Regression

Question 1:

Dataset : https://www.cs.toronto.edu/~delve/data/boston/

For Entire Dataset

RMSE	Value
For Training	4.396188144698281
For Testing	5.783509315084541

K Fold Cross Validation (K=5)

Degree = 1

Fold	RMSE Value
Fold 1 – Testing Fold 2,3,4,5 -- Training	5.112463170270926
Fold 2 – Testing Fold 1,3,4,5 – Training	5.90823583454406
Fold 3 – Testing Fold 1,2,4,5 – Training	6.198467467052379
Fold 4 – Testing Fold 1,2,3,5 -- Training	6.733586634589787
Fold 5 – Testing Fold 1,2,3,4 -- Training	6.343679668511275

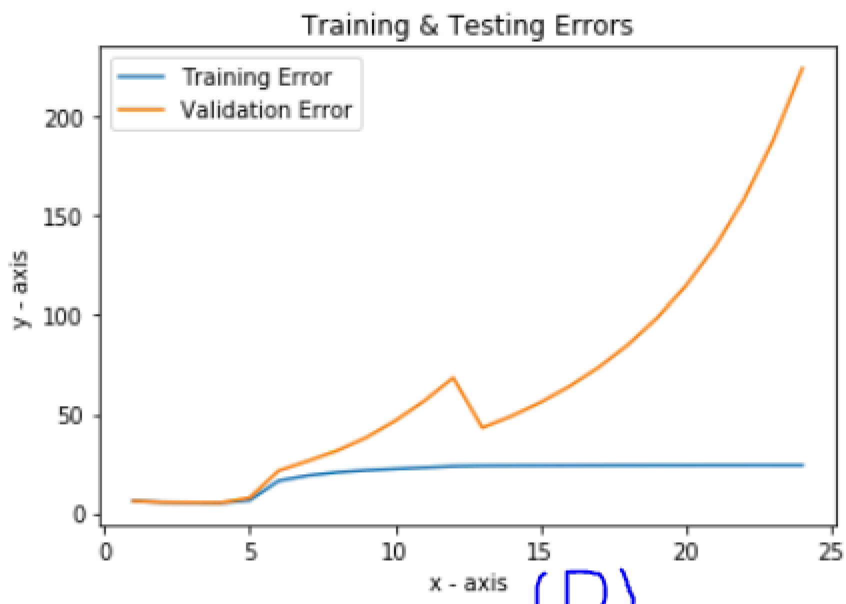
Mean RMSE : 6.059286554993685

Degree	Mean RMSE for Training & Validation
Degree 1	E_avg_k_cross_fold 6.059286554993685 E_train 6.0579657078521
Degree 2	E_avg_k_cross_fold 5.335866177597398 E_train 5.319824203678452
Degree 3	E_avg_k_cross_fold 5.213412805506195 E_train 5.183802178021808
Degree 4	**E_avg_k_cross_fold 5.148708568550794** **E_train 5.089239299774836**
Degree 5	E_avg_k_cross_fold 6.56518436870554 E_train 6.342202564593217
Degree 6	E_avg_k_cross_fold 17.124132556704993

	E_train 16.338598863745304 E_avg_k_cross_fold 20.352046521172387 E_train 19.06637951843012
Degree 8	E_avg_k_cross_fold 22.604212212043194 E_train 20.702467479099457
Degree 9	E_avg_k_cross_fold 24.633439580699896 E_train 21.73316775312481
Degree 10	E_avg_k_cross_fold 26.91394367009885 E_train 22.422743818149577

Minimum Training & Validation mean RMSE was reported at degree =4

Below is the graph for the following:



(D)

Analysis:

- From the above graph we can conclude that Training and validation error is minimum at Degree =4.

- Ideally Training error should decrease with increase in degree. Here it is slightly increasing and then staying constant .The reason can be that here we have used only one feature out of 13 features to make prediction therefore we are making significant amount of assumptions while prediction. This may introduce some constant amount of training error even if we increase the degree of the polynomial.

- Validation error showing the regular pattern. It is decreasing first till degree=4 then start increasing due to overfitting.

RMSE at Degree= 4

Training set: 4.058436146853399

Testing set: 5.084233753365987

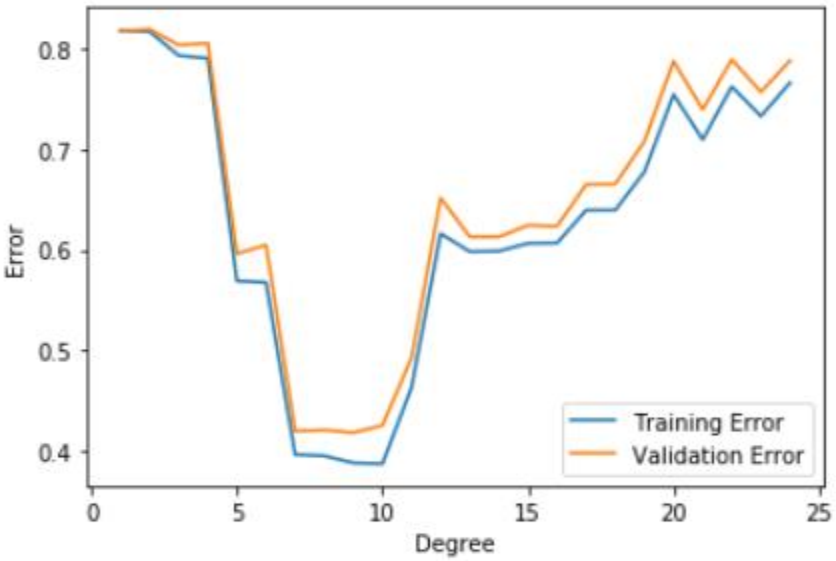
Question 2:

K Fold Cross Validation (K=5)

Degree = 1

Fold	RMSE Value
Fold 1 – Testing Fold 2,3,4,5 -- Training	0.887270815334462
Fold 2 – Testing Fold 1,3,4,5 – Training	0.8995450394804589
Fold 3 – Testing Fold 1,2,4,5 – Training	0.8264201866362071
Fold 4 – Testing Fold 1,2,3,5 -- Training	0.8729892053653323
Fold 5 – Testing Fold 1,2,3,4 -- Training	0.5992407298026305

Mean RMSE : 0.8170931953238183



Analysis:

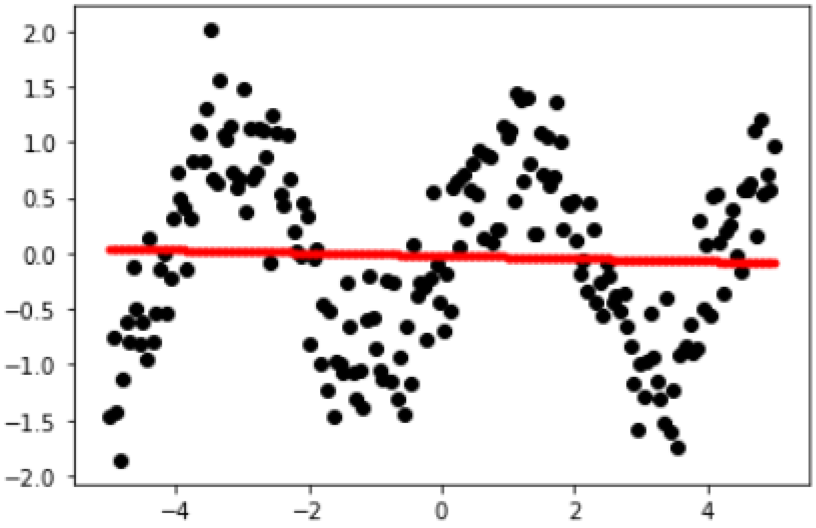
Validation Error is showing the usual pattern. This is minimum at Degree=10.After this it increases due to over fitting.

Training error is showing some unusual behavior. This decreases till degree=10 then increases. The reason can be understood by the following regression fit plots. For higher degrees the polynomial is not able to fit for the data well which in turn introducing training error.

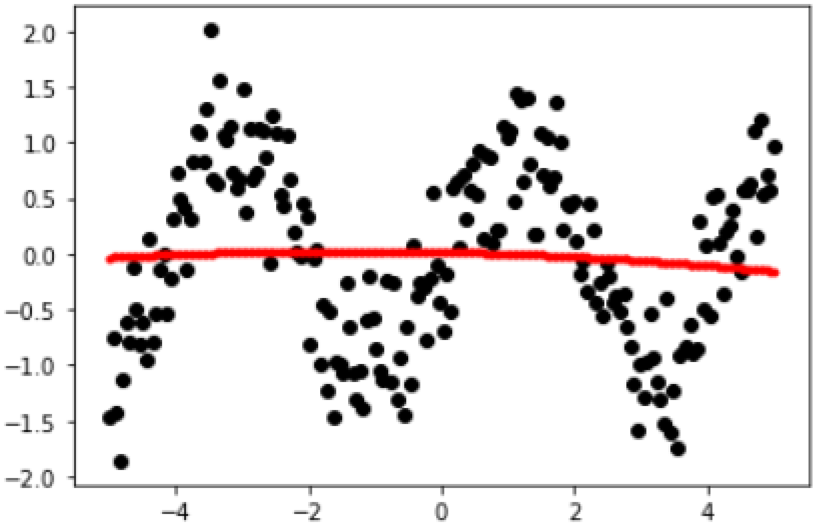
Degree	Mean RMSE for Training & Validation
Degree 1	E_avg_k_cross_fold 0.8170931953238183 E_train 0.8176720967625666
Degree 2	E_avg_k_cross_fold 0.8190244501114975 E_train 0.8166985940822238
Degree 3	E_avg_k_cross_fold 0.8032839139880288 E_train 0.7928840667413464
Degree 4	E_avg_k_cross_fold 0.8049923750221316 E_train 0.7901380534460938
Degree 5	E_avg_k_cross_fold 0.5961684256861994 E_train 0.5691097656608093

Degree 6	E_avg_k_cross_fold 0.6046935812417988 E_train 0.5676549759786366 E_avg_k_cross_fold 0.420069074264193 E_train 0.39684245673237434
Degree 8	E_avg_k_cross_fold 0.4211547394415929 E_train 0.3954621625855902
Degree 9	E_avg_k_cross_fold 0.41845161925982144 E_train 0.388414423705013
Degree 10	E_avg_k_cross_fold 0.42569668174484915 E_train 0.3875638521364226

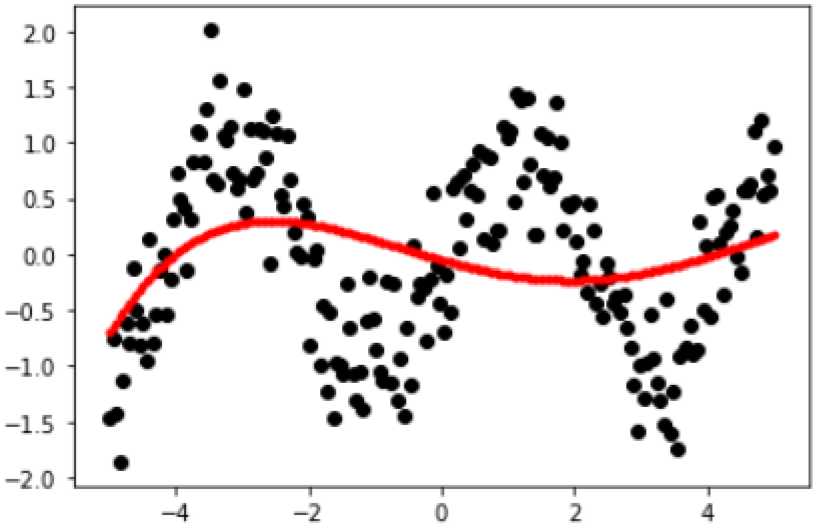
Degree : 1



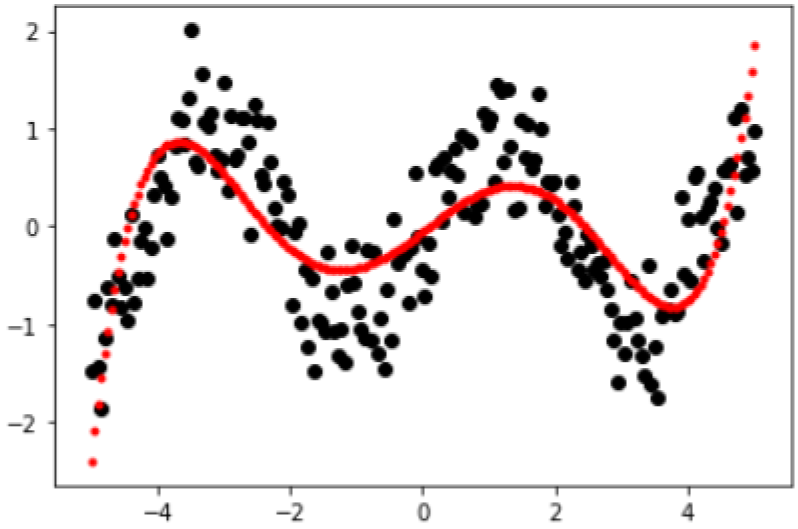
Degree 2



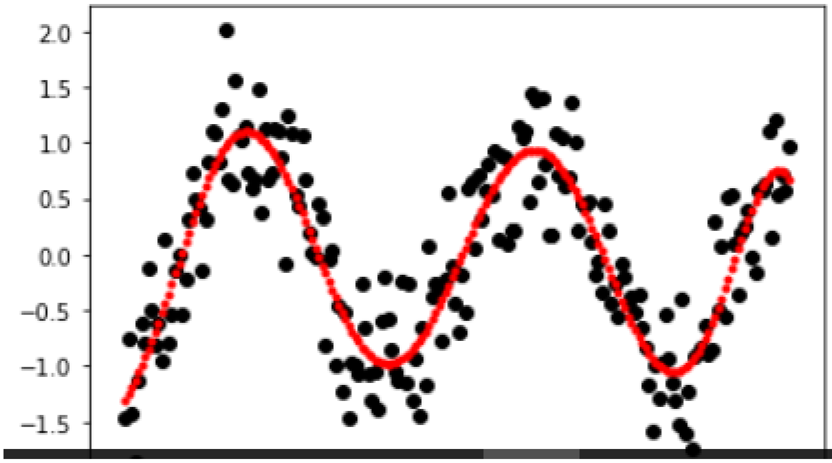
Degree 4



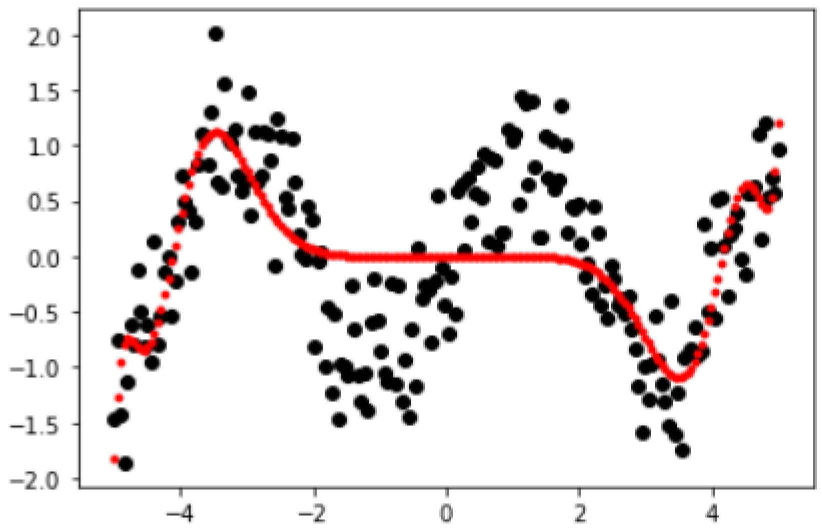
Degree 5



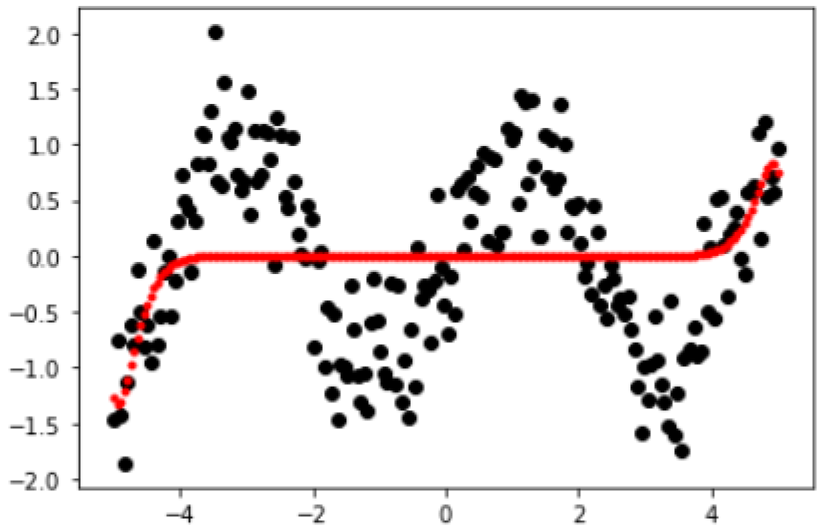
Degree 10



Degree 15



Degree 30 :


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−4
−2
0
2
4