Logistic Regression

Group 2

02 12 2019

1. Introduction

The aim of this project is to predict whether a College Football Player will be drafted into the NFL. This ReadMe lines out, how the Logistic Regression model was used to classify the players into the to categories "Drafted" and "Not Drafted" based on game statistics that quantify the player's performances. In a first step, the theoretical idea behind the Logistic Regression classifier will be elaborated on and then the application on our case will be described. In the end, the results will be displayed.

2. Logistic Regression - Theory

In the logistic model, it is assumed that the relation between the explanatory variables (in our case the statistics about the player's performance) and the conditional probability $P(Y = 1|X)$ or $P(Y = 0|X) = 1-P(Y = 1|X)$ is given by a logistic function. In other words, the Logistic Regression classifier predicts the probability of an observation to belong to class 1 or 0 (in the case of a binary Logistic Regression as in our case). Logistic functions have a sigmoid shape and are given by

$$f(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

This makes sense, since the numerator of a sigmoid function (in this case 1) is equivalent to the maximum of the function and the probability of an observation belonging either to class 1 or to class 0 can never be more than 1. We then set

$$P(Y = 1|X) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}$$

where X' denotes the features of a specific observation and β denotes the coefficients. In a last step, the Logistic Regression algorithm will assign the observations to a discrete set of classes, 0 or 1. A threshold is defined (or by default set at 0.5). If the probability predicted by the model/ the logistic function is smaller than the threshold, it will assign the class 0 and if the predicted probability is larger than 0.5, it will assign the class 1.

3. Logistic Regression - Application to our case

3.1 Training

We have built individual models for the different player positions, as well as for different data samples (he specifics as to why we did this are further elaborated on in the Read Mes about Data Sampling and Data Handling) We have used the respective data for the years 2007 to 2013 for training and applied a 10-fold cross validation. We use the `train()` function from the caret package to do so. The following example displays the code run for building the model for the unsampled data set and all players together.

```
model_logit_tog = train(Drafted ~ .,
                        data = Data_tog_train,
                        trControl = trainControl(method = "cv", number = 10),
                        method = "glm",
                        family=binomial())
```

In the end we have run a Logistic Regression model for all combinations of player positions and data samples. This amounts to 20 (4 positions including all together * 5 data samples) models.

3.2 Testing

After having trained the models, we test them on the 2014 data. For reasons of comparability, this is always done for the unsampled data.