STA3010
2019-2020 term 2
Course Project: California Housing Price Prediction

1 Problem

You are given a real data set of California housing price. Descriptions as well as a download link of this data set can be found at `https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html`. To be brief, this data set contains $n = 20,640$ observations on 9 variables. The output variable is ln(median house value), while the 8 input variables are respectively, median income, housing median age, total rooms, total bedrooms, population, households, latitude, and longitude. Note that all the input variables are treated as quantitative/continuous variables. You are supposed to apply (the taught, to be taught, and self-taught) regression techniques for accurate housing price prediction.

2 Tasks

You are supposed to write an **open-ended technical report**, in which the following aspects should better be addressed. **Submission deadline is 23:59 PM, May 1st, 2020**.

- Divide the complete data set into a training data set with 60 percent of the samples (i.e. $n_{train} = 12384$) and a test data set with 40 percent of the samples (i.e., $n_{test} = 8256$), without data overlap. You can randomly select which data samples go to which set. Let us call the training data set $\mathcal{D}_{train} = \{X_{train}, \boldsymbol{y}_{train}\}$ and call the test data set $\mathcal{D}_{test} = \{X_{test}, \boldsymbol{y}_{test}\}$. The training data set is used to train your regression models, while the test data set is used to evaluate the prediction performance. Comment on the characteristics of the data set.

- Do you encounter multi-collinearity problem for your training data set? If there is no severe multi-collinearity problem, then you don't need to perform standarization on the original data set. Otherwise, please use unit length scaling for standarization.

- Fit a multiple linear regression model with the aforementioned 8 input variables to the training data set. Please report R^2 value and R^2_{adj} after the least-squares (LS) fitting. Please also report the test mean-squared-error (MSE) of the trained LS model applied to the test data set. The test MSE is defined in general as follows:

$$MSE = \frac{1}{n_t} \sum_{i=1}^{n_t=8256} (y_{test,i} - \hat{y}_{test,i}(\boldsymbol{x}_{test,i}))^2, \tag{1}$$

  where $\hat{y}_{test,i}(\boldsymbol{x}_{test,i})$ is the fitted value for the i-th input vector $\boldsymbol{x}_{test,i}$ in the test data set. The above test MSE can be regarded as an indicator of the prediction performance of your regression model.

- For the above multiple linear regression model, please compute the R-studentized residuals and plot them versus the fitted values, $\hat{y}$. In addition, please construct a normal probability plot or a Q-Q plot. By combining the two residual plots, what can you say about the zero mean Gaussian i.i.d. assumption on the random error terms, ε_i, $i = 1, 2, ..., n_{train}$?

- Fit a polynomial regression model with the above 8 inputs to the training data set. You can freely choose the order K of the polynomial model as well as the interacting terms. Please comment on the merits and demerits of using the polynomial regression model. Please report R^2 and R^2_{adj} values for the traning phase and test MSE for the test phase. Does it help by adding an L_2 regularization term to the cost function?

- Until now, you have applied two linear models to the data and obtained some results. Next, please try to apply a nonlinear regression model, namely the neural network (NN) model, to fit the data. Please read more about the NN model by yourself with the aid of our carefully prepared document (NN-help.pdf in the folder) or some other reference books. Please use your own words and a few equations (less than 5) to describe the idea of the NN model for regression and explain why it is a nonlinear model and how its model parameters can be trained. Please compare the resulting R^2_{adj} and test MSE with those of the linear models. Please also comment on the merits and demerits of the NN model for regression tasks from optimization and model complexity perspectives but not limited to these.

- **Optional 5 Bonus Points**. Please read the first 8 pages of the paper (Breiman85.pdf in the folder) on alternating conditional expectation (ACE) algorithm, which applies optimal transformations on both the output and input variables of the original data set. Search for the existing ACE algorithm implemented in R or Python. Use for instance `https://github.com/partofthething/ace`. Execute the algorithm and compute the R^2_{adj} and test MSE and compare the results with those of the previous models. Compare the residual plots for the ACE model with the Ordinary LS without performing data transformation. Is the transformed model superior to the original multiple linear regression model? Please also comment on the merits and demerits of the ACE model for regression tasks.

- Are there any outlier points in the traning data set? Is the fraction of outliers large as compared to the ordinary data points? How can we achieve robustness against outliers if any.

- A summary of the main findings obtained from solving the above tasks.

- Comparison of different regression models in terms of, for instance, training performance, prediction performance, overfitting of data, time complexity, etc.

3 Format of the Report

The report should be written in English using WORD or LATEX. The report should contain text for describing your mind flow as well as tables and figures for demonstrating the experimental results. Don't just pile up binary answers and shabby results to the above questions. You don't need to append your codes. The report is limited to 10 pages in length with a fontsize 12 pt.

4 Evaluation

This course project takes 20+5 points, while the assignments takes 30 points and the final exam takes 50 points. You will be receiving one of the levels: 0, 5, 10, 15, 20, 25 points.