CSCI433/CSCI933: Machine Learning - Algorithms and Applications
Assignment Problem Set #2

Lecturer: Prof. Philip Ogunbona (philipo@uow.edu.au)
School of Computing and IT
University of Wollongong

Due date: Saturday 1800 Hrs of Week 9

Motivation

The goal of this assignment is to design and compare the performance of three regression predictors using the data set provided. To facilitate this learning process, the assignment has been set up as a competition. We are seeking the best prediction accuracy rate obtainable by any group in the CSCI433/CSCI933 class. You are to use the regressiion predictors studied in the class. Each group will have to study the data carefully by reading about the features (variables), particularly the range of plausible values, meaning, method of measurement, etc. It is expected that a good deal of effort will need to be expended on data preparation (scaling, imputation, etc.). The Machine Learning/Python books provided on Moodle will be of great help in this regard. These books could also be used as de facto reference manual for Python modules (ScikitLearn, matplotlib, numpy, scipy, etc.) for Machine Learning. You should refer to the books on Machine Learning (also provided on Moodle) for the theory underlying the various regression predictors used in your experiment. You may also find information at Kaggle website [1] useful.

About the data

This dataset was taken from the Kaggle website [2]. The aim of gathering this data was to design a predictor of house sale prices and practice some feature engineering. You are given three files namely train.csv, test.csv and data_description.txt (all taken from kaggle).

Features/variables The dataset is organised such that each row contains the features for a house. The columns contain 79. Please see the data_description.txt file doe details.

1. (125 Marks) **Task**

   The following should be taken as the specifications of this assignment.

   1. Form a group of no more than 5 and no less than 4 to work on this assignment. Give your group a nice name. Send me the names of members of your group by the Friday of Week 7.
   2. Select the three different regression models studied in the class and design a predictor using the dataset provided.

[1] `https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview/tutorials`
[2] `https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data`

3. Report your best prediction accuracy regularly on Moodle Forum.
4. Submit a ten-page report on your results for grading. See specifications of the report below. Based on the marks awarded to each section of the report, it is hoped that teams will allocate their efforts.

It is advisable that each team divides the work effort in such a way that everyone in the team has opportunity to deepen their theoretical and programming skills in Machine Learning.

Report

Your report should be according to the following format (i.e. headings):

Title (5 marks) - Give your report a nice title and write the names of the members of your team as well as their student numbers.

NO CSCI...

Introduction (10 marks) - Describe the data in your own words and highlight various statistics (mean, variance, etc.) along with any significant observation that could be gleaned from the data. You may include some graphs. But they must be described in your report.

Data preparation (20 marks)- Describe the various methods and implications of the data preparations you undertook. Note that this is very important as it would have significant impact on the accuracy obtained from your predictor. You should discuss how you split the data for training, validation and testing.

不同的回归模型

Predictors (40 marks) - Describe the various predictors you have tested in your experimentation. This is very important because it shows how well you understand the properties of the predictors. It is expected that you will write equations that describe the predictor model.

Evaluation (20 marks) - Describe and justify the methods of performance evaluation you have adopted. State the comparative evaluation estimates and justify the differences. There is a pool of 20 **extra bonus** marks to be shared by the winning teams. This implies that if there are several teams that obtain similar winning accuracy, they will share 20 bonus marks. If there is only one winning team, a bonus mark of 7 is awarded to the team.

Conclusions (30 marks) - You are required to reflect and write about the differences amongst the various predictor models relative to their parameters, amount of data required for training, nature/format of data required and the accuracy obtained. In addition, you are required to reflect and describe any significant trend/observation you discovered with regards to what features may be dominant in determining the sale price of a house. For example, for a given house type, is there a subgroup of features that are more likely to fetch high sales value?

What needs to be submitted?

PLEASE READ VERY CAREFULLY

You are required to submit your 10-page report according to the format specified above. The report should be typed (or typeset using LaTeX) with 11-point font, one-and-half spacing and 1.5 cm all round margin. Submitted report MUST be a PDF file. Any WORD document should have been converted to PDF before submission. Non-PDF reports will not be marked.

You must submit the code for all the predictors used in your experimentation. You must prepare the code that provided the best accuracy in such a way that it can be tested easily.

In addition you must archive or “zip” or “rar” your source code and submit along with your report. Place your report and your archived source code in a folder with your group name and and ”zip” or “rar” the folder before submission. This is **important** because of the way Moodle submission works.

The most popular programming language used in industry for machine learning is currently Python. You must use Python for this assignment. As previously indicated, you should use Python 3.xx and jupyter notebook for your machine learning studies. This will facilitate easy sharing of codes and will allow the markers to be able to run your code easily. Ensure that your code includes documentation and reasonable naming convention for variables and functions. This is worth a bonus 5 marks!

Only one submission is expected from each group.

This assignment is due on the Saturday of Week 9 at 1800 Hours.