CSE5ML: Machine Learning – Assignment Part 2

Semester 1, 2020

Overview

- This assignment contributes **20%** of your final mark in the subject. Please read this sheet carefully before doing your assignment.
- The assignment aims to consolidate your knowledge on **Regression models** and develop your Machine Learning skills.

Policies

- This is an individual assignment.
- Plagiarism is the submission of somebody else's work in a manner that gives the impression that the work is your own. The Department of Computer Science and Information Technology at La Trobe University treats plagiarism very seriously. When it is detected, **penalties are strictly imposed**.

Submission

- The submission of assignment is due on Monday 1st of June 10 am.
- The assignment is consist of a python code and a report of min 1000 words.
- As the assignment contributes over 15% of your final mark, you need to apply for **Special Consideration** to the University if requiring an extension. Please refer to the link below for more details.
  https://www.latrobe.edu.au/students/admin/forms/special-consideration
- Unless a special consideration is given by the University, late submission is **NOT** accepted and the corresponding group will **NOT** be allowed to give the presentation. In this case, **zero mark will be assigned**.
- All submission need to have your student name and number included.

Problem Description

The assignment requires you to develop 3 end to end regression models: SVM, Linear Regression and K-Nearest Neighbors (all have been studied during the course both in lectures and labs). The models should be developed to accurately predict median value of Boston houses in any given suburb, where the suburbs are described based on their provided characteristics:

- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per $10,000
- PTRATIO pupil-teacher ratio by town
- B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in $1000's

Therefore, the 13 first columns would generate X, and the last column (MEDV) would be the label. The dataset is provided for you in the .csv file format in the same folder (it is originaly from UCI Machine Learning repository, at: http://lib.stat.cmu.edu/datasets/boston).

Your code needs to:

- ✓ Load the dataset (you need to use this library : from pandas import read_csv)
- ✓ Describe the data: printing the shape of the dataset, the type of data, and related correlation correlation
- ✓ Visualize the data (at least with pyplot)
- ✓ Split the data into train and test (20% of the entire data for test). You can also use 10% of data for validation if you want.
- ✓ Train, test and compare the performance of the three developed models

Your report needs to

- ✓ be in the pdf format
- ✓ Describe each model. Starting by explaining the basic of SVM, Linear Regression and K-Nearest Neighbors. Then explain what steps you have taken for fine tuning your model (changing the parameters).

- ✓ Compare the performance of each model individually using different parameters; then compare the performance of the best of each model with other models.
- ✓ It also requires you to have references if you have used any

Marking Criteria

Criterion	Contribution
Understandings on the problem	20
Knowledge on 3 regression models	30
Systems setup and learning parameter setting	20
Meanings and interpretation of used metrics for system performance evaluation	10
Explanation and comments on each method	20
Total	**100**