Summer 2018 EE690 Machine Learning Dr. L. Jololian

Final Project

**Due Date: Tuesday, August 7th**

Please download the dataset for this project from the Kaggle Web site using the following URL: <https://www.kaggle.com/christophercorrea/dc-residential-properties> . The name of the file containing the dataset is **DC\_Properties.csv**. The dataset consists of real property information, including most recent sales price as of July 2018, for properties located in Washington, D.C.

The total number of rows in the file is approximately 160,000. There are forty-nine columns (features) for each data sample. Not every data sample is complete; there might be missing values in some of the rows.

For this project, to make the data size more manageable, we will focus only on the data with SOURCE=’Residential’. This will reduce the data size to about 107,000 rows.

Write a Python program that can predict the price of a house based on a set of provided features.

**Implementation**

1. Implement a sensible approach for dealing with the missing data to get a high level of confidence in your predictive models.
2. Use linear regression to create a predictive model.
3. Convert the PRICE column into a categorical value, such that you replace the original value with one of ten (10) categories. For example, 0 will be substituted for the price range of 0.0-30,000.
4. Use Logistic regression to create a predictive model using the modified data described in (c).
5. Use a neural network to build a predictive model for the same data used in (d).

**Work to Submit**

For (a) above, provide an explanation of the work that you have done.

For (b) above, provide the Python code for the implementation using the Scikit-Learn library. Describe the performance of your answer by using appropriate metrics.

For (c) above, provide the code for transforming the values in the PRICE column.

For (d) above, provide the code for implementing logistic regression using your own Python code and without using the Scikit-Learn library. Using appropriate metrics, show the effectiveness of your model. Show the value of the cost as a function of the number of iterations in your algorithm.

For (e) above, you have a choice of implementing your solution by either using the Scikit-Learn library or your own Python code implementation. In either case, provide the code and the performance metrics.