DS203: Programming for Data Sciences

Assignment 6: Linear and Logistics Regression

Exercise 1 (Linear Regression)**.** *This exercise aims to help you learn the application of Linear Regression for real estate evaluation. It involves understanding what data means, how to handle data, training the model, prediction, and testing your model. We will try to do the complete flow in this assignment.*

Dataset: *The market historical data set of real estate valuation is collected from Sindian District, New Taipei City, Taiwan. The dataset can be downloaded from here.*

Features Information:
The features of the dataset are as follows
X_1 = *the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)*
X_2 = *the house age (unit: year)*
X_3 = *the distance to the nearest MRT station (unit: meter)*
X_4 = *the number of convenience stores in the living circle on foot (integer)*
X_5 = *the geographic coordinate, latitude. (unit: degree)*
X_6 = *the geographic coordinate, longitude. (unit: degree)*

The output is as follow
Y = *house price of unit area (3000 New Taiwan Dollar/meter squared)*

More details about the dataset can be found in this web-page. You can use linear regression, Ridge regression, and LASSO models for this. Perform the following steps.

1. *Download the dataset from above shared link.*
2. *Load the dataset into your python program and do pre-processing (remove the first row and first column as they are not useful).*
3. *Split loaded dataset into train and test dataset by keeping 80% samples in train dataset and remaining 20% samples in test dataset.*
4. *Now train linear regression model on train dataset (note that the last column (house price) is the output).*
5. *Report coefficients (weights corresponding to features) and intercept of trained model.*
6. *Predict price for every house (sample) in test dataset.*
7. *Compute mean squared error and r^2 value using predicted price and true price.*
8. *Repeat Step 4-6 for following train and test split: 60:40, 70:30, and 90:10. Report mean squared error and r^2 value for each split.*
9. *Use Ridge regression and Lasso models with following λ values (regularization parameter): 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5. Report mean squared error and r^2 value for each alpha value with all train and test split ratios given in Step 7.*

Exercise 2 (Logistic Regression)**.** *In a study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast*

cancer. The dataset can be downloaded from here. In this exercise we will predict the survival status of patients using logistic regression. The data has the following features:

Features Information:

1. *Age of patient at the time of operation (numerical)*
2. *Patient's year of operation (year - 1900, numerical)*
3. *Number of positive axillary nodes detected (numerical)*
4. *Survival status (class attribute)*
   - *1 = the patient survived five years or longer*
   - *2 = the patient died within five year*

You can find more details about the dataset in this web-page.

Perform the following steps.

1. *Download the dataset from above shared link.*
2. *Load the dataset into your python program and do pre-processing (like separating features and class labels).*
3. *Split loaded dataset into train and test dataset by keeping 80% samples in train dataset and remaining 20% samples in the test dataset.*
4. *Now train logistic regression model on train dataset (note that the last column (Survival status) is the class label).*
5. *Report coefficients and intercept of the trained logistic model.*
6. *Predict the survival status of patients in the test dataset.*
7. *Compute classification error using predicted survival status and true survival status as follows:*

$$\gamma = \frac{1}{m} \sum_{s=1}^{m} \mathbb{1} \{\hat{y}_s \neq y_s\}$$

   where m is the number of samples in the test dataset, $\hat{y}_s$ and y_s are the predicted label and the true label for the sample 's' of the test dataset, respectively.
8. *Repeat Step 4-7 for following train and test split: 60:40, 70:30, and 90:10. Report classification error for each split.*