Machine Learning: Lab 2 – Linear and Logistic Regression

Download the House Rent Prediction dataset from
https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset

Prerequisites: Python basics, numpy, pandas, matplotlib, sklearn, etc.

Importing Data:

1. Randomly shuffle the dataset by taking a random seed of “42”. Create a testing set from the last 1000 rows of the dataframe (these must be the same for all the students). The remaining rows will be the training + validation set, with training : validation ratio of 80% : 20%. Determine

A) number of rows in training, validation and test sets, along with the structure, datatypes and value counts of the dataframes.

Data Cleaning:

1. Analyse the data and identify which columns are not relevant for house rent prediction task. Drop those columns from the dataframes.

2. Check for missing values and logically impute the dataset.

3. Identify any categorical valued columns (non-numeric) and convert them to numeric.

Exploratory Analysis (On training set):

1. Plot the house rents against the dependent variable of “size”. See if there is a uniform linear trend between the dependent and independent variables. Make accurate axis and legend. Save the plot in a png file.

2. Find average rent prices in different cities and report which city has the highest average rent.

Regression:

1. Train a linear regression model on the training set partition by taking only one dependent variable of “size”. Calculate the error on the validation set.

2. Plot the model predictions of rent values alongside the actual rent values taken for the validation set. Show the legend, axes and color-coded predictions and ground truth for differentiating.

3. Create a function for calculating the RMSE values for the predictions Vs the actual ground truth rent values. **RMSE = SQRT(Σ ((F(x_i) - y_i)^2)/N),** Here F(x) are the prediction values, N are the number of rows.

4. Train a logistic regression model and check the score for different training iterations. Plot the validation results by varying max_iter as 10, 20, 30, ….

5. Try to improve accuracy (on validation set) by considering more features and retraining.

6. Make predictions on the test set by taking 3 of your best models. Report these 3 accuracy values.