Project 2: Classification Algorithms

Code and report due: November 12, 2019, 11:59pm

General Introduction:

In this project, you are asked to implement classification algorithms. Each team should submit codes and a report via UVA Collab.

Dataset Description:

Three datasets (*project2_dataset1*, *project2_dataset2*, and *MNIST*) can be found on Piazza. Here is a short description of the first two datasets:

Each line represents one data sample.
The last column of each line is class label, either 0 or 1.
The rest columns are feature values, each of them can be a real-value (continuous type) or a string (nominal type).
project2_dataset1: 569 observations, 31 attributes
project2_dataset2: 462 observations, 10 attributes

Complete the following tasks (only *project2_dataset1*, *project2_dataset2*):

- Implement four classification algorithms by yourself: **Nearest Neighbor, Decision Tree, Naïve Bayes,** and **SVM**. (Normalize the data to avoid scaling issue, and/or apply regularization to avoid overfitting if needed.)
- Implement **Random Forests** based on your own implementation of Decision Tree.
- Implement **Boosting** based on your own implementation of Decision Tree.
- Adopt 10-fold **Cross Validation** to evaluate the performance of all methods on the provided two datasets in terms of **Accuracy, Precision, Recall**, and **F-1 measure**.
- Note: **All six methods must be implemented by yourself**. Existing packages or online codes for the algorithms are not allowed (mathematical computation packages that are not related to the algorithms, like numpy, are allowed).

Bonus:

Implement neural network (two hidden layers, sigmoid activation function, softmax output layer, and cross entropy loss) with the MNIST dataset. You must implement a neural network from scratch (without tensorflow or pytorch) (20 points).

MNIST dataset description:

The MNIST dataset (Modified National Institute of Standards and Technology dataset) is a dataset of handwritten digits which used to serve as a benchmark dataset for various image processing tasks. Here is the visualization of some examples within this dataset. Current dataset consists of 50k training images, 10k validation images, and 10k testing images. Each image has 28 by 28 pixels (equivalently, 784 features).

Your task is to train a neural network with 50k training samples to classify 10 digits (0-9) and report its classification results on 10k testing images (ignore validation images for now).

We have uploaded a piece of code (mnist_loader.py) on Piazza. You could use the following lines of code to import the mnist dataset with mnist_loader.py.

```
import mnist_loader
training_data, validation_data, test_data = mnist_loader.load_data_wrapper()
training_data = list(training_data)
test_data = list(test_data)
```

(Try different number of hidden units, different weight/bias initializations, different learning rates and discuss whether they affect the performance.)

Project Submission:

- Prepare your submission. Make a zipped folder named "*CompID[-CompID]-Classification.zip*", where "CompID[-CompID]" refers to the list of your group members' computing IDs. In the folder, you should include:

    1. Report: A pdf file named *Classification_report.pdf*. Describe the flow of all the implemented methods, and briefly describe the choice you make (such as parameter setting, pre-processing, post-processing, how to deal with over-fitting, etc.). Compare their performance, and state their pros and cons based on your findings.

    2. Code: A zipped folder named *code.zip*, which contains all codes used in this part (preferably, each algorithm has a separate .py file with informative file name). Inside the folder, please also provide a README file which describes how to run your code.

    3. A collab assignment page has been created for Project 2. Please submit your zipped folder there. One team only needs to provide one submission on collab.

Note that copying code/results/report from another group or source is not allowed and may result in an F in the grades of all the team members.