# CS584. Machine Learning: assignment 1

## Parametric Regression

A20364639 Yung-Chi Liu

## 1. Problem statement

The problem in this assignment is to implement techniques parametric regression. Base on two kinds of regression, single variable and multivariate regression. Separate data set to training data and testing data, and evaluate the performance use 10 fold cross validation.

First, use linear modal for each data set with single variable regression. Observe training error, testing error and performance, and classify the model fit data or not. Then, use different polynomial models on data, and do the same process as linear modal

Second, do multivariate regression model. Map to higher dimension space using combination of space

## 2. Proposed solution

Single variable regression:

Linear model:

With data set Xi and Yi. Separate to Xi-training, Yi-training and Xi-testing, Yi-testing.

Calculate θ.

$$\theta = \begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix} = \begin{bmatrix} M & \sum Xi \\ \sum Xi & \sum (Xi)^2 \end{bmatrix} \begin{bmatrix} \sum Yi \\ \sum Xi \cdot Yi \end{bmatrix}$$

M is size of training data set. Than calculate $\widehat{Y}$

$$\widehat{Y}_i = \theta_0 + \theta_1 X_i$$

And get error by

$$\frac{1}{N}\sum\left(\widehat{Y}_i - Y_i\right)^2$$

Polynomial model:

Set Z , n is the polynomial number.

$$Z = \begin{bmatrix} 1 & {X_1}^1 & {X_1}^2 & \cdots & {X_1}^n \\ 1 & {X_2}^1 & {X_2}^2 & \cdots & {X_2}^n \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & {X_n}^1 & {X_n}^2 & \cdots & {X_n}^n \end{bmatrix}$$

Calculate θ

$\theta = (Z^{-1} \cdot Z)^{T} \cdot Z$

and calculate $\hat{Y}$ same as liner model but in higher dimension.

# 3. Implementation details

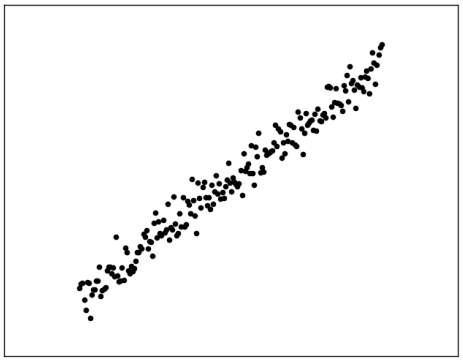
1. Program design
    1. Data Structure: In this program I use numpy.array as data structure. With numpy.array I can reshape the data to the matrix I want and get the value more specifically.
    For example: Get value of θ, the equation $\theta = (Z^{-1} \cdot Z)^{T} \cdot Z$ could be done by

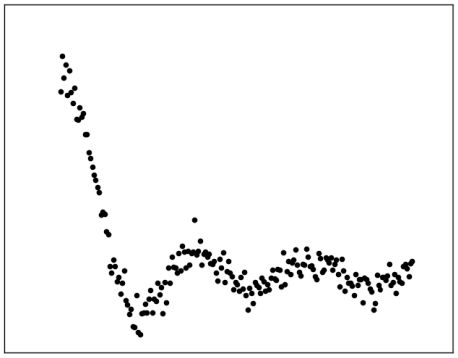
```
bigZ  = np.dot( inv(np.dot(np.transpose(x), x))
                ,np.transpose(x))
theta = np.dot( bigZ, y)
```

    2. Read Data design: I use read line method, and put the last value in y[], the rest put in x[] in order. So, the size of data could be get by size of y[]. And reshape x[] to the fit matrix for calculation.

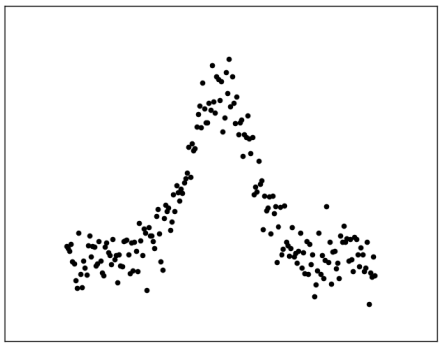
# 4. Results and discussion

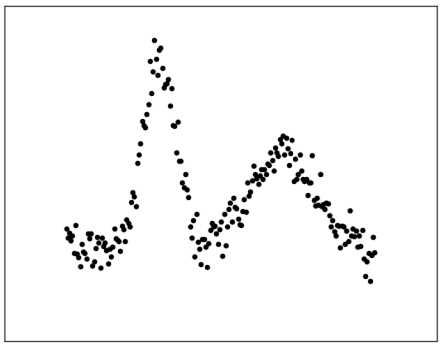
1. Plot the single feature data





svar-set1.dat – data is a rigid line shape, which can be use linear model to fit the data.
svar-set2.dat- data with wave, which may use polynomial model to get close values.

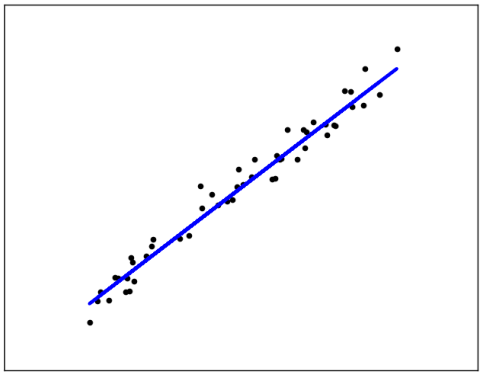


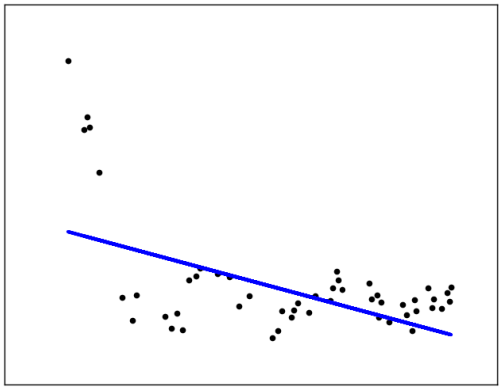


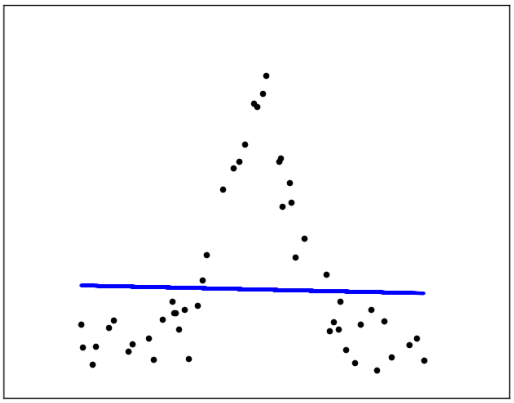
svar-set3.dat – data set looks like normal distribute.

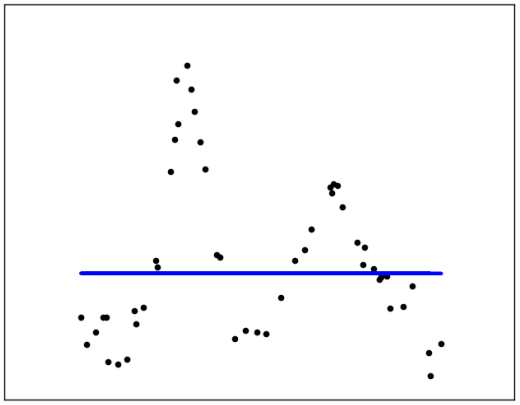
svar-set1.dat – data looks like complex sign wave, which use polynomial model.

2. Fit with leaner model.





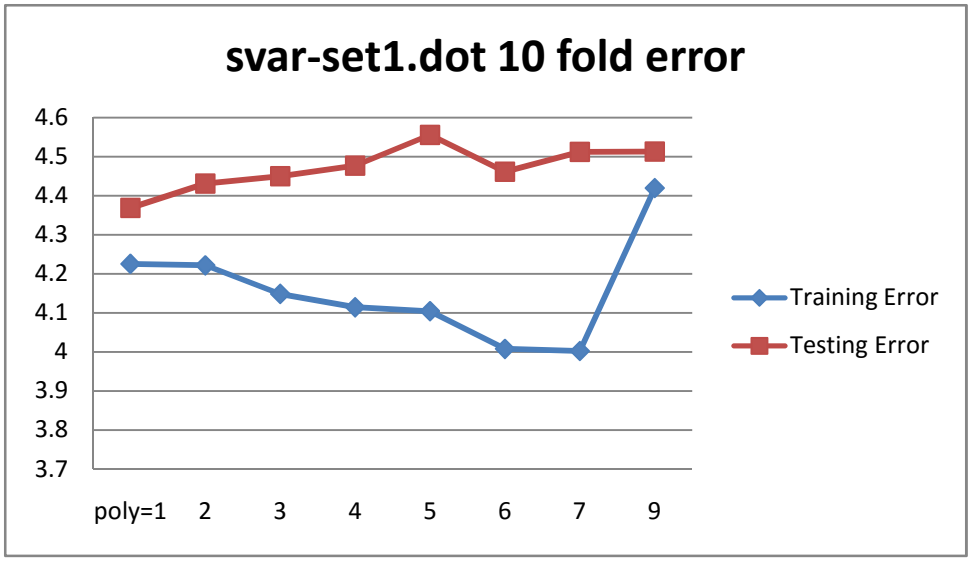




| | DataSet1 | DataSet2 | DataSet3 | DataSet4 |
|---|---|---|---|---|
| Training Error | 4.225492682 | 0.059487187 | 0.498417166 | 1.200202502 |
| Testing Error | 4.368752565 | 0.061200693 | 0.504339248 | 1.212287048 |

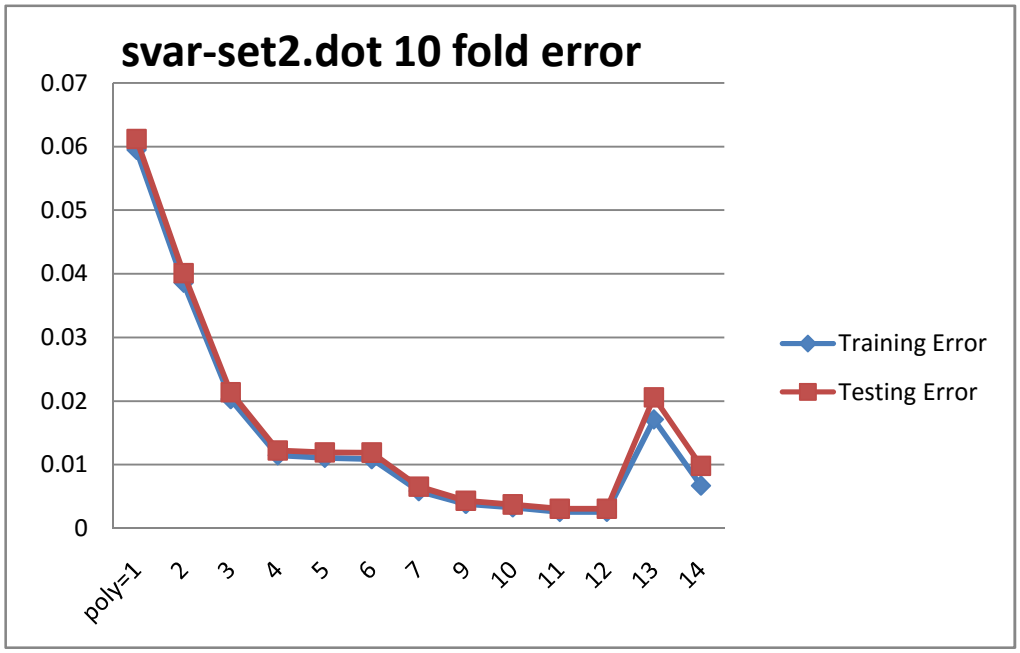
With picture above, data set 2, 3 and 4 is not fit for linear model.

3. Polynomial model

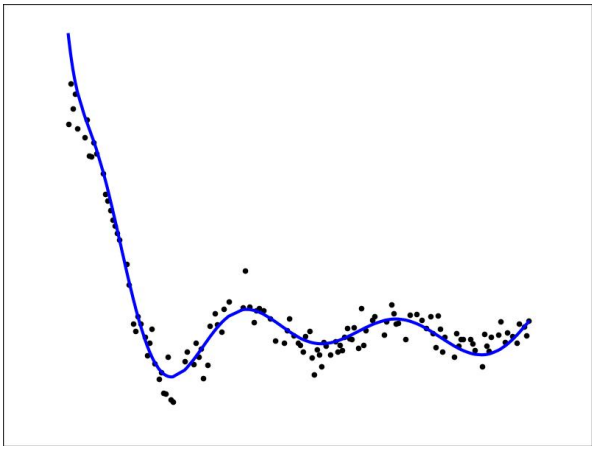



As the result svar-set1.dat 's data is best fit when using polynomial model when number of polynomial is 1, which is same as linear model.

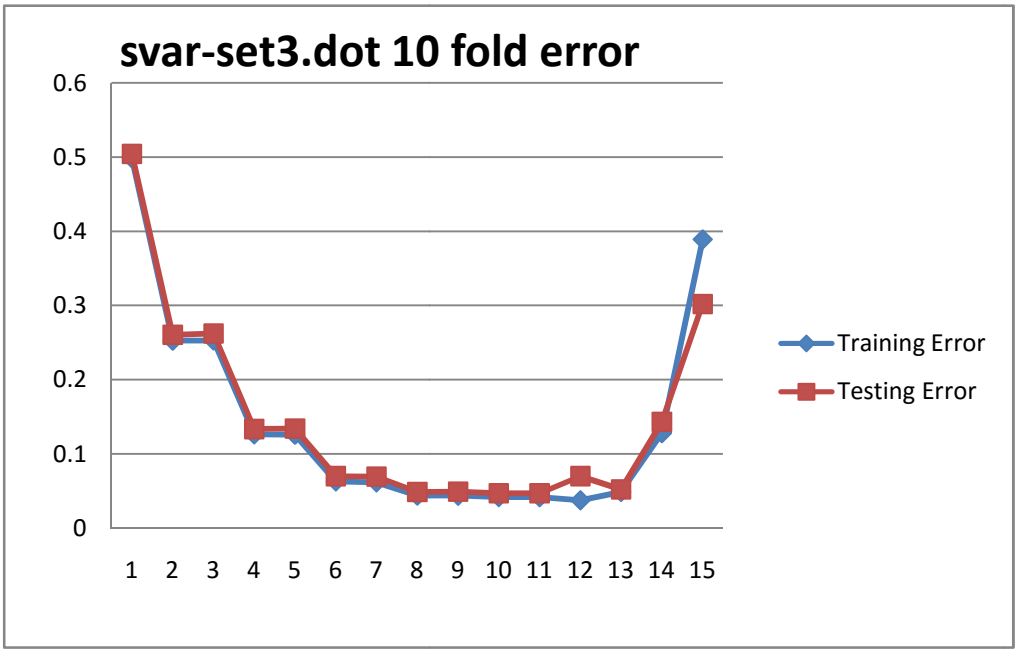
◆ Data set ( svar-set2.dot)



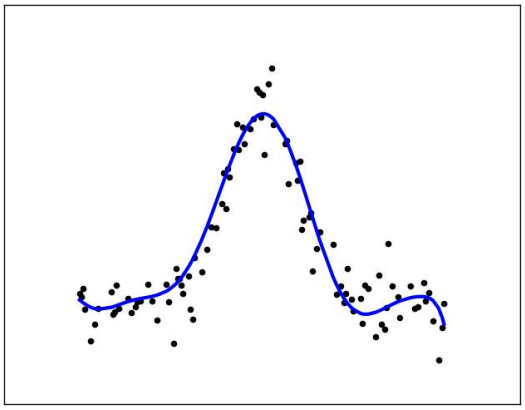

As the svar-set2's data, when polynomial 11 get the lowest Testing Error. So I choose polynomial 11 as model.



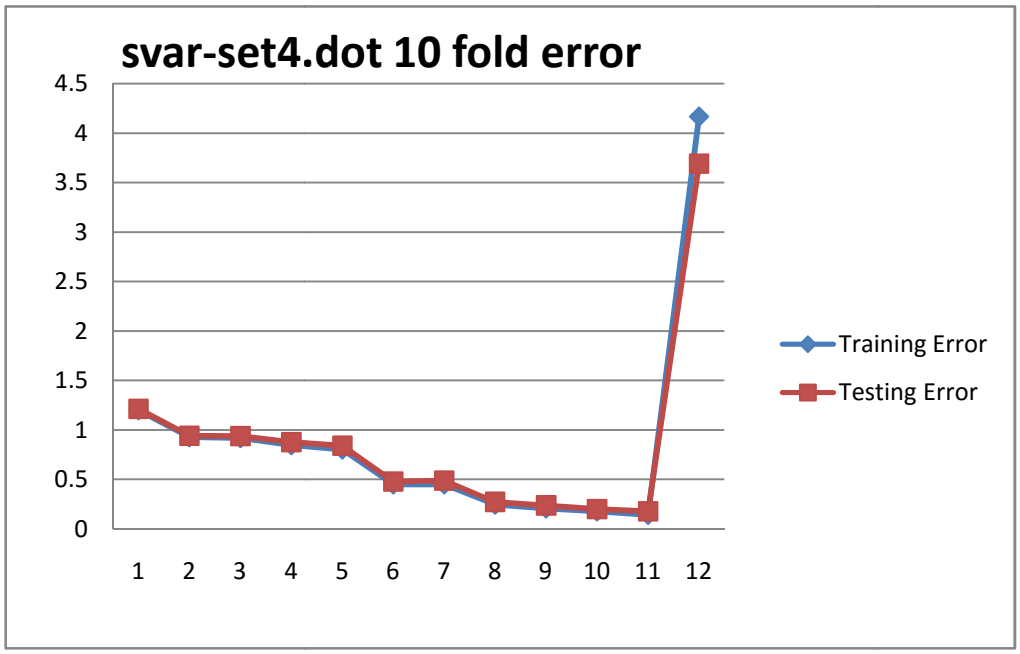
- Data set ( svar-set3.dot)



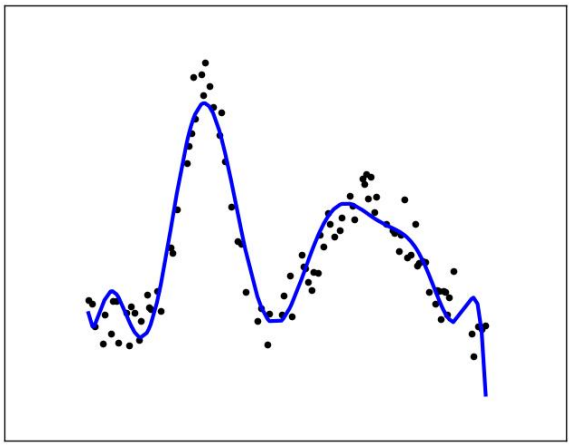

At this case I choose polynomial 11 as model. Although polynomial 12 has lower Training error, but with higher Testing error.



- Data set ( svar-set4.dot)




As svar-set4's data I choose polynomial 11, since polynomial over 12 the error is grow gigantically. And I can't find out why this happened.



4. Multivariate regression:

In this section, my code still had problem with combination the features. The code still could test 0 and 1 combination.

## 5. References


1. http://data.princeton.edu/wws509/notes/c2s4.html
2. CS584: Introduction to Python. G.Agam. January 14,2016
3. http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

4. http://myshare.dscloud.me/scipydoc/numpy_intro.html
5. http://docs.scipy.org/doc/numpy/reference/