Advanced computational methods - Problem set 1

Hrvoje Stojic

January 8, 2016

1. Think of the data that linear models would have troubles with fitting well, it can be simple 2-dimensional data set with 2 categories or more complex. After giving it some thought, scour the machine learning literature for benchmark classification datasets. Then write a file with a function that will generate such dataset. Name the file `genData.R`. The arguments to the function should be as flexible as possible (number of observations for example), think of somebody else using your function. The output of the function should be the following:
   - Main output is an R object - dataframe with the data.
   - But it should also save the dataset in `.csv` format under the name `dataset.csv` in the same folder. Note that this means you should not set the directory to any specific path. Saving the dataset should be governed by an argument to the function that is set to `TRUE` by default.
   - It should also save an illustrative plot of the data in `.pdf` format under the name `dataPlot.pdf`. Of course, if it has more than 2 dimensions, plot only one or two dimensions (three is also a possibility). Saving the plot should be governed by an argument to the function that is set to `TRUE` by default.
   - If you did not dream up that dataset (you have seen it in some article instead), in comments of the function write the reference to the article and URL to it. I will collect all the interesting functions, put it in a single `.R` file and provide it to all of you, so that you will be able to use them for testing your classification algorithms.
2. Use the simple dataset function I have created, `loanData`, and add a third category - “Undecided”. Train a discriminant function on the dataset. Be careful with the coding scheme (check “More general implementation” section in my handout and Section 4.1.2 in Bishop)! Name the file `loanData3C.R`. The main outputs of the file should be the following:
   - Main output is an R object - dataframe with the dataset, where you added columns for predictions for each class and final decision to which class does the observation belongs to.

- Running the file should also save the final dataset in `.csv` format under the name `predictions.csv` in the same folder.

- Similarly, running the file should also save a plot of the data and decision boundaries in `.pdf` format under the name `discFunction3C.pdf`. Decision boundaries will, of course, look differently than the two class example in the handout.

Important details:

- Deadline is January 15, at 12h.

- Recall that you have to keep your code under version control in a single repository at Github, I will check your submission there. Place all the files for this problem set in a folder under the name "PS1".

- Please, pay attention to the names of the files I instruct you to use. Names facilitate examining the code greatly, there is 30 of you and it takes me far more time if everybody produces different set of files with different names.

- Create a simple text file with the name `Readme.md` in the problem set folder, if you want to leave me any instructions or messages regarding the problem set.

- I will give extra points for extra nice solutions!