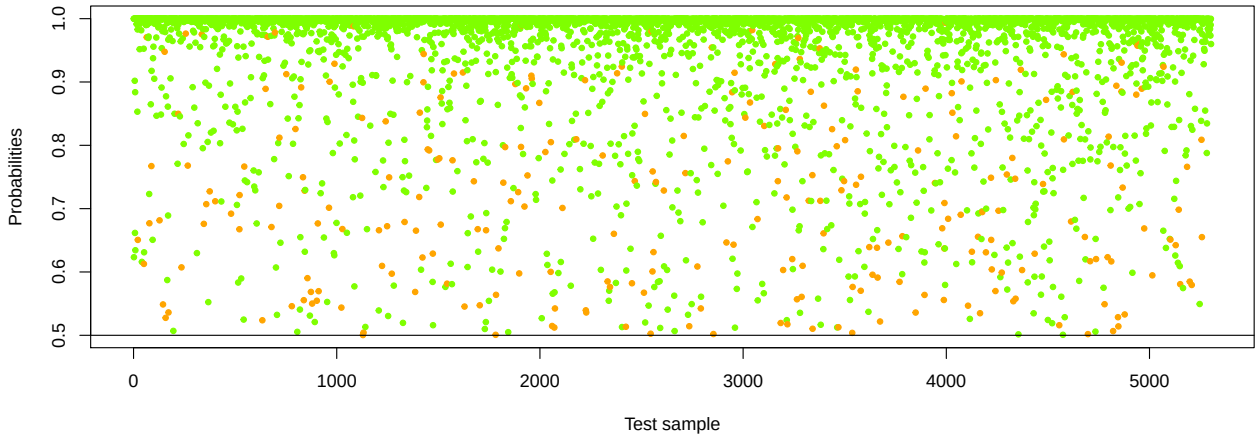


2.3 Logistic Regression

In the last step we applied a multiple logistic regression.

	A	DF	GK	M
A	262	0	0	101
DF	0	845	0	65
GK	0	0	601	0
M	52	45	0	3332

From the confusion matrix we can see that the method is able to perfectly classify all the goalkeepers, whereas it misclassifies almost half of the attackers and some defenders into midfielders. However this model is the best among those trained since it has the lowest Test Error Rate.



The above plot shows that almost all of the observations with high probability are well classified (green), while there are some misclassifications when the probabilities get closer to 0.5.

Now let's try to reduce our logistic model through stepwise selection without loosing the accuracy of the method.

	Backward	Forward	Both
TER	0.04978314	0.04959457	0.04978314

We notice that the forward method isn't actually discarding any variables, while the directions both and backwards select 9 over the 12 variables with little loss in the accuracy of the model; in particul Wage, Value and Potential are deleted.

Therefore due to complexity reasons we would prefer to use the model obtained with *both stewise selection* since the accuracy is not decreasing.

3 Conclusion

The category goalkeeper is always correctly classified since players that play in this position are very easily detected by the algorithms due to their special characteristics. Regarding the atacker and defender category, we have different error rates depending on which techinques we use. We found these results linked to the distinct features that midfilders have among themselves. This makes difficult for the algorithm to categorize some players, which based on their features, could be either midfielders or defenders, as well as, midfielders or attackers. For our dataset the classification methods tend to fail in different manners depending on the size of midfielders observations. The higher the number of midfielders, methods tend to classify more observations as midfielders, and more failures are found in the attacker and defender categories. When we used our balanced sample, the failures are found in the midfielder category, classifying some of them as attackers and defenders. However none of the methods are confusing attackers with defenders or viceversa, so we must highlight midfielders category as a classification issue.

When comparing the different methods that have been used, based on error rate and simplicity, the best model is logistic regression.