Sing Variable Regression

1 Problem statement

Based on the given data, finding the best solution that is more close to the real data and spend less time.

2 Proposed solution

Solve this problem by using polynomial regression. We suppose the model as follows and try to get θ

$$h_0(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \bullet\bullet\bullet + \theta_n x^n$$

3 Implementation detail

1 Read file from “svar-test1.txt” and save x to Z

2 Read file from “svar-test2.txt” and save y to Y

3 We suppose the equation is as follows:

$$h_0(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \bullet\bullet\bullet + \theta_n x^n$$

4 Using follows equation to get θ

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

4 Results and discussion

4.1 The Fig shows results when X’s order is 3 and 10 fold cross validation

	Θ_0	Θ_1	Θ_2	Θ_3	Training Error (variance)	Testing Error (Variance)
Svar-set1.txt	-1.06573387e+01	3.82148170e+00	-9.63774686e-02	1.59486974e-03	4.17233278	4.11927854
Svar-set2.txt	0.99734198	-1.07305468	0.32673071	-0.02968398	0.02126954	0.01268473
Svar-set3.txt	-6.24677249e-01	6.59524725e-01	-6.23000783e-02	-3.39807518e-04	0.25508276	0.24752208
Svar-set4.txt	-0.34322644	0.89846299	-0.12285142	0.00362178	0.87024017	1.37029609

4.2 The Fig shows results when X’s order is 3 and using 80% data to train and use 10% percent to test

	Θ_0	Θ_1	Θ_2	Θ_3	Training Error (variance)	Testing Error (variance)
Svar-set1.txt	-8.24471074e+00	3.48734466e+00	-8.29692716e-02	1.43252015e-03	4.06363974	4.21223616
Svar-set2.txt	0.98537131	-1.07627282	0.33090212	-0.03027732	0.02105925	0.01375936
Svar-set3.txt	-6.12297396e-01	6.76622917e-01	-7.32364093e-02	6.42840170e-04	0.23820268	0.22784889
Svar-set4.txt	-0.3189511	0.86730288	-0.11849535	0.00345807	0.87354475	1.38331694

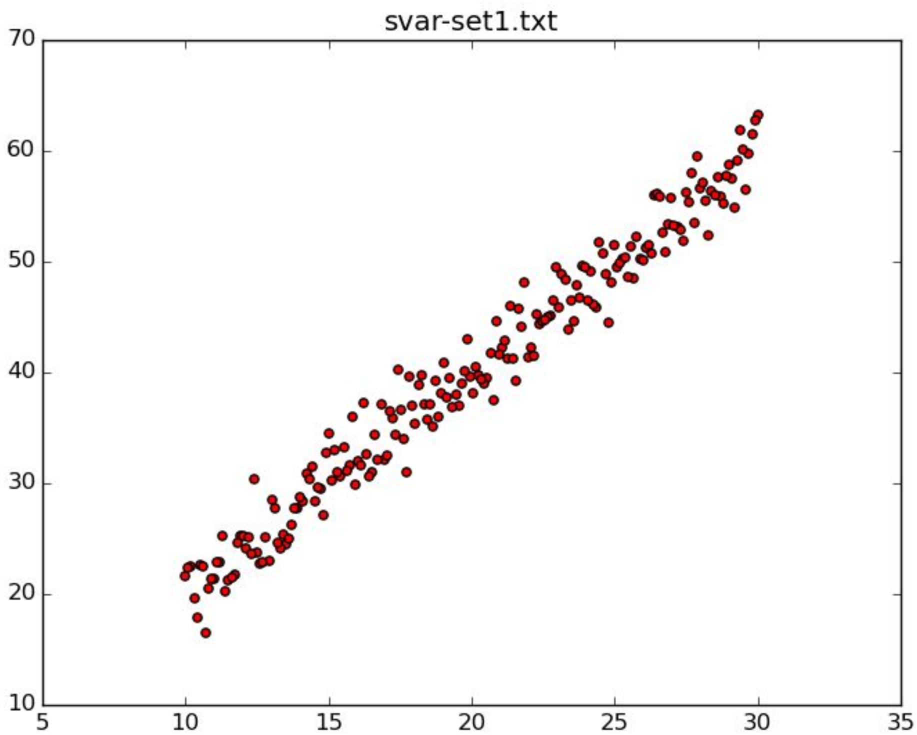
4.3 The Fig shows results when X’s order is 3 and using 70% data to train and use 10% percent to test

	Θ_0	Θ_1	Θ_2	Θ_3	Training Error (variance)	Testing Error (variance)
Svar-set1.txt	-9.39882724e+00	3.72853538e+00	-9.76857175e-02	1.71340575e-03	4.15598437	4.107713
Svar-set2.txt	0.9780343	-1.04832334	0.32124691	-0.02943264	0.02064599	0.01345135
Svar-set3.txt	-0.60799149	0.69481227	-0.08021437	0.0011849	0.24156065	0.22164338
Svar-set4.txt	-0.25529233	0.7914385	-0.10319713	0.00264736	0.88796329	1.40056508

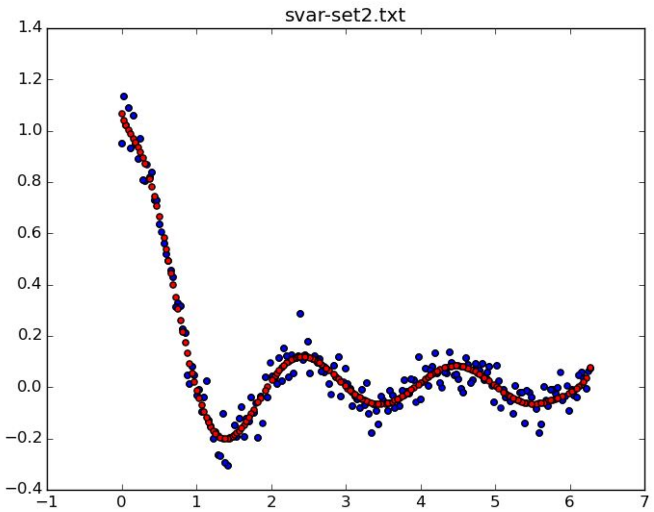
4.4 The Fig shows results that using 10 fold validation to test "svar-set2.txt"

	Training Error (variance)	Testing Error (variance)
Order is 10	0.00393432	0.00314877094907
Order is 12	0.00257015	0.00270113663934
Order is 13	0.00344792	0.00314918966149
Order is 15	0.00278876	0.00303002065937

4.5 Show the raw data that reads from "svar-set1.txt"



4.6 For "savr-set2.txt" using 10 fold validation,green color points is raw data, red color points are expected data



From “4.4” and “4.6” we can see, when order is 10,the solution is best. About the time ,it is almost the same with the other orders. So I think order 10 is the best.
From “4.1”,”4.2” and “4.3” we can see , if we have more training date, the error is smaller. So I think we use as much training data as possible.

Multivariate regression

2.1 Problem statement
Based on the given data, try to find the best multi-variables solution that is more close to the real data and spend less time. Using polynomial method, iterative method and Gaussian Kernel function to solve the problem.

2.2 Proposed solution
2.2.1 Using follow Polynomial model to create the function

$$h_0(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2$$

2.2.2 Iterative solution
For i =1,2,⋯,m

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \eta(\theta^{(i)^T} x^{(i)} - y^{(i)}) x^{(i)}$$

2.2.3 Create Gaussian Kernel function Model

$$k(x^i, x) = \exp(-\frac{1}{2\sigma^2}) \| x - x^i \|^2$$

3.3 Implementation Details

3.3.1 Polynomial Model

1. Read date from “mvar-set1.txt” and save x in variable z
2. Read date from “mvar-set1.txt” and save y in variable y

3. Create model as follows

$$h_0(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2$$

4. Depends on follows equation to get θ

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

3.3.2 Using Iterative solution

For i =1,2,⋯,m (m is the number of example)

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \eta(\theta^{(i)^T} x^{(i)} - y^{(i)}) x^{(i)}$$

3.3.3 Using Gaussian Kernel function

1 Read data from “mvar-set1.txt” and save x in variable x.
2 Read data from “mvar-set2.txt” and save y in variable y.
3 Find kernel function:

$$k(x^i, x) = \exp(-\frac{1}{2\sigma^2}) \| x - x^i \|^2$$

$$G = \begin{bmatrix} k(x^1, x^1) & \bullet & \bullet & K(x^1, x^m) \\ \bullet & \bullet & & \\ \bullet & & \bullet & \\ K(x^m, x^1) & \bullet & \bullet & k(x^m, x^m) \end{bmatrix}$$

$$\alpha = G^{-1} y$$

$$\hat{y} = \sum_{i=1}^{m} \alpha_i k(x^i, x)$$

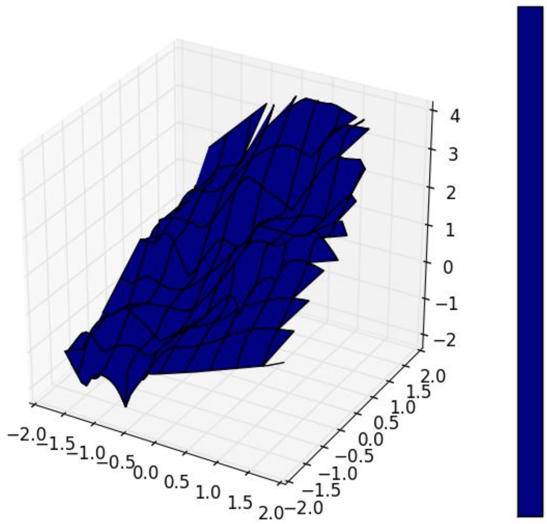
4 Results and discussion

4.1 The figures shows results using Polynomial Model with 10 fold validation

	Θ0	Θ1	Θ2	Θ3	Θ4	Θ5	Training Error (RSE)	Testing Error (RSE)
mvar-set1 .txt	1.020496 64	0.997872 41	0.984256 37	-0.00998 818	-0.01625 034	-0.00246 613	0.257446979 223	0.2665881545 5

mvar-set2 .txt	-6.31590 614e-05	6.459676 37e-02	-2.62858 352e-04	-1.09600 244e-03	-3.51815 636e-05	9.247761 81e-0	0.020141416 3004	0.0178305190 415
mvar-set3 .txt	-0.00227 05	-0.00110 772	-0.00034 975	-0.00050 837	0.000231 04	0.001273 52	1.674509622 92	1.6613004040 2
mvar-set4 .txt	0.007891 96	-0.00149 741	0.000422 64	0.000403 25	-0.00087 041	-0.00301 665	1.671976200 23	1.6547864164 4

Real data for mvar-set1.txt



For Gaussian Kernel method, we use more than 738 seconds for “mvar-set1.txt” and the training error is 1.08833414.

Because Gaussian Kernel method using so much time, I think polynomial method is better.

Reference

1 Elements of Statistical Learning (Second Edition)

2 http://www.stat.wisc.edu/~mchung/teaching/MIA/reading/diffusion.gaussian.kernel.pdf.pdf

3 http://www.numpy.org/

4 http://stackoverflow.com/questions/211160/python-inverse-of-a-matrix