adaboost

Submitters:

Yishay Seroussi 305027948, Samuel Bismuth 342533064.

Python version:

3.9

Configuration

This repository includes an implementation of the adaboost alogithm in python.

We use the docker environement. Make sure docker is installed in you machine. That is the only dependency of the project.

According to your distribution, run:

```
sudo <yum/apt-get> install -y docker
```

Then, to run the script run:

```
bash start.sh
```

To enter in the container terminal (only for developement purpose):

```
bash bash.sh
```

If you don't want to use docker, you are able to run the code in any machine by folowing the next steps:

Install python3.9.

Install numpy by running:

```
pip3 install numpy
```

Run the main python file:

```
python3 main.py
```

Code structure:

The code is composed of three folders.

- The data folder containing the two txt files of data received for the assignment.
- The packages folder containing the requirement txt file with the pip lib we used (numpy).
- The src folder containing the code source.
    - main.py -> The main file of the code. This is our entrypoint. This is also the file were the prints are done.
    - data.py -> The file handle the txt data to convert it into objects.
    - geometry.py -> Here the classes Line and Point are defined.
    - model.py -> Here the Feature and Label classes are defined.
    - rule.py -> Here the Rule class is defined.
    - adaboost.py -> Here you can find the main algorithm with the computation of the final error and error.

Example of outputs:

NUMBER OF RULES 1

Hc Body Temperature error

hc body temperature train error: 32.31
hc body temperature test error: 32.69

##

Iris error

iris train error: 25.11

iris test error: 24.89

##

NUMBER OF RULES 2

########## Hc Body Temperature error ##################

hc body temperature train error: 23.1
hc body temperature test error: 30.79

##

########## Iris error ##################

iris train error: 3.22
iris test error: 5.86

##

NUMBER OF RULES 3

########## Hc Body Temperature error ##################

hc body temperature train error: 23.1
hc body temperature test error: 30.79

##

########## Iris error ##################

iris train error: 3.22
iris test error: 5.86

##

NUMBER OF RULES 4

########## Hc Body Temperature error ##################

hc body temperature train error: 19.32
hc body temperature test error: 30.02

##

########## Iris error ##################

iris train error: 2.63
iris test error: 5.51

##

NUMBER OF RULES 5

########## Hc Body Temperature error ##################

hc body temperature train error: 19.3
hc body temperature test error: 30.12

##

########## Iris error ##################

iris train error: 2.54
iris test error: 5.45

##

NUMBER OF RULES 6

########## Hc Body Temperature error ##################

hc body temperature train error: 16.58
hc body temperature test error: 29.5

##

########## Iris error ##################

iris train error: 2.44
iris test error: 5.52

##

NUMBER OF RULES 7

########## Hc Body Temperature error ##################

hc body temperature train error: 16.84
hc body temperature test error: 29.76

##

########## Iris error ##################

iris train error: 2.35
iris test error: 5.37

##

NUMBER OF RULES 8

########## Hc Body Temperature error ##################

hc body temperature train error: 13.41
hc body temperature test error: 29.08

##

########## Iris error ##################

iris train error: 2.25
iris test error: 5.41

##

Do you see overfitting?

Let first focus on the Hc Body Temperature data set.

Obviously, the result are less impresive than the Iris's data set results. Both error are relatively high.

The difference between the error of the train and the error of the test is growing up with the number of rules, anotherly said, with the number of VC dimension.

That is, we can say that there is overfitting.

Regarding the Iris data set.

The result with this data set is much better that the one we have with the Hc Body Temperature data set. Also, while the VC dimension is growing, there is no a big side effect on the errors. That's why, there is no reason to conclude that there is a big overfitting with the Iris data set. Nevertheless, it is not excluded that a little overfitting is actually exist.

Work division

We worked on this code together using one computer as a pair programming. That is, we've handle and understand together the complexity of the adaboost implementation and the code design in python. There is nothing in this work that have been done only by one submitter. Notice that we worked only on one github account since we used only one computer.