Data analysis and data Mining
Report 2

Mauro Angelini
Alessio Gilardi

November 28, 2018

Contents

First exercise

We define a multiclass classification problem.
Being a multiclass problem it is necessary to use a one-vs-all or all-vs-all method. In this case we use a linear classifier with an all-vs-all method, in which each class is compared to all the others with each new iteration of the algorithm.
We use a λ as regularization parameter. To get the best solution we divide the dataset into two sets: a learning set, corresponding to the 70% of the initial data and a validation set, in order to calculate the error committed by our model, and run a loop on different λ values in order to identify the best value.
To mediate the result and minimize the influence of variance we repeat this procedure a number k of times, with k equal to 30 for example.

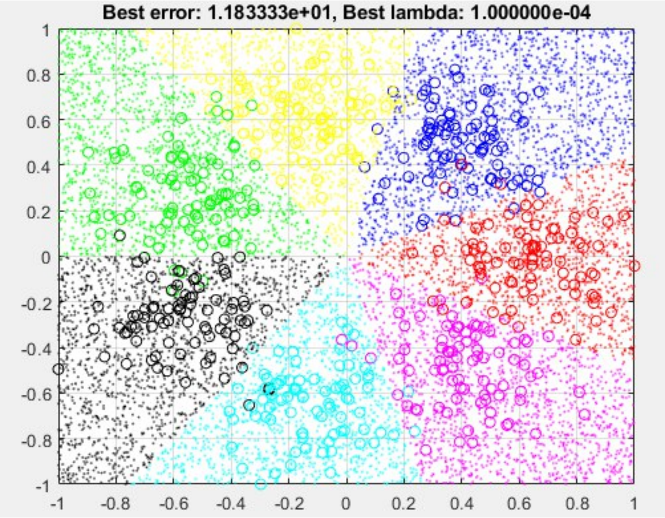

Figure 1: Regression Function

Second exercise

We define another classification problem and we deal with it using SVM. The regularization term that we can optimize this time is the C parameter, taking into account that if we use low C values then the solution will be smooth, whereas if we use big C values we are only interested in minimizing the error and will have a more complicated solution . To optimize the term on C we loop on it for various values to find the most sparse $alpha$ vector, ie with the greatest number of zeros, and the corresponding w model. At this point with the optimal value of C and the respective model we plot the solution.

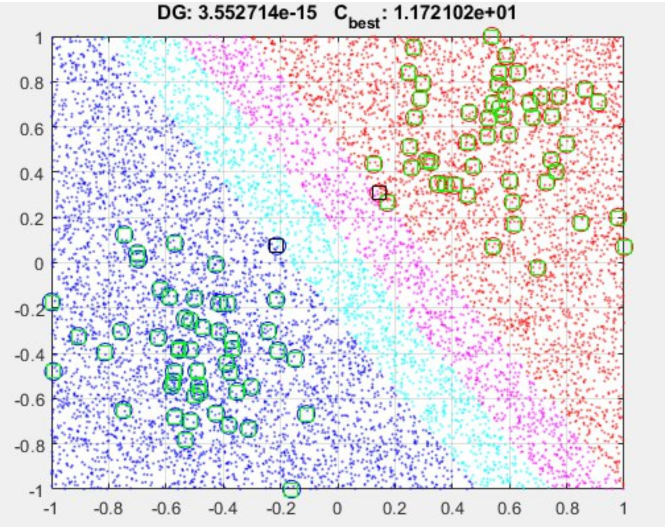

Figure 2: Regression Function

Third exercise

In this exercise we implement a multiclass classification problem with a Decision Tree. We start from the binary case and extend it to the multiclass problem.
The parameter to be optimized in this problem can be the depth of the tree.
Also in this case we divide the dataset into two sets, the learning set and the validation set and using the validation set we calculate the error as the number of points classified incorrectly. Here again we repeat the product k times, with k equal to 30.

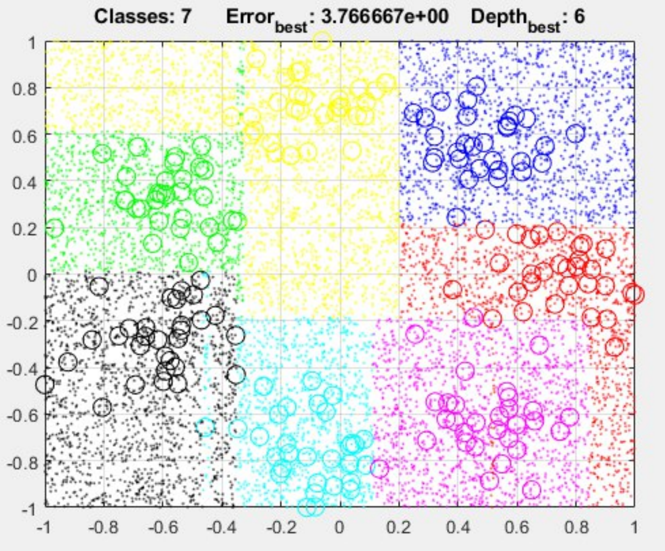

Figure 3: Regression Function

Finally, we took a dataset with 3 features, in order to be plottable in 3D, and 3 classes. As before we divided dataset in training set and validation set, we used 70% of data in training and the remaining for validation. Then we ran a loop on k and on depth to optimize the depth parameter.

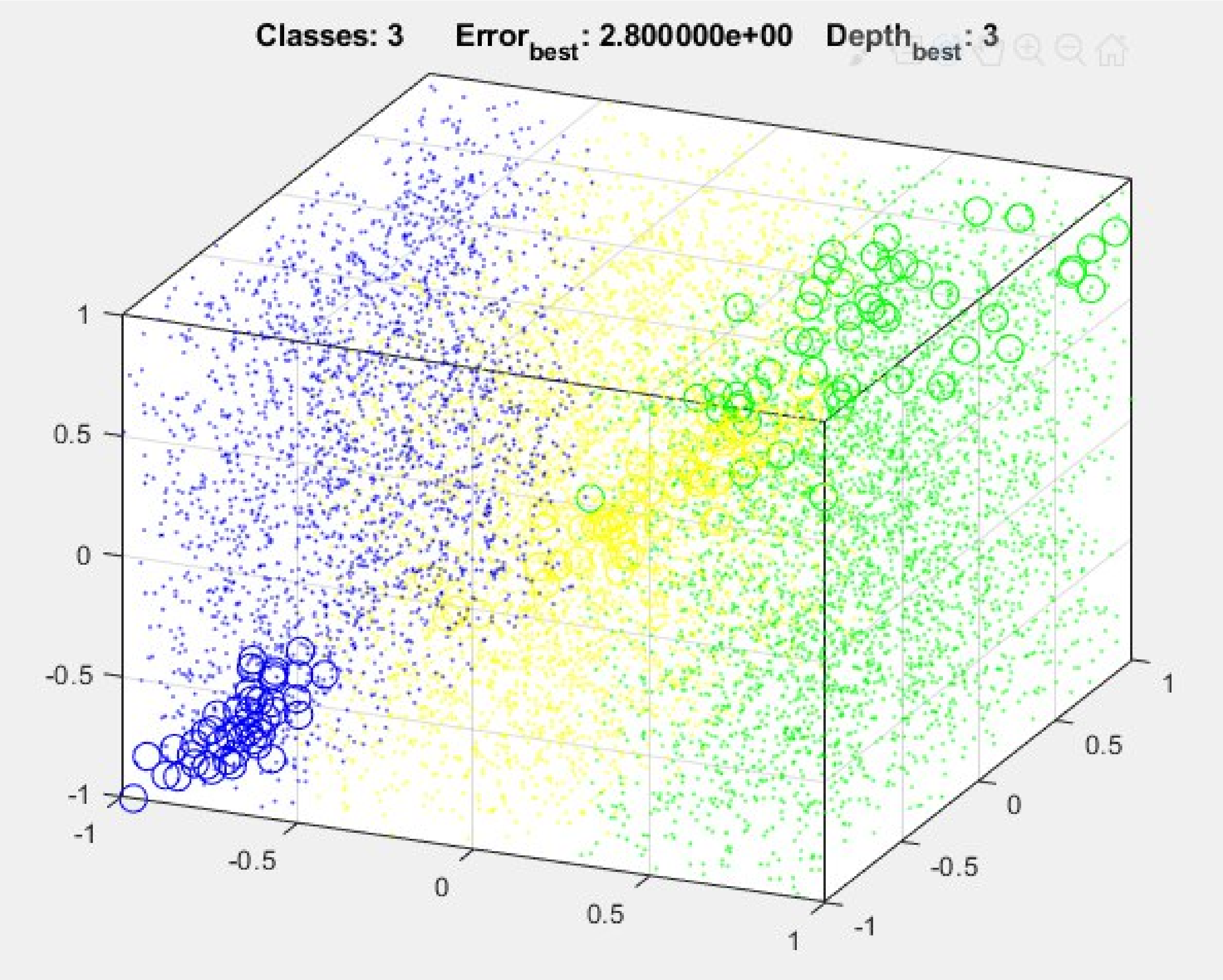

Figure 4: Regression Function

List of Figures