

Name: Harshitha Komatineni

Student Id: 1001968082

MACHINE LEARNING PROJECT -1 Report

Introduction:

The iris data set is part of sklearn and it consists of 3 different types of irises, petal and sepal length stored in a 150 *4 numpy.ndarray. The rows of the iris data set are the rows being the samples and columns are sepal length, sepal width , petal length and petal width .

Problem Statement:

I have applied gradient descent in which five parameters including bias have been optimized.

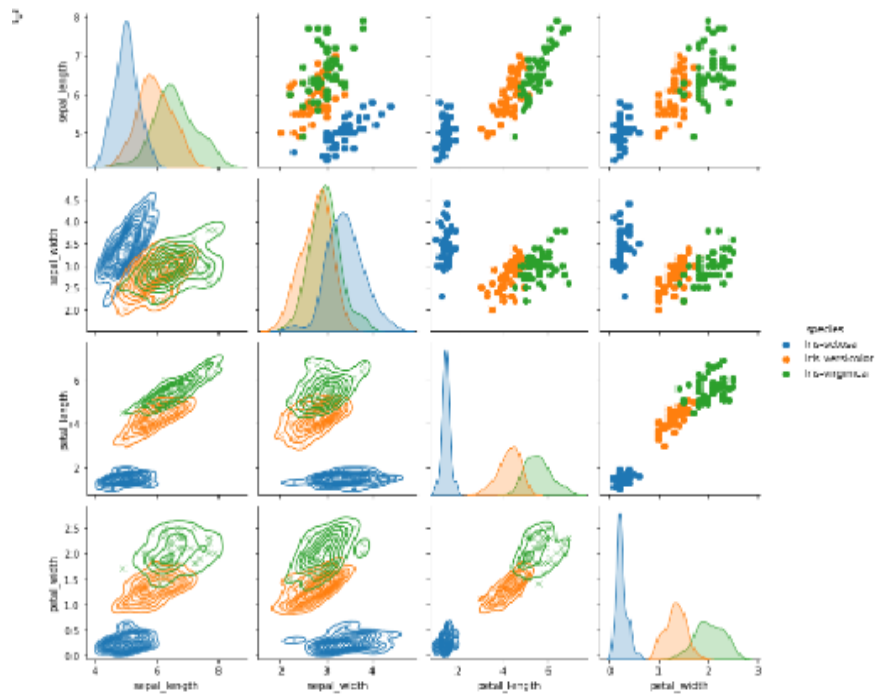
I have shuffled the dataset according to train and test each time training data was made and gradient descent was called and on optimized parameter test data was tested and mean square error was given.

Data:

IRIS data set where 150 samples are over there and 3 classes are over there for setosa, vericolor, Virginica in species data.

The species was categorical therefore, I changed it to ordinal feature by mapping to 1,2 and 3.

Preparation Methods:



Methods:

1. Load the data in a Pandas data frame, which is nothing but data preprocessing.
2. we will be analyzing the data.
3. I have initialized theta with random 5 values including bias then with those theta values I have fitted my line and calculated predictions.
4. I have got error which I used to make changes in parameters and used those new parameters to update the theta for next iteration. I have used learning rate and 1000 iterations to get updated theta.
5. Cross validation:
I have 10 times shuffled the data for training and testing to get trained parameters for trained data and used updated one for prediction.

Results are as follows:

```
theta, cost_history = gradient_descent(X, y, theta, alpha, iterations)
```

```
print('Final value of theta =', theta)
```

```
print('First 5 values from cost_history =', cost_history[:,5])
```

```
print('Last 5 values from cost_history =', cost_history[-5
```

```
:]theta, cost_history = gradient_descent(X, y, theta, alpha, iterations)
```

```
theta, cost_history = gradient_descent(X, y, theta, alpha, iterations)
print('Final value of theta =', theta)
print('First 5 values from cost_history =', cost_history[:, 5])
print('Last 5 values from cost_history =', cost_history[:, -5])
```

```
Final value of theta = [[1.73347054]
```

```
[0.21447671]
```

```
[0.88662585]
```

```
[2.10955625]
```

```
[1.82394774]]
```

```
First 5 values from cost_history = [105.08991218  0.          0.          0.          0.          ]
```

```
Last 5 values from cost_history = [105.08991218  0.          0.          0.          0.          ]
```

The mean square error for each iteration is shown below when model is trained for each train data and tested on test data.

```
8.763157894736842
```

```
cross validation score for 2 comparisons: 8.76 (+/- 0.00)
```

```
8.078947368421053
```

```
cross validation score for 3 comparisons: 8.08 (+/- 0.00)
```

```
7.815789473684211
```

```
cross validation score for 4 comparisons: 7.82 (+/- 0.00)
```

```
7.368421052631579
```

```
cross validation score for 5 comparisons: 7.37 (+/- 0.00)
```

```
7.052631578947368
```

```
cross validation score for 6 comparisons: 7.05 (+/- 0.00)
```

```
6.631578947368421
```

```
cross validation score for 7 comparisons: 6.63 (+/- 0.00)
```

```
6.394736842105263
```

```
cross validation score for 8 comparisons: 6.39 (+/- 0.00)
```

```
6.0
```

```
cross validation score for 9 comparisons: 6.00 (+/- 0.00)
```

References:

- 1) [Linear Regression | Kaggle](#)
- 2) <https://scikit-learn.org/stable/>
- 3) [Classification Basics: Walk-through with the Iris Data Set | Towards Data Science](#)
- 4) [How to Configure k-Fold Cross-Validation \(machinelearningmastery.com\)](#)