Introduction to Machine Learning Program Assignment #3

TA's name: 王培霖 Pei-Lin Deadline: 2018/12/07 (Fri) 23:59:59

TA's email: barry84090371@gmail.com

I. Problem

1. Linear regression with single variable by built-in function

   To choose the most relative features, you should plot all the features with the target first. Then calculate the weight and the bias to make the simple prediction, also record your accuracy.

2. Linear regression with single variable by your own gradient descent

   In this case, you should implement linear regression with your own gradient descent model. With the random initial weight and bias(you should give a reasonable number), updating the learning rate to get your result.

3. Linear regression with multi-variable by your own gradient descent

   In this case, you should build a multi-variable model. Record the MSE(mean squared error)and the R^2(coefficient of determination) for both training and testing. Also compare each iteration only update w_j and each iteration updates w.

4. Polynomial regression by your own gradient descent

   In this case, you should build a quadratic or higher(i.e. the degree of polynomial is >=2) polynomial model. Make your R^2 higher as you can. Record the MSE and the R^2 for both training and testing.

5. (Bonus) Making different regression model to make the accuracy > 0.87

   Hint: You can also change your loss function

II. Dataset

Wine-quality dataset

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Number of Instances: white wine - 4898.

Number of Attributes: 11 + output attribute

Attribute information:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Testing data = 20% of the whole dataset

III. Report & Scoring

This is a team-based program assignment, so one team should only submit one report and one source code to E3.

The report should contain the following:

1. What environments the members are using (5%)
2. Visualization of all the features with the target(5%)
3. The code, graph, accuracy, weight and bias for problem 1(10%)
4. The code, graph, accuracy, weight and bias for problem 2(25%)
5. Compare Problem1 and Problem2, show what you got.(5%)
6. The code, MSE, R^2, and the accuracy for problem 3(20%)
7. Compare the performance between two different update method.
8. The code, MSE, R^2, and the accuracy for problem 4(20%)
9. Answer the question(5%)
    1. What is overfitting?
    2. Stochastic gradient descent is also a kind of gradient descent, what is the benefit of using SGD?
    3. Why the different initial value to GD model may cause different result?
    4. After finishing this homework, what have you learned, what problems you encountered, and how the problems were solved?
10. Bonus(10%)

There are some rules to follow:

- C / C++ / Java / Python / Matlab are allowed to use. For visualization, Excel or other programs are allowed.
- Report format should be PDF.
- Attach your code when you are submitting.
- No cheating and plagiarizing.
- Delay：Your score *= 0.8