COMP10200: Assignment 4 – Part 1

Overview

For this assignment, you will explore the use of perceptrons and multi-layer perceptrons on both artificial and real data sets.

This is Part 1, in which you will develop your own code for the perceptron learning algorithm and then test it on the data sets provided.

The Data

You will use 4 data sets for this assignment, which you can download as a zip file from Canvas:

In the zip file, look for the sets named with your student id. The number of features and instances varies in each data set. The last column in each data set holds the classification targets (1 or 0).

Please do not modify the data or change the names of these files.

The Code

The code you use for this part of the assignment should be written by you in Python using numpy. The instructor might help you get started in class, but the rest of the code should be written by you and not sourced, in whole or part, from somewhere else.

Try to make the code economical by using the array and list processing features of python and numpy. You ought to be able to write a loop that loads each data set and tests it without having to hard code any values for the number of columns or the name of each data set.

The Task

Your task is to create an implementation of the perceptron learning algorithm as described in class and in this week's handouts. This week's handout contains an example and an exercise with a solution that you should be able to use to check the performance of your perceptron.

Once you have developed your perceptron, test it on all four data sets. Your code should load each data set, separate it into training and testing data, train a perceptron on the training data, test it on the testing data, and report the accuracy along with the final weights and threshold. Here's an example of what your output could look like:

```
000123456_1.csv: 78% W:[-0.2 0.2 0.2 -0.0] T:-1.3
000123456_2.csv: 99% W:[0.8 -1.2 0.1 1.5 0.9 0.1 1.9 0.3] T:0.4
000123456_3.csv: 49% W:[-1.5 -1.7 -0.5 -0.4] T:0.0
000123456_4.csv: 62% W:[-18.2 -15.3 -11.7 -13.3 -14.6] T:-60.0
```

You should notice a difference in performance between the data sets. **Using the results you got, which sets would you say are linearly separable and which are not linearly separable?** Put your answer to this question into the docstring at the top of your python file.

Extra Stuff

It might be a good idea to figure out how to use the sklearn implementation of perceptron. Then you can check to make sure you're getting about the same results. It might also be interesting to run the data through a decision tree, kNN, or Naïve Bayes learner to see the difference in performance between a linear classifier and a non-linear classifier on each dataset. If you decide to do this, please output results for all the systems you tested and make it clear in the output of your program which result is which.

Handing In

Zip up your code and data files for both parts of the assignment and hand it in to the dropbox. It should be possible for me to unzip your folder and run your code easily (i.e. without having to move files around or make changes to the code). If necessary, include instructions in the docstring with the correct command to run the code.

See the drop box for the exact due date.

Evaluation

Your code will be evaluated for Performance, Structure, and Documentation using the Rubric attached to the drop box. Your answer to the question about the linear seperability of the data sets is worth 0.5 marks and will be included in your "documentation" mark.