IE 7374 Machine Learning in Engineering

LAB as Final
Due in Canvas by EOD Friday April 29, 2022

Application of Linear Models for Regression

The US Census Bureau has published California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The purpose of the lab/final is to predict median house values in Californian districts, given many features from these districts. The dataset attached in this lab is adapted from kaggle[1]. This lab aims at building models of housing prices to predict median house values in California using the provided dataset. Districts or block groups are the smallest geographical units for which the US Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the dataset. You are allowed to work in a team and use any Python libraries such as Scikit-learn to implement all steps below. Please cite your references should you use any.

1. **Visualize and pre-process the data**. Identify missing or null values in the dataset and then either impute or disregard. Further create 5 bins for variable of median income. Identify categorical variables in the dataset and further apply one-hot encoding.

   1). Plot the histograms of all variables.

   2). Do the scatter plots of y=”latitude” vs. x=”longitude” by population size and median house values.

2. **Build an ininitial model**. Split data to train and test and further build a linear regression model including all identified variables to predict median house value. Perform cross-validation and report your performance metrics.

3. **Enhance your model**. Further improve the linear regression model you have built to include regularization terms and enhanced features. Report your performance metrics. As a kind reminder, your results will be compared with peers in terms of the improvement journey you took and final prediction accuracy.

Submission

When you turn in your assignment, please submit electronically (on Canvas) **BOTH** your results and your Python code that you used to solve the problems. Please make sure you codes are reproducible.

[1] https://www.kaggle.com/subashdump/california-housing-price-prediction/data