COMP10200: Assignment 5 – Part 2

Overview

This assignment is about machine learning techniques beyond supervised classification. You will explore unsupervised clustering in part 1, and regression in part 2.

Part 2

For this part of the assignment, you will obtain a data set suitable for regression, and apply linear regression plus one other regression technique.

The Data

Choose a data set that is suitable for "Multivariate Regression" from the UCI Machine Learning repository. Clicking on the link below will take you directly to the relevant data sets. Do not use the Air-Foil set, or any other set we looked at in class. Do not use a set that you have already used in another assignment.

https://archive.ics.uci.edu/ml/datasets.php?format=&task=reg&att=num&area=&numAtt=&numIns=&type=mvar&sort=nameUp&view=table

If you know of a good data set that is not in the UCI, feel free to use it. But whatever data set you use:

- make sure that it is clear from the description of the data set which column is meant to be the output (the target value you are predicting)
- make sure that you remove or convert any non-numeric inputs.

The Code

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

The Task

Your task is to test different kinds of regression on your chosen data set. You should write a python module that does the following tasks automatically when it is run:

- Load the data and split it into training and testing sets (you only need one split for this). Make sure it's the same split on every run, and make sure it's a reasonably balanced split (i.e. a full range of values for each feature and output is represented in the testing and training sets.)
- Output the size of the training and testing sets and the number of features.
- Output the RSS error (*e*), correlation (*r*), weights, and intercept for a linear regression.
- Output *e* and *r* for one other regression technique (kNN, Tree, or MLP).

    - You should experiment beforehand to find parameters that will show off the best achievable RSS error rate.
    - It might be a good idea to normalize the data for some of these algorithms.
    - For MLP, it might be a good idea to perform several runs on the same testing/training split and report the best result.

- In the comments at the top of your code, state which algorithm performed the best and offer at least one reason why you think it performed better than the other.

Handing In

Zip up your code and your data files for both parts of the assignment and hand them in together. 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). Simply running your code file with Ctrl-Shift-E should be enough for me to see all text and graphical results.

See the drop box for the exact due date.

Evaluation

This assignment will be evaluated based on the coding rubric shown in the drop box. To get full marks, you must follow the documentation standards for the course, in the Student Resources section of Canvas.