COMP10200: Assignment 4 – Part 2

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 2, in which you will run sklearn's MLPClassifier class.

The Data

You will use 5 data sets for this assignment. The first 4 are the ones you used for Part 1.

The 5th data set should be from the UCI Machine Learning Repository. It should be a classification set with more than 100 numerical features and more than 1000 to 5000 instances, and it should be one that you have not used before in any other assignments. It does not have to be a binary classification task.

Here's a link to the repository with these parameters set:

http://archive.ics.uci.edu/ml//datasets.html?format=&task=cla&att=num&area=&numAtt=greater100&numIns=greater1000&type=&sort=nameUp&view=table

The Code

The code you use for this part of the assignment should be written by you in Python using numpy and sklearn.

The Task

Your code should read in the data files, run a multi-layer perceptron to solve each one, and output the accuracy of the result, the number of iterations required, and the configuration of the network (nodes in each hidden layer, plus any other non-default settings you used). You should experiment with parameters to try to get the best performance you can (best accuracy with smallest network and lowest number of iterations required), but you only have to hand in code for the best configurations you can find.

You should also run another classifier as a benchmark to guide you in your search for a good MLP configuration. For example, if you run a decision tree on the same testing/training split and you get 80% performance, but the MLP is only achieving 60% then you should keep searching for a better configuration.

Here's an example output just for one file:

```
File: mydata.csv
Decision Tree: 85% Accuracy
MLP: hidden layers = [5,6], LR = 0.05, tol = 0.0000001
87% Accuracy, 123 iterations
```

Keep in mind that some of the artificial data sets are linearly separable, so you should not won't need a hidden layer. Also keep in mind that the UCI data set you choose might be very difficult to solve. Expect to spend at least an hour trying different configurations once you have the basic code running.

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 separability of the data sets from part a is worth 0.5 marks and will be included in your "Documentation" mark.