ML Assignment

Logistic regression: - Gradient Descent Method

Dataset: - Breast Cancer

Prediction: - Predict diagnosis (B/M)

Data Analysis: -

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
…

Data Preprocessing: -

- Removing the column Id as it doesn't add anything to the dataset.
- Check for any null or NaN values (No Null values Found)
- Replaced the characters 'M' and 'B' from the dataset with 0 and 1
- Check for correlated values and removed them if correlation is greater than .9
- Standardize using StandardScalar()

Code Approach: -

- Created a class named Logistic Regression
- Implemented the code as per the theory told in the class.
- Divided the dataset manually using random, sample and seed methods
- Fit the model using training dataset

Code Evaluation: -

- Created methods to check f1 score
    - Calculated recall
    - Calculated Precision
- Checked accuracy of training data and testing data using f1 score
- F1 score of the testing data is :- 0.856269113149847
- F1 score of the testing data is :- 0.8901734104046243

Linear regression: - Gradient Descent Method

Dataset: - Boston Dataset

Prediction: - Predict MEDV

Data Analysis: -

data: contains the information for various houses

target: prices of the house

CRIM: Per capita crime rate by town

ZN: Proportion of residential land zoned for lots over 25,000 sq. ft

MEDV: Median value of owner-occupied homes in $1000sDESCR: describes the dataset

Data Preprocessing: -

- Removing the column Id as it doesn't add anything to the dataset.
- Check for any null or NaN values (No Null values Found)
- Check for correlated values and removed them if correlation is greater than .9 (Found one column 'TAX')
- Check if column contains zeroes (Found 'CHAS' = 93% and 'ZN'=73%) removed them.
- Manual Normalization using mean and standard deviation
- Divided the dataset manually using random, sample and seed methods

Code Approach: -

- Defined cost function and gradient descent method
- Plot the reducing error values which nearly becomes constant after 80 iterations
- Defined a Predict method and mean square error method
- Fit the model using predict method

Code Evaluation: -

- Predicted the values of training and testing dataset using mean square error and the result comes out to be :-
- The Mean Square Error of the training data is :- 0.253951246073690
- The Mean Square Error of the testing data is :- 0.33788074492747827

Classification: Naive Bayes

Dataset: - Breast Cancer

Prediction: - Predict diagnosis (B/M)

Data Analysis: -

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
…

Data Preprocessing: -

- Removing the column Id as it doesn't add anything to the dataset.
- Check for any null or NaN values (No Null values Found)
- Replaced the characters 'M' and 'B' from the dataset with 0 and 1
- Check for correlated values and removed them if correlation is greater than .9
- Normalize using mean and standard deviation

Code Approach: -

- Defined class NaiveBayesClassifier that has following methods :
    - **prior_probability :** calculate prior probability P(y)
    - **mean_std** : calculate mean, variance for each column and convert to numpy array
    - **gaussian_density :** calculate probability from gaussian density function (normally distributed) we will assume that probability of specific target value given specific class is normally distributed.
    - **posterior_probability :** calculate posterior probability for each class
    - **fit :** Fit the model using
    - **predict:** Predict the output from given input
    - **accuracy :** Check the accuracy
- Splitting the data manually using random seed methods
- Fitting the model using fit function

Code Evaluation: -

- Predicting the values of y using predict method and evaluated using the method predict
- The accuracy comes out to be :
    - The accuracy of the model is 0.956140350877193

Linear regression: - Closed Form Method

Dataset: - Boston Dataset

Prediction: - Predict MEDV

Data Analysis: -

data: contains the information for various houses

target: prices of the house

CRIM: Per capita crime rate by town

ZN: Proportion of residential land zoned for lots over 25,000 sq. ft

MEDV: Median value of owner-occupied homes in $1000sDESCR: describes the dataset

Data Preprocessing: -

- Removing the column Id as it doesn't add anything to the dataset.
- Check fo r any null or NaN values (No Null values Found)
- Check for correlated values and removed them if correlation is greater than .9 (Found one column 'TAX')
- Check if column contains zeroes (Found 'CHAS' = 93% and 'ZN'=73%) removed them.
- Manual Normalization using mean and standard deviation
- Divided the dataset manually using random, sample and seed methods

Code Approach: -

- Defined a class named LinearRegression that contains following methods:
    - compute_theta : computes the optimal values of theta
    - fit_model : fit the model using training features and training output
    - coef_ : Store the coefficient values
    - intercept_ : Store the intercept values

    - predict_output : Predict the output on training and testing data
- Defined train test split model to divide the sample into training and testing sets
- Fitting the model and store the optimal value of theta
- Predicting the values of y using predict method

Code Evaluation: - :

- Predicted the values of testing dataset using mean square error and the result comes out to be :-
    - The root mean square value is 0.33775570691502577