Regression-My First Assignment of Pattern Recognition

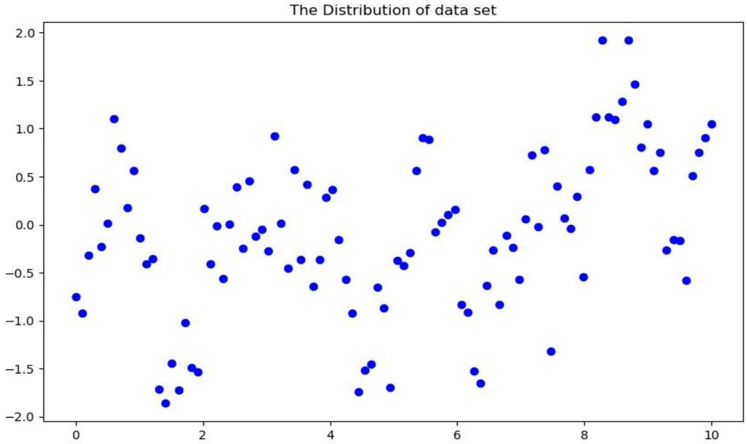
--Peilin Feng

Dear Mr.Qin:

In order to urge myself connect Chinese proper nouns to English academic vocabulary ,I want to finish this homework in English.Thanks for your consideration!

Problem1:The Analysis of The Data Set

A) Draw the data set distribution initially



What we can get from the draft is that the data distribution isn't linear!So there is need for us to do some improvements on the linear model that we just learned.By the way,I find that the trendy of distribution looks like trigonometric functions and polynomial function.

Besides,it's hard for us to separate some specific part individually cause the distribution of the set is uneven.If we extract a certain part of the data,the model we get ultimately will be inaccurate!

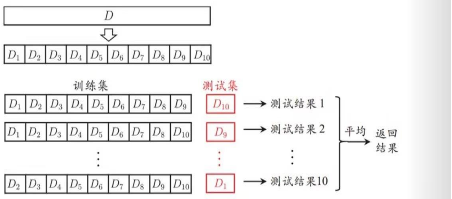
These two messages we get from this draft are useful,which will direct us to take a better strategy to classify the data set and adapt a proper model.

Problem2:The Classification of The Data Set

B) Data set classification

In *Machine Learning* written by Zhou zhihua,there delivers me three methods to classify the data set we get : Hold-out method, Bootstrap method and K-fold Cross Validation.Here we have 100 samples.So we K-fold Cross Validation method is useful.We take K=10 as an example.

The process of 10-fold Cross Validation is as follows:

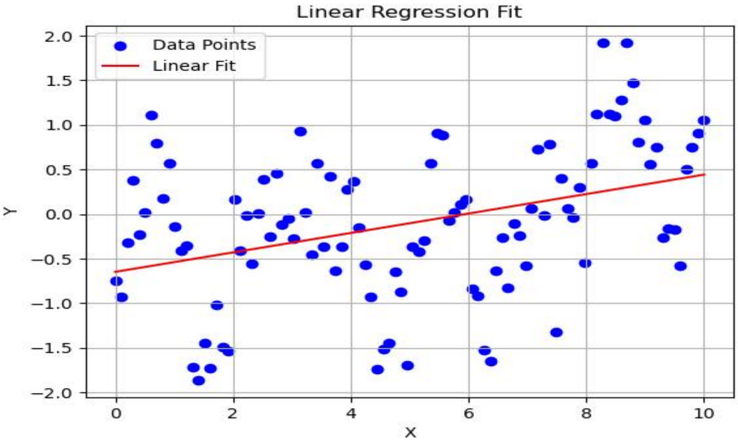

D
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
训练集
测试集
D1 D2 D3 D4 D5 D6 D7 D8 D9
D10
测试结果 1
D1 D2 D3 D4 D5 D6 D7 D8 D10
D9
测试结果 2
D2 D3 D4 D5 D6 D7 D8 D9 D10
D1
测试结果10
平均
返回结果

Problem3:Three Solutions to This Problem

C) Using linear model as an beginning

Although it's unreasonable to use linear model to fit these data points,it's helpful for me to know the error basically.

Here is the "Linear Regression Fit" using MSE loss.



The Mean Squared Error is 0.8179.

Specifically,the result of these training processes is as follows:

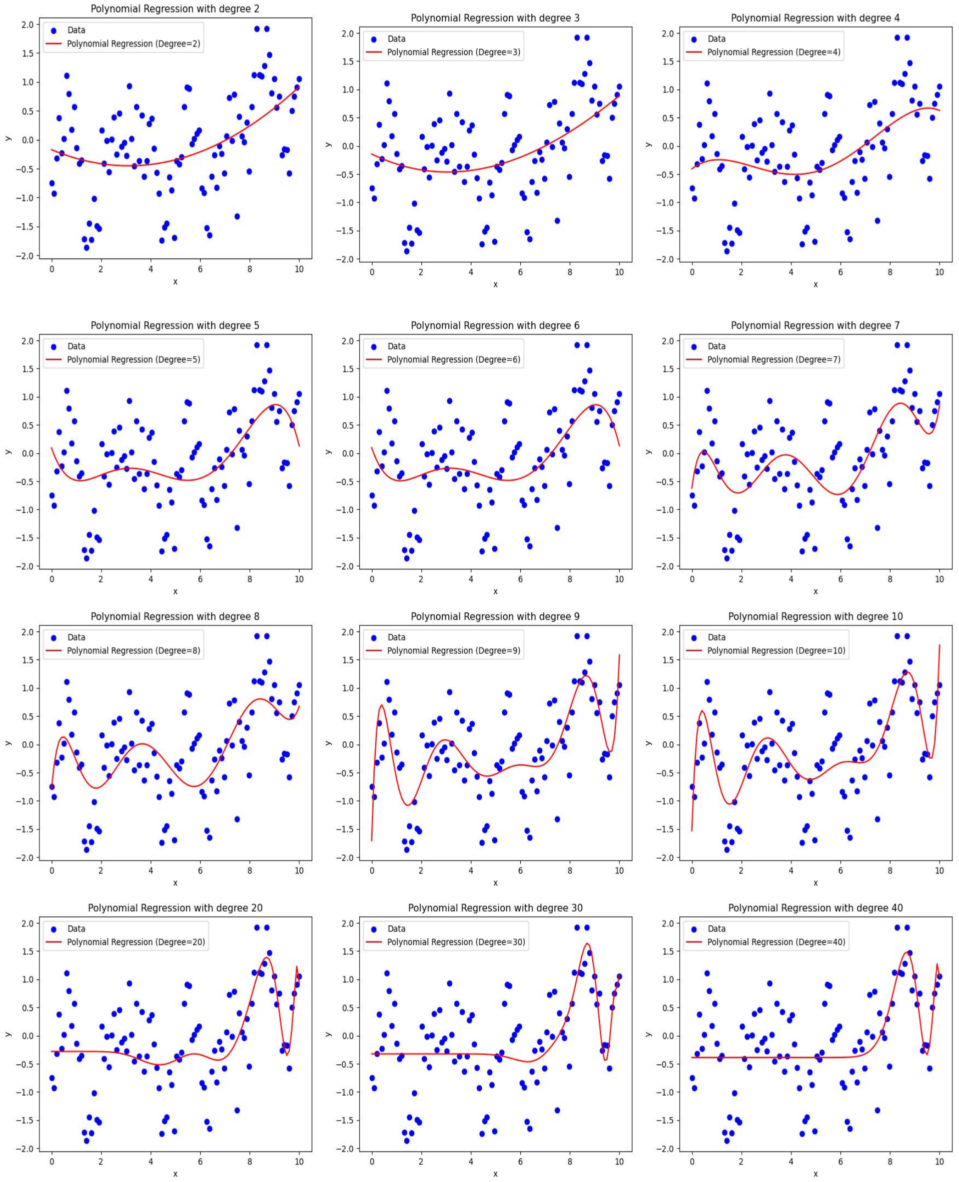
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Ave
1.392	1.195	0.260	0.351	1.183	0.256	1.099	0.381	1.772	0.290	0.8179

D) Using polynomial model to fit better

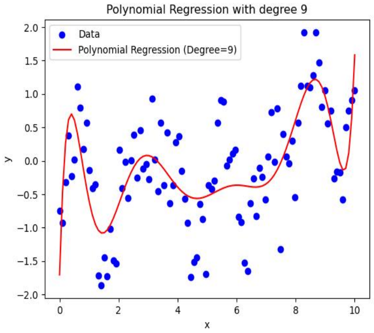
I tried a great number of degrees to fit the data set.So it's hard to record every result of training process.So here I just calculate the average of MSE loss as follows:

Degree of Polynomial	Training/Testing Average Loss
2	0.5637/0.5980
3	0.5636/0.5985
4	0.5527/0.6192
5	0.5194/0.6762
6	0.5194/0.6470
7	0.4615/0.5343
8	0.4623/0.5401
9	0.3488/0.4805
10	0.3486/0.4831
20	0.4498/0.9082
30	0.4448/0.5139
40	0.4478/0.6801

Here are the excerpts of the result:



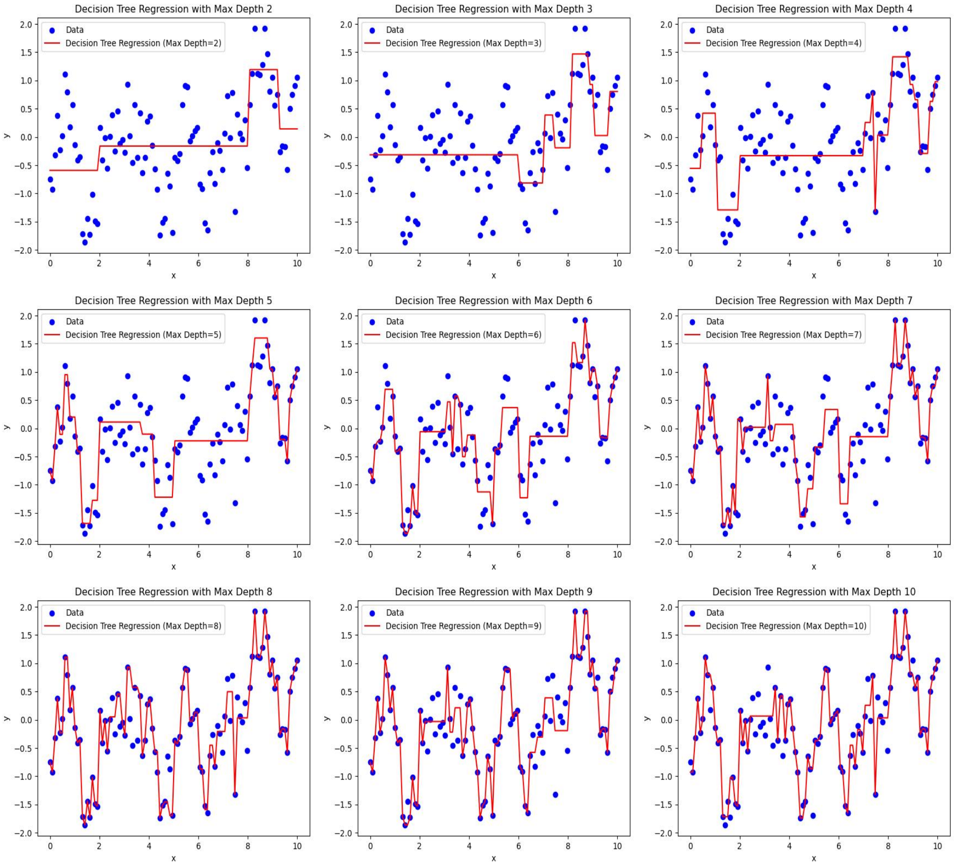
We can find that the result will be better when we use proper degrees of polynomial.However,we can easily find that it is not that the higher the polynomial degree, the better the fitting effect.When the degrees become ridiculously high,the error will increase because of the flat part that occurs when x is small.And the best degree of polynomial we get is 9.The model is drawn in the following picture:



Perhaps this result is already a satisfactory one for us because the loss of test data based on 10-fold cross validation method.However,this model is limited to the former of the polynomial.I believe that there are still some better models based on the *"No Free Lunch Theory"*.

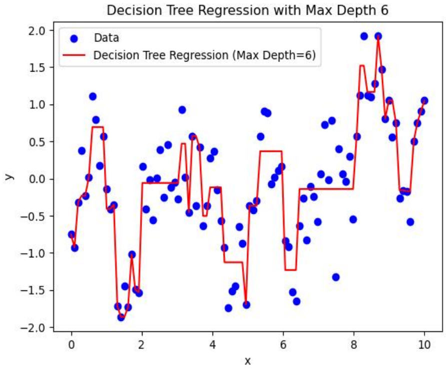
E) Using decision to get more proper result

The max depth of decision tree	Training/Testing Average Loss
2	0.4521/0.5980
3	0.3430/0.4652
4	0.2458/0.4573
5	0.1824/0.4136
6	0.1450/0.3597
7	0.0786/0.4190
8	0.0438/0.3808
9	0.0245/0.3980
10	0.0102/0.4408



Based on the table,we can find that the decision tree model performers better than the polynomial model.Nevertheless,this method is easy to encounter over-fitting phenomenon.Specifically,when the max depth of decision tree is more than 6 ,the variance of the model will be greatly high,which is called over-fitting phenomenon.

The best solution we get is 6-max-depth decision tree now.Here is the result：



We can find that this model simulate the trendy of the data well actually.Whereas there is still some better solutions.And I need to accumulate them in the future.

Summary:

In a nut shell,in this homework ,I tried linear regression,polynomial regression and decision tree to fit the data we get.Ultimately,I find the 6-max-depth decision tree is the best solution I find now.Looking at the final result I made,I find it valuable to move forward step by step during the process of this work!