Assignment 1. Part 1 – Trade-off Between Overfitting and Underfitting

In this assignment part 1, several regression models have been compared to illustrate the trade-off between overfitting and underfitting. There is only one feature and the target *t* satisfies the following relation:

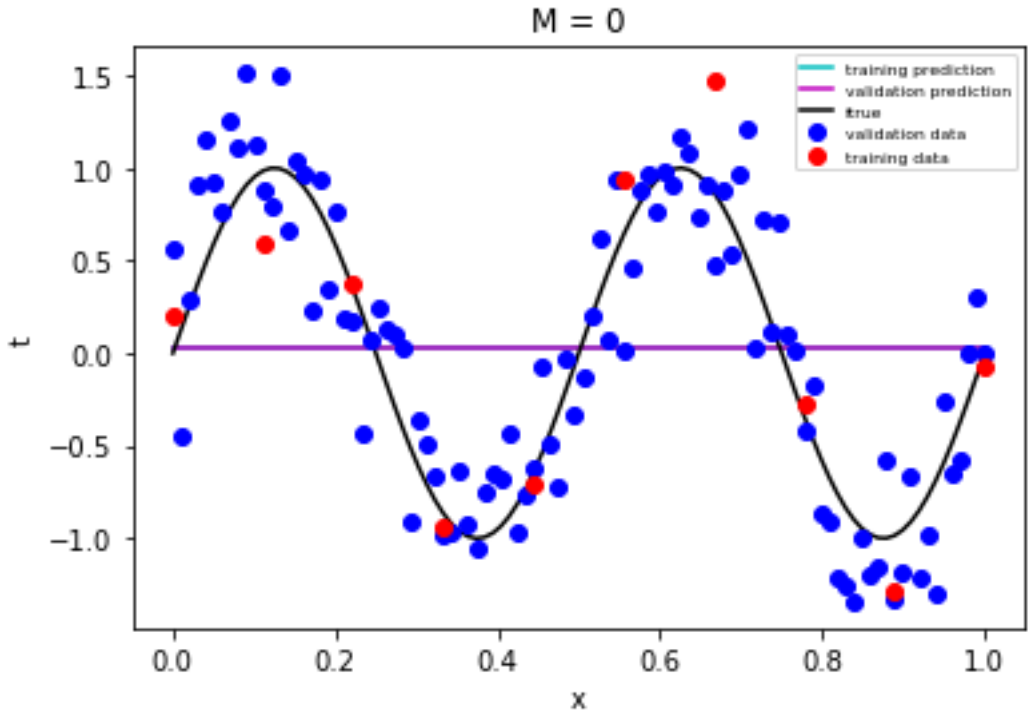
$$t = \sin(4\pi x) + \in$$

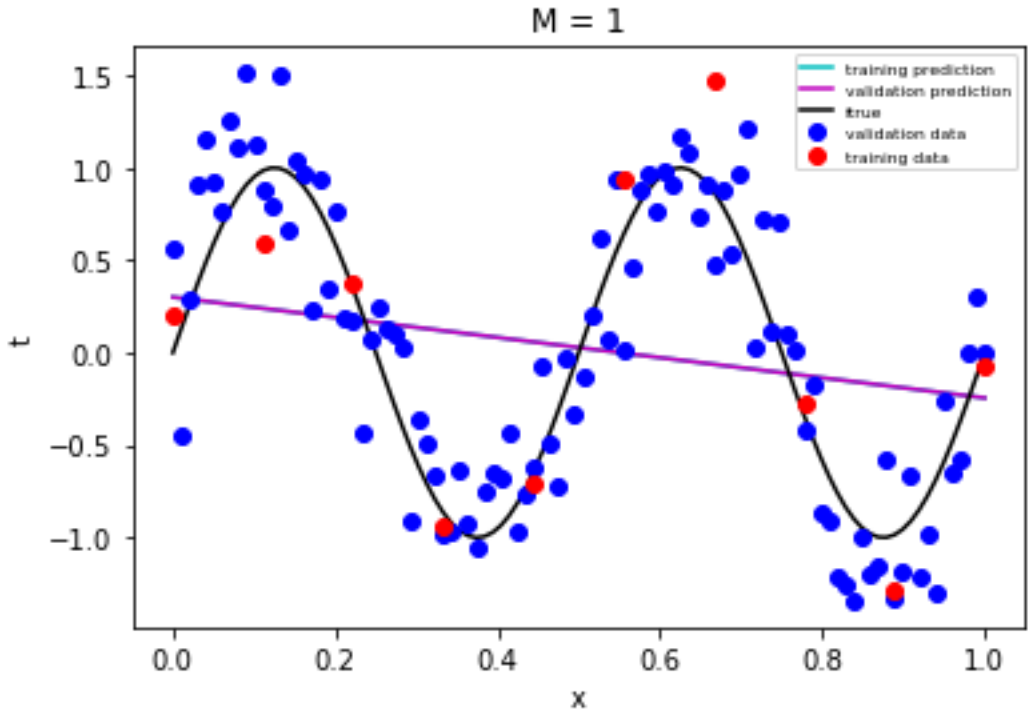
Where $\in$ is random nose with a Gaussian distribution with 0 mean and variance 0.09. There are 10 examples generated in the training set and 100 examples generated in the validation set. The number used to generate random data is 5007.

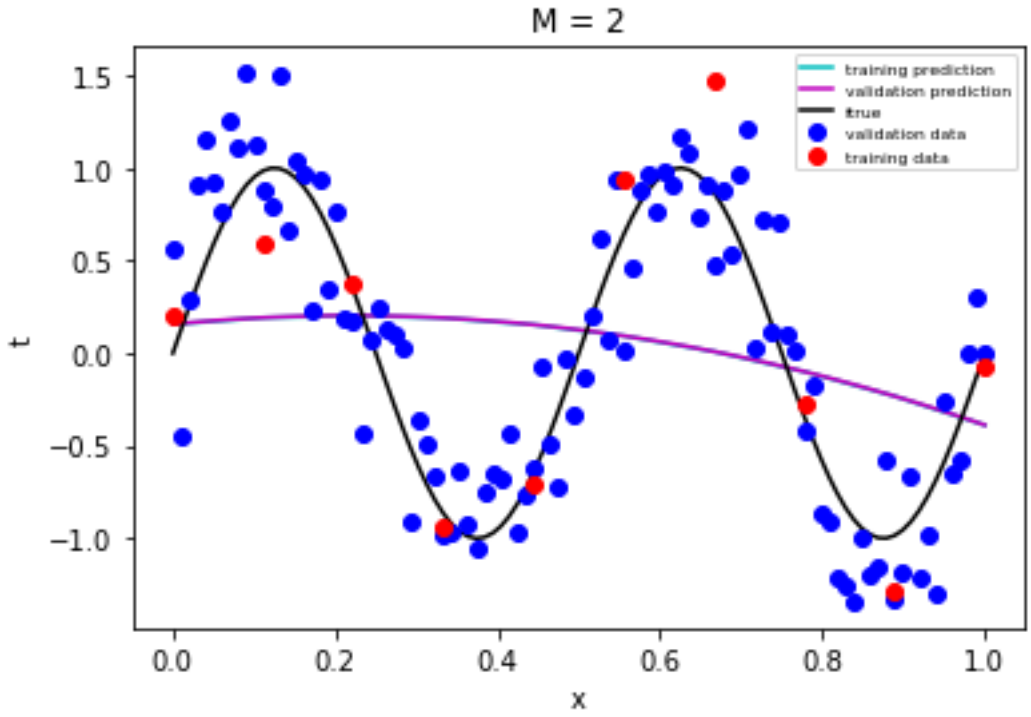
Ten regression models of increasing capacity (corresponding to M from 0 to 9) have been generated using least squares. Below are the training and validation errors recorded for each M.

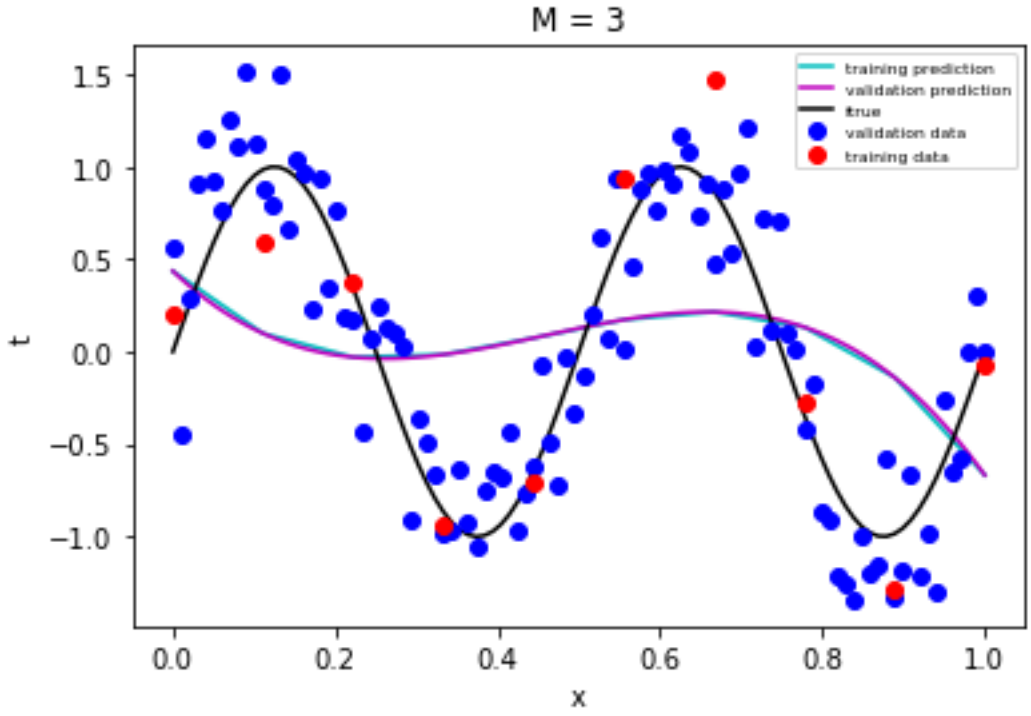
M	Training Errors	Validation Errors
0	0.6664948962165589	0.6253151390962123
1	0.6362870371614407	0.5438671886654354
2	0.6287315411072403	0.5525442006701698
3	0.5908571802702978	0.5251505915437131
4	0.5747169668116779	0.5394303894455582
5	0.10781130566248671	0.18571801038481478
6	0.0916335051015977	0.211276557842437
7	0.0007259881299826567	0.23352406423117436
8	1.3367573520634344e-05	0.21472225305221238
9	2.3565852162940628e-08	0.224773245919178

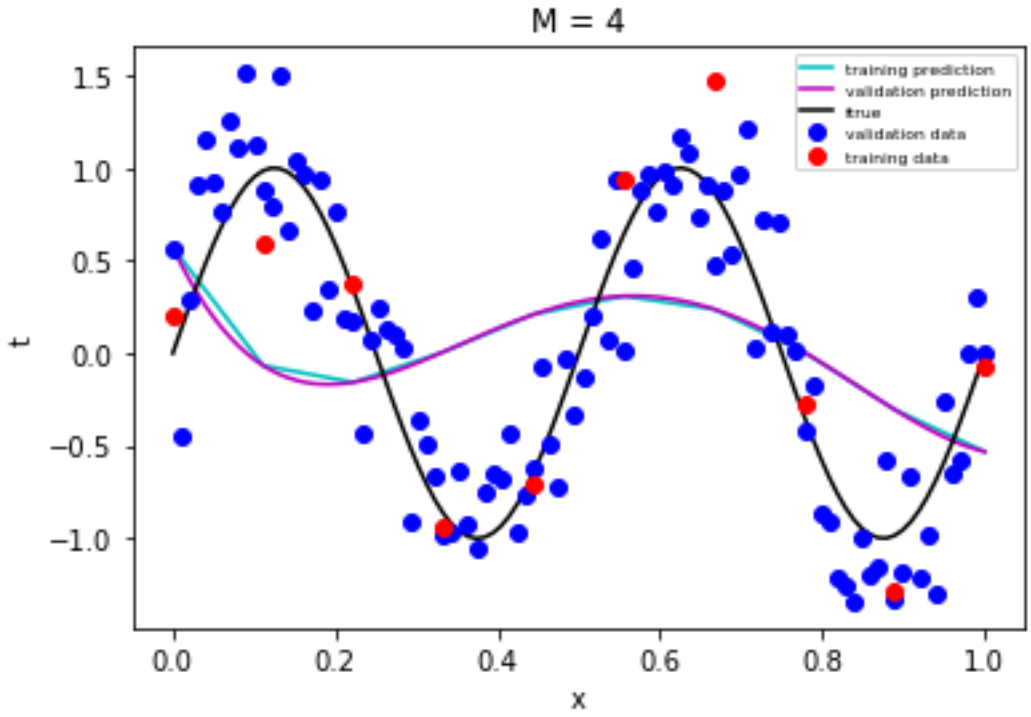
The following figures below show the plot of the prediction $f_M(x)$ versus x, all the points in the training and validation sets, as well as the curve $f_{true}(x)$

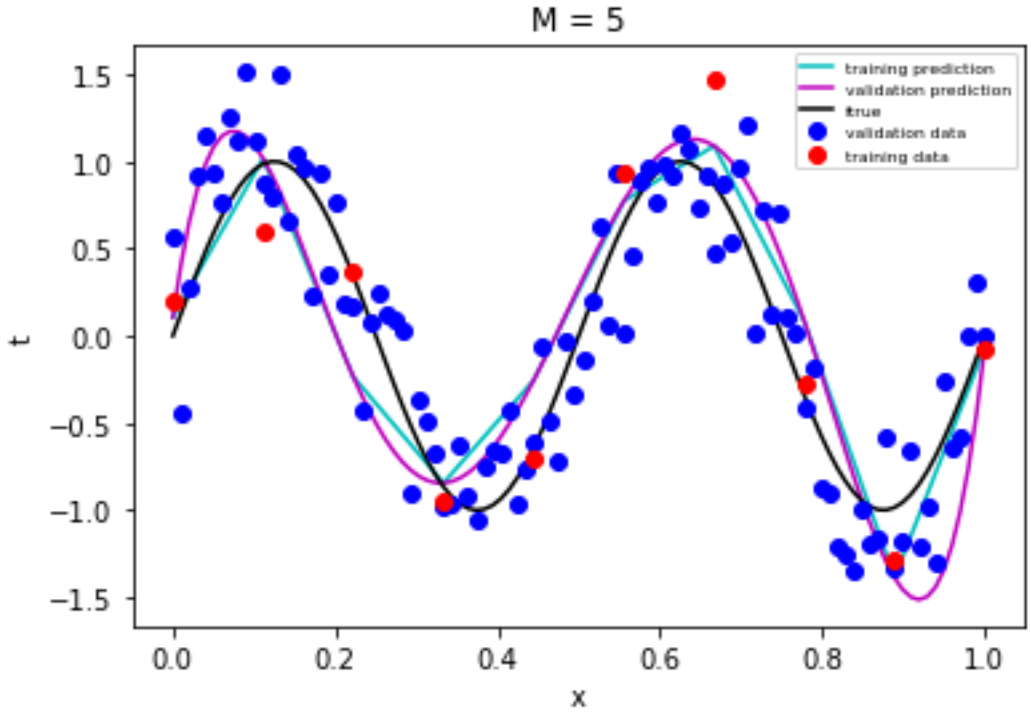

M = 0
training prediction
validation prediction
true
validation data
training data
1.5
1.0
0.5
0.0
−0.5
−1.0
t
0.0
0.2
0.4
0.6
0.8
1.0
x

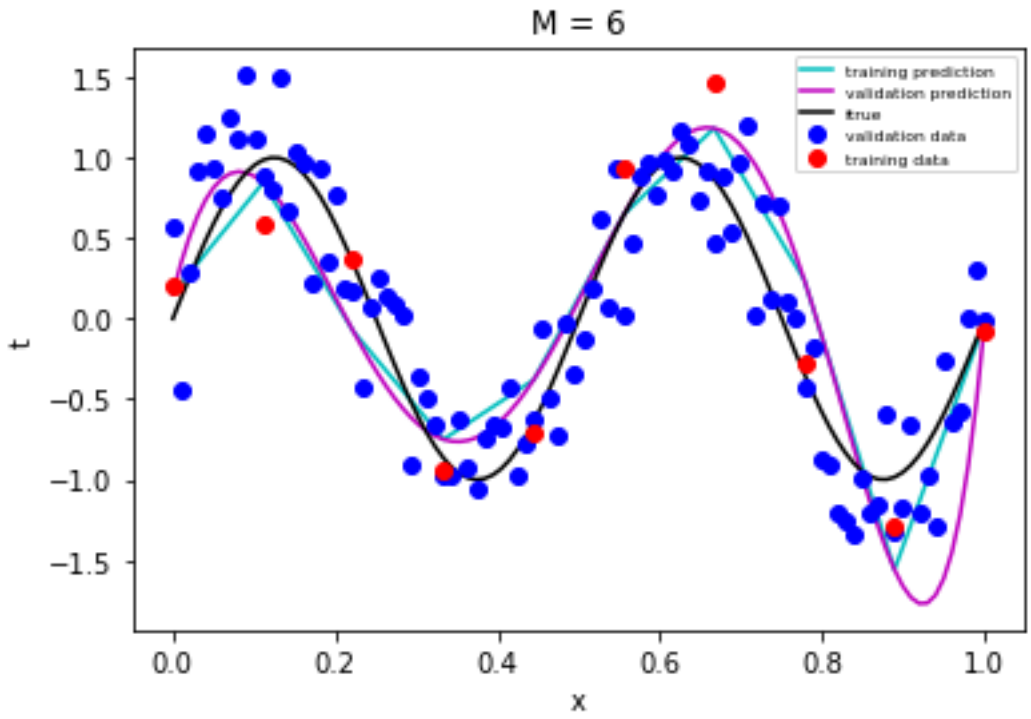

M = 1
training prediction
validation prediction
true
validation data
training data
1.5
1.0
0.5
0.0
−0.5
−1.0
t
0.0
0.2
0.4
0.6
0.8
1.0
x

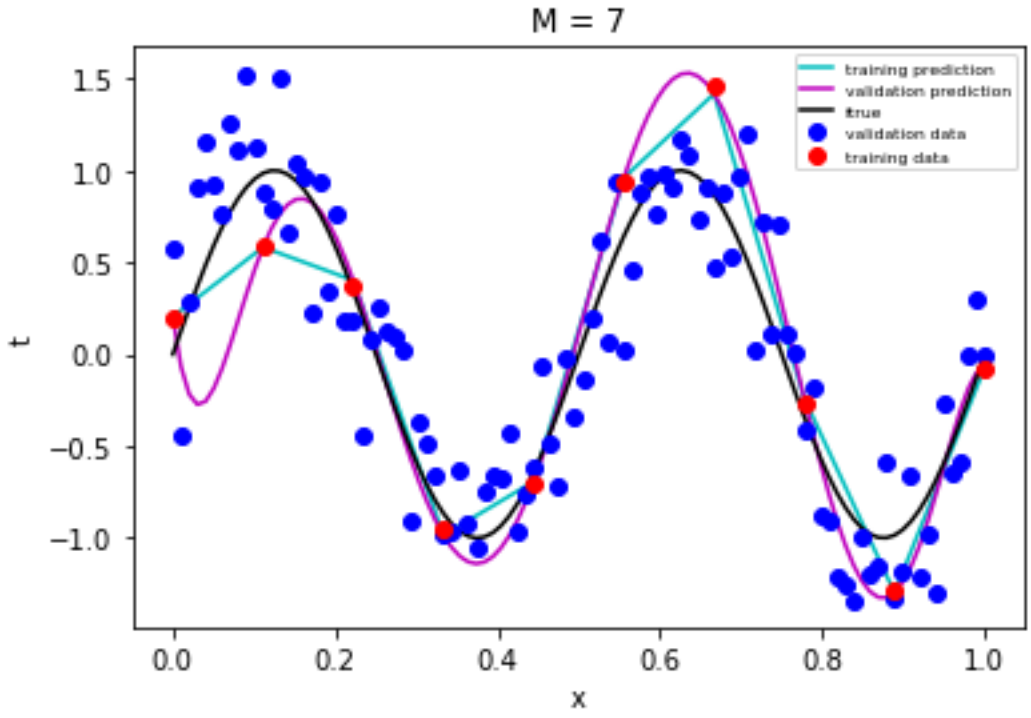

M = 2
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

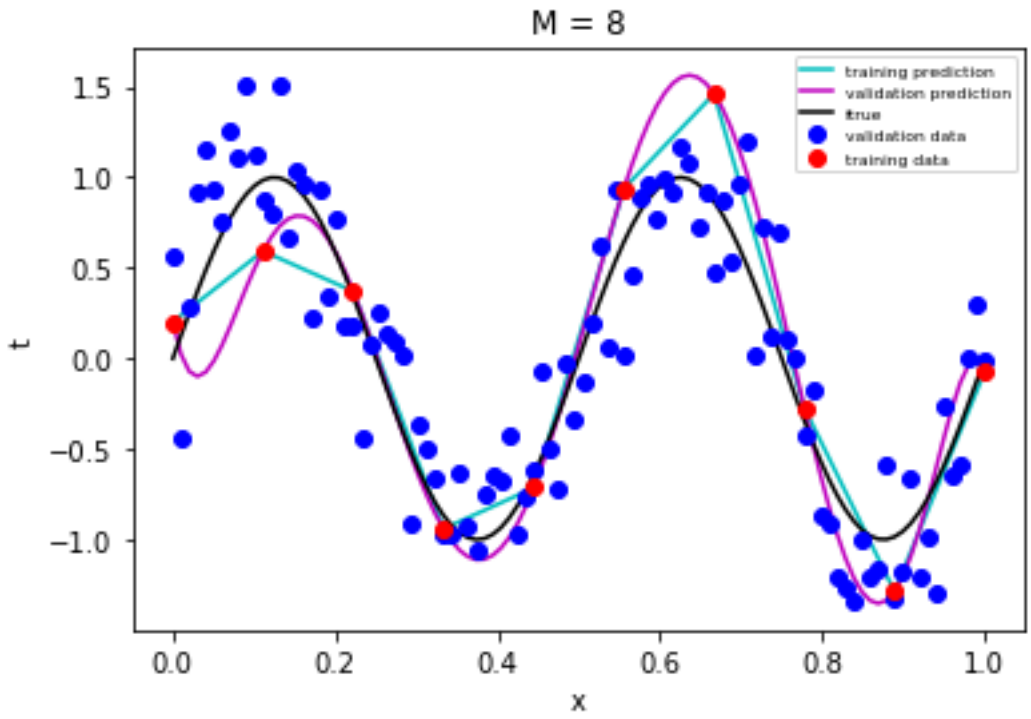

M = 3
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

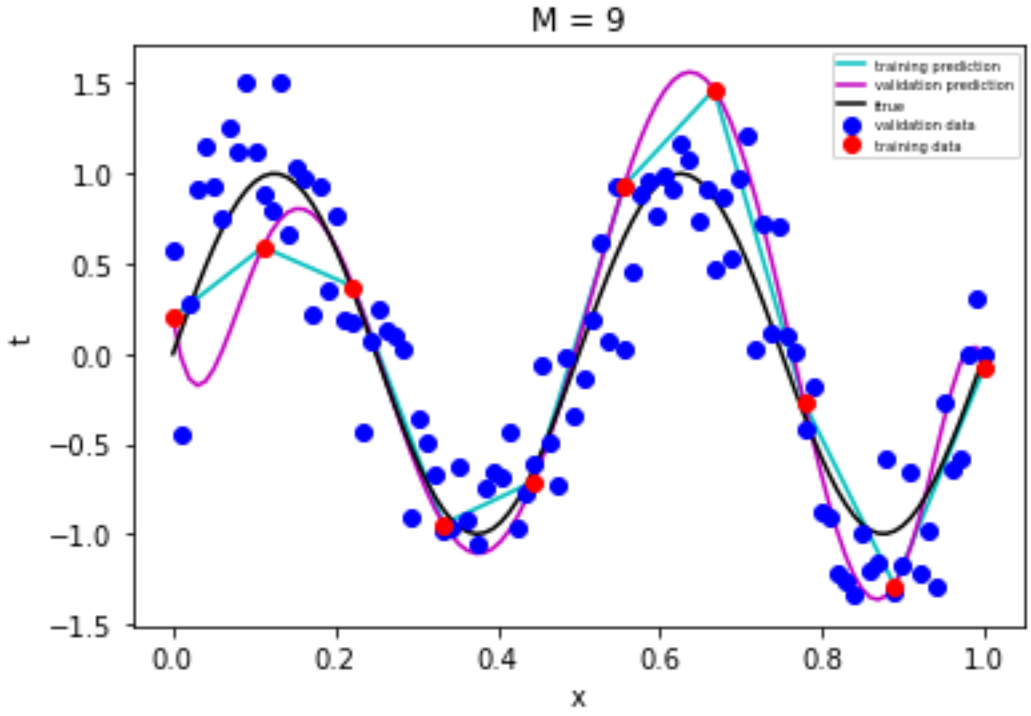

M = 4
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0


M = 5
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

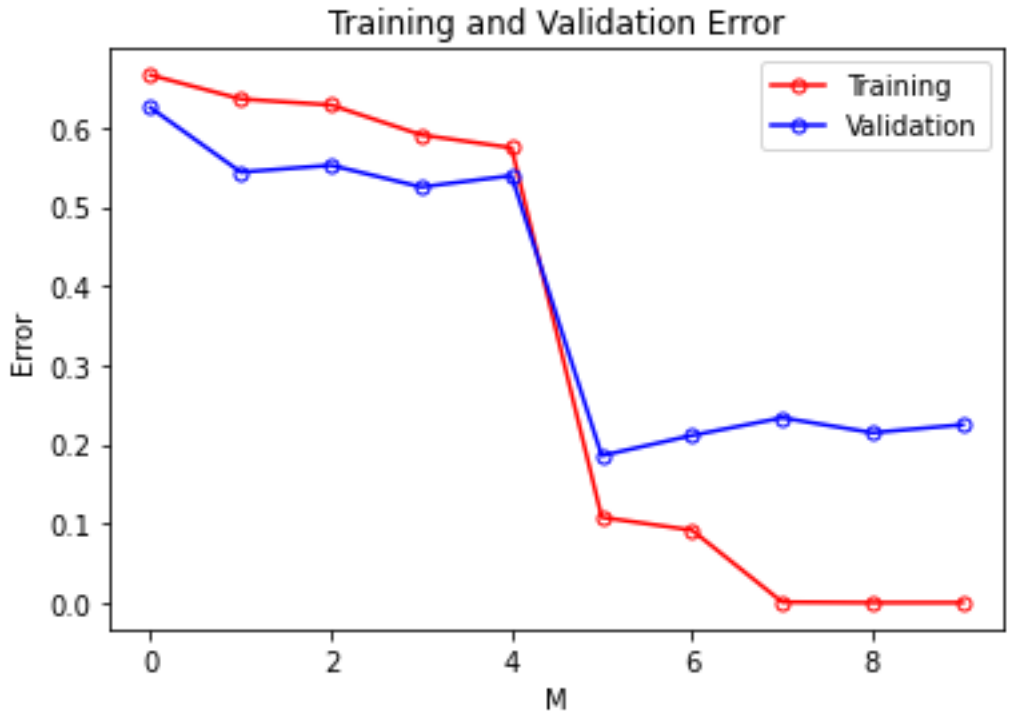

M = 6
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0


M = 7
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0


M = 8
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0


M = 9
training prediction
validation prediction
true
validation data
training data
t
x
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

The following figures show the comparison of the training and validation errors. From the plot, we can see that for M = 0 to 4, it is underfit as the error is high. The best fit is at M = 5. Then as M increases, training error becomes even less as expected because the polynomial contains high degrees of freedom to tune exactly to the points in the training set. However, the validation error increases, and we can see it in the plot as it exhibits wild oscillations. This shows overfitting.



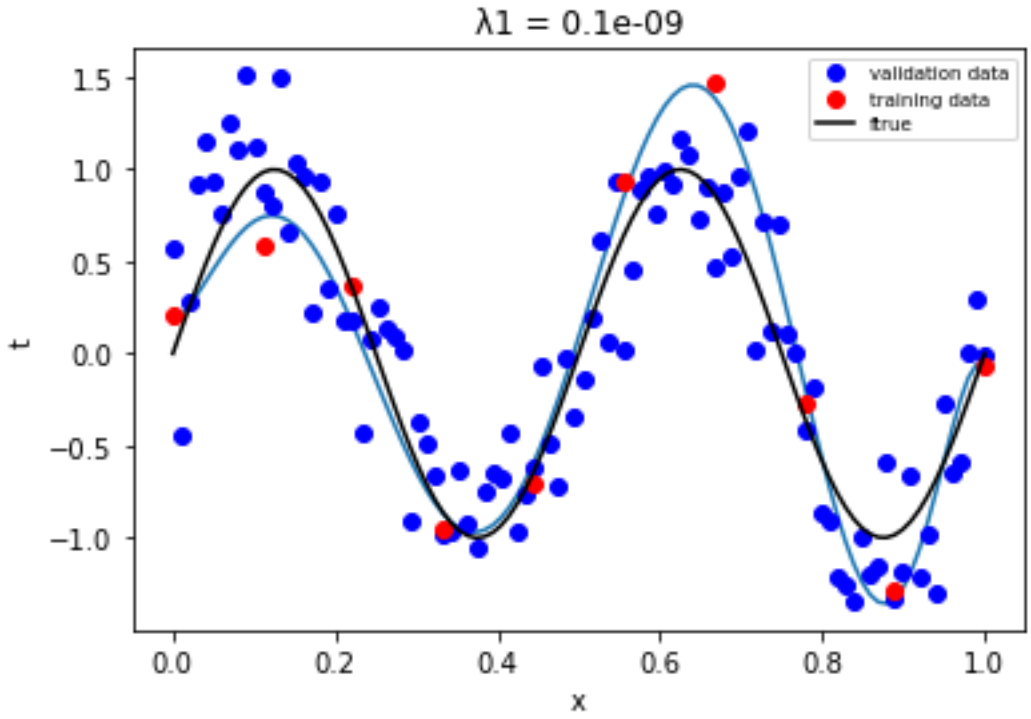
Due to the overfitting at M=9, this model has been trained with regularization in order to control the overfitting. The method used is **ridge regression.** Several values of λ have been tested to find the number that would eliminate overfitting and a value that shows underfitting occurring.

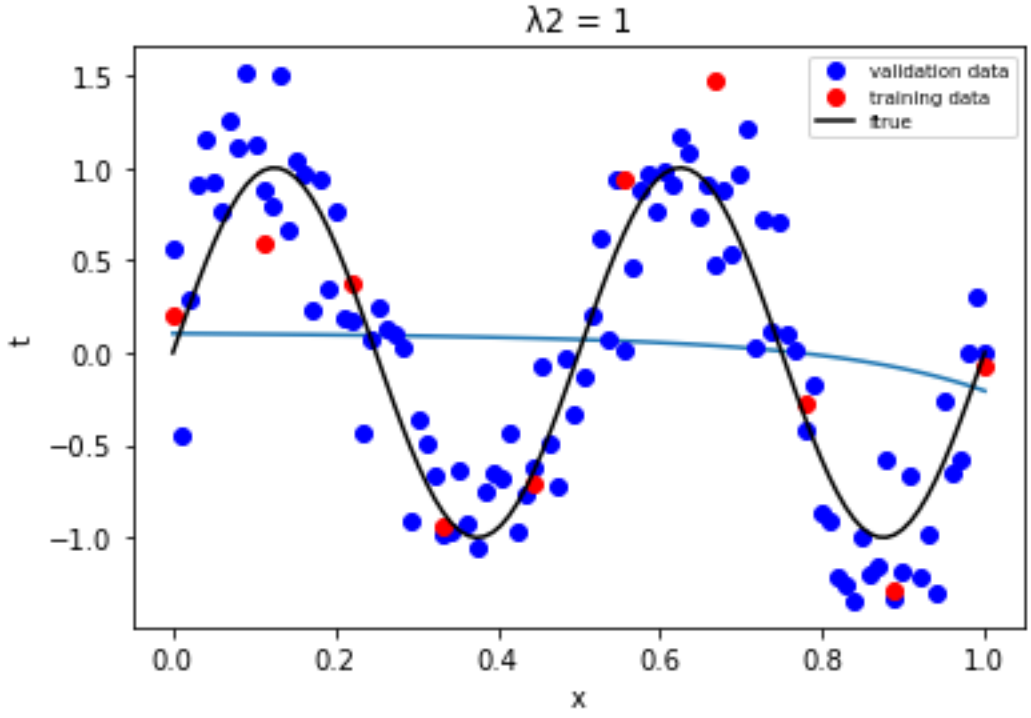
Without regularization, at M=9, the validation error is 0.224773245919178.

The chosen λ_1 that eliminates overfitting is 0.1e-09. With this λ, the validation error becomes 0.1502171583888331. This decreased the original validation error.

The chosen λ_2 that shows underfitting occurring is 1. With this λ, the validation error becomes 0.5818145445465288. This increased the validation error.

The following figures shows λ_1 and λ_2, respectively with the prediction $f_M(x)$ versus x, all the points in the training and validation sets, as well as the curve $f_{true}(x)$





As a result, the 12 plots with an additional plot of the errors illustrated in part 1 of this report has given a good comparison of the trade-off between overfitting and underfitting. The addition of the regularization also illustrates clearly how overfitting can be controlled.

Assignment 1. Part 2 – Linear Regression, Basis Expansion and Feature Selection

In part 2 of this assignment, linear regression, basis expansion, and feature selection have been experimented using the Boston housing data set.

The approach used for feature selection is the greedy algorithm where it starts from an empty set and grows gradually by selecting at each step the feature that increases the performance of the predictor the most. In this assignment, least squares regression is used and K-fold cross-validation error is computed to measure performance.

There are 13 features in the Boston housing data set. To see the result of each possible size *k* of the selected subset of features, the greedy algorithm is applied until subset *S* reaches the maximum size (i.e. 13). Below shows the recording of all the cross-validation errors and the test errors at each *k,* for each feature and the parameter vector for the *k*-feature model chosen.

For k = 1

error: 71.64526322246665
error: 49.69413098767019
error: 122.32195661112019
error: 62.63796954102938
error: 66.92339341414704
average error of f1 is74.64454275528668

error: 67.96800678494195
error: 53.47209804394972
error: 125.3464361741119
error: 60.5716767677559
error: 64.87023310616712
average error of f2 is74.44569017538531

error: 57.52132083872495
error: 45.8199927993924
error: 110.74329865871655
error: 54.262261784105405
error: 69.03297979379768
average error of f3 is67.4759707749474

error: 80.91200737362847
error: 53.17538913502616
error: 131.08648595935634
error: 79.61061093723966
error: 72.83259989467456
average error of f4 is83.52341865998503

error: 67.92529700358473
error: 44.069091160428805
error: 116.78788206859176
error: 59.650952791345105
error: 67.45596756523295
average error of f5 is71.17783811783667

error: 37.06170155336991
error: 34.43283874316904
error: 51.216579450952054
error: 30.68658932675317
error: 70.98937156433632
average error of f6 is44.8774161277161

error: 69.03569704045275
error: 44.64651440588948
error: 123.99163421312377
error: 62.22918488172283
error: 72.79707971807025
average error of f7 is74.54002205185182

error: 77.28853821937184
error: 49.54916835693122
error: 135.58337327092025
error: 72.32600941164505
error: 69.96088302799554
average error of f8 is80.94159445737277

error: 73.27638272296787
error: 41.84129725798218
error: 129.76237586490848
error: 61.66298383894511
error: 73.60418684810554
average error of f9 is76.02944530658183

error: 63.53991194113722
error: 37.058234194447806
error: 116.54979815063935
error: 56.331774880206964
error: 72.30597506683846
average error of f10 is69.15713884665396

error: 67.65950300966078
error: 40.51249115912848
error: 110.21585211548572
error: 48.73515914266877

error: 64.71843795489956
average error of f11 is66.36828867636866

error: 79.09518595933429
error: 47.89998961761599
error: 123.69749987046032
error: 68.20792788413777
error: 66.82962911101065
average error of f12 is77.14604648851181

error: 34.172058201512456
error: 22.922897365579264
error: 66.48680617326706
error: 36.251435022859624
error: 49.45570816946632
average error of f13 is41.857780986536945

Smallest cross-validation error is 41.857780986536945
The test error is 30.895552903066726
S1: [f13]
W parameters: [34.47966394 -0.93461692]

For k = 2

error: 33.697633007848644
error: 22.42645175828073
error: 67.62282981219842
error: 35.95940069478116
error: 49.25747948551157
average error of f1 is41.7927589517241

error: 33.02271483018797
error: 24.020533704663038
error: 67.6667638981232
error: 34.87885743826627
error: 48.87554785336043
average error of f2 is41.69288354492018

error: 33.660396465489434
error: 22.95349361595661
error: 67.14184720188642
error: 35.74732330871607
error: 49.86240599313301
average error of f3 is41.87309331703631

error: 31.511452705494293

error: 19.663928812345244
error: 61.358191345478325
error: 37.25628571655466
error: 52.14117202497183
average error of f4 is40.38620612096887

error: 34.429399510798156
error: 22.971217615762857
error: 66.79982820766696
error: 36.44921881317405
error: 49.66379446070291
average error of f5 is42.06269172162099

error: 23.759100555228482
error: 19.971166863109538
error: 45.24073171364638
error: 25.090462855450827
error: 55.81667462299963
average error of f6 is33.97562732208697

error: 33.77001453481012
error: 22.47423855498389
error: 65.33780296934431
error: 37.093173479896464
error: 49.2904275000955
average error of f7 is41.59313140782605

error: 33.889492539266435
error: 22.605715048197915
error: 62.46556258927312
error: 35.73832837156965
error: 48.46035617411989
average error of f8 is40.6318909444854

error: 34.2615764236459
error: 22.92151958325968
error: 67.58236916857804
error: 36.31327045006192
error: 49.46598090702862
average error of f9 is42.10894330651483

error: 33.37377405671619
error: 22.294556896237193
error: 67.36914326416571
error: 35.47460665943188
error: 49.51001381548636

average error of f10 is41.60441893840747

error: 30.38988568014693
error: 20.40073982689924
error: 60.418799076620594
error: 28.00570173141643
error: 47.24240786871521
average error of f11 is37.29150683675968

error: 33.39861039392363
error: 23.181220602161744
error: 66.61565697509579
error: 35.25171356813321
error: 49.96604555297979
average error of f12 is41.68264941845884

Smallest cross-validation error is 33.97562732208697
The test error is 24.823609311671348
S2: [13, 6]
W parameters: [-3.24401394 -0.60111645 5.33716355]

For k = 3

error: 23.118211860355178
error: 19.33535098881211
error: 46.31558323635069
error: 24.404418880286514
error: 55.51895484683706
average error of f1 is33.738503962528306

error: 23.25342104138343
error: 20.814075938099865
error: 46.351174359729306
error: 24.52991204218474
error: 55.436875676028784
average error of f2 is34.07709181148523

error: 23.370208631924168
error: 19.804567956688846
error: 45.86897244151222
error: 24.66581060818448
error: 56.054699716527
average error of f3 is33.95285187096734

error: 21.595761937073764
error: 18.45353313928068

error: 41.665608873417284
error: 26.09468198783808
error: 59.904894421012635
average error of f4 is33.542896071724485

error: 23.522958954772133
error: 20.503991715827834
error: 45.74316878802819
error: 24.86293433511931
error: 55.544867432761706
average error of f5 is34.035584245301834

error: 23.79665204691526
error: 20.022840028617548
error: 45.3097988702381
error: 25.696311357910826
error: 55.82965772391235
average error of f7 is34.13105200551881

error: 24.40505547457751
error: 18.982482984614542
error: 43.50683972317386
error: 25.575234735038713
error: 54.901787454511954
average error of f8 is33.474280074383316

error: 23.157177171408147
error: 19.87112921783562
error: 46.45443198874936
error: 24.473866782940583
error: 56.047458156264135
average error of f9 is34.00081266343957

error: 22.425321200035626
error: 19.020227718421147
error: 45.79370624535858
error: 23.650362090913198
error: 55.85532621668352
average error of f10 is33.34898869428242

error: 21.402439743645527
error: 17.325255937669755
error: 42.94055415153059
error: 20.09777159687035
error: 52.13606980742069
average error of f11 is30.780418247427384

error: 22.385509262893766
error: 21.4866165383888
error: 43.91663070110405
error: 22.392332690514028
error: 55.54835227349986
average error of f12 is33.145888293280095

Smallest cross-validation error is 30.780418247427384
The test error is 20.641612168024857
S3: [13, 6, 11]
W parameters: [15.50137937 -0.53230991 4.84073611 -0.89637111]

For k = 4

error: 21.206384048724626
error: 17.09312630052426
error: 43.93524404801454
error: 19.906208443500162
error: 52.12125559382879
average error of f1 is30.85244368691848

error: 21.629960225636452
error: 17.50034816190107
error: 43.07907764259571
error: 20.427835792619664
error: 52.13170980649461
average error of f2 is30.9537863258495

error: 21.68431268002697
error: 17.328076889172518
error: 43.16534832545165
error: 20.279857162637217
error: 52.42337331189528
average error of f3 is30.976193673836725

error: 19.351031618947527
error: 16.65239633623222
error: 39.95473265809943
error: 21.22193183414108
error: 56.24275292288576
average error of f4 is30.684569074061205

error: 21.045182321946687
error: 17.758357206453926
error: 43.18089258067237

error: 19.74798816796596
error: 51.957449049386085
average error of f5 is30.737973865285007

error: 21.545697747590953
error: 17.218755164871066
error: 42.889382924884245
error: 20.913747141158506
error: 52.13013586258434
average error of f7 is30.939543768217824

error: 21.902100790512804
error: 16.414116946381455
error: 41.063026596045766
error: 20.406783014861208
error: 50.492843181332645
average error of f8 is30.055774105826778

error: 22.234800223703086
error: 17.318335168420496
error: 43.024974224019864
error: 20.425570596591083
error: 51.982493385485824
average error of f9 is30.997234719644077

error: 21.30612752986868
error: 17.194267567093757
error: 43.56691023812061
error: 19.967707035726637
error: 52.56767063977973
average error of f10 is30.92053660211788

error: 20.392429042090992
error: 18.812900510913884
error: 41.572282865041224
error: 18.081830499231803
error: 52.2287158724897
average error of f12 is30.21763175795352

Smallest cross-validation error is 30.055774105826778
The test error is 19.063722623348596
S4: [13, 6, 11, 8]
W parameters: [20.82530935 -0.6223631 4.56704814 -0.92508598 -0.51293971]

For k = 5

error: 21.466053826001758
error: 15.852843393098128
error: 41.992814554778825
error: 20.039032078483768
error: 50.270953290027535
average error of f1 is29.924339428478003

error: 21.604609682566853
error: 17.31004961451134
error: 40.74773627484908
error: 19.505650168293894
error: 49.723314728659304
average error of f2 is29.778272093776092

error: 20.832512669516724
error: 15.782640533403788
error: 40.90301654832177
error: 19.62229520819223
error: 50.539142548045525
average error of f3 is29.535921501496006

error: 20.039221861142412
error: 15.903605027390821
error: 38.71068550125249
error: 21.50036220637249
error: 54.76825624656281
average error of f4 is30.184426168544206

error: 19.22849846537394
error: 16.797686594413463
error: 39.2089937185506
error: 18.60976246017567
error: 47.110141168494316
average error of f5 is28.191016481401597

error: 21.632228530148968
error: 17.22600405140869
error: 40.29534661177782
error: 19.69730011217068
error: 50.61310337705587
average error of f7 is29.892796536512407

error: 22.10575383275175
error: 16.414116526290407
error: 41.591290225456454
error: 20.432666281361026

error: 50.607924169727156
average error of f9 is30.230350207117358

error: 21.096920191019798
error: 15.678950861740763
error: 41.52465671869203
error: 19.906841166630482
error: 50.66074716357636
average error of f10 is29.77362322033189

error: 20.52070283914705
error: 17.887533214729373
error: 39.302953549410056
error: 18.034017389726895
error: 50.45683991336939
average error of f12 is29.24040938127655

Smallest cross-validation error is 28.191016481401597
The test error is 18.696732941282427
S5: [13, 6, 11, 8, 5]
W parameters: [33.68973358 -0.5256199 4.61544159 -0.9953466 -1.18505501 -18.96816344]

For k = 6

error: 19.03119113837244
error: 16.54393527822353
error: 39.98582720680207
error: 18.43052123038217
error: 47.105639623221485
average error of f1 is28.21942289540034

error: 19.092032232146256
error: 17.681497902603727
error: 38.78684322932258
error: 17.70938402191107
error: 46.40969672321497
average error of f2 is27.935890821839724

error: 19.190616921542635
error: 16.77925074781808
error: 39.33569720878894
error: 18.576285722490923
error: 47.573115743757945
average error of f3 is28.290993268879703

error: 17.45019952503334
error: 16.06811104810356
error: 36.663336472022486
error: 19.54284400132725
error: 50.957713893585534
average error of f4 is28.136440988014435

error: 19.25610378322116
error: 17.238535143904976
error: 39.13032683687953
error: 18.558426932961225
error: 47.52904231945235
average error of f7 is28.34248700328385

error: 19.95144443347013
error: 17.24479262786886
error: 38.35709716627519
error: 18.321690509264315
error: 46.27745360442002
average error of f9 is28.030495668259704

error: 19.31391632689232
error: 17.14614967787095
error: 39.430298499637566
error: 18.613391099061793
error: 47.486473781013004
average error of f10 is28.39804587689513

error: 18.353779365574724
error: 17.339706808282784
error: 38.07684381707185
error: 17.271961766293913
error: 47.41972453392596
average error of f12 is27.69240325822984

Smallest cross-validation error is 27.69240325822984
The test error is 18.093740159028776
S6: [13, 6, 11, 8, 5, 12]
W parameters: [2.69006025e+01 -4.88067931e-01 4.81696141e+00 -9.57756553e-01
-1.16805121e+00 -1.71179267e+01 9.18095200e-03]

For k = 7

error: 18.341904004218787
error: 17.963993350200685
error: 38.707568894987375

error: 17.263391400711185
error: 47.41285290455891
average error of f1 is27.93794211093539

error: 18.128950584703244
error: 18.257014522063002
error: 37.54679367787536
error: 16.21676958413045
error: 46.65472895877793
average error of f2 is27.360851465509995

error: 18.365386746724592
error: 17.45092940440722
error: 38.189759638820995
error: 17.263738417142076
error: 47.7667583182202
average error of f3 is27.807314505063015

error: 16.74072069175358
error: 16.74289573953337
error: 35.91169024360174
error: 18.404625931717266
error: 50.916197089344905
average error of f4 is27.74322593919017

error: 18.35588252654588
error: 17.936781521174215
error: 37.91241351536637
error: 17.149303096814148
error: 47.91746261725501
average error of f7 is27.854368655431124

error: 18.634685342811288
error: 18.09785643051573
error: 36.47389602852724
error: 16.487543774132575
error: 45.96975010137124
average error of f9 is27.132746335471616

error: 18.445533571722265
error: 18.20863469923806
error: 38.04898439263867
error: 17.197582838049634
error: 47.417614766103576
average error of f10 is27.863670053550443

Smallest cross-validation error is 27.132746335471616
The test error is 19.462765078975302
S7: [13, 6, 11, 8, 5, 12, 9]
W parameters: [3.27615366e+01 -5.03283679e-01 4.58929457e+00 -1.16056224e+00 -1.18620468e+00 -2.18679523e+01 1.17654439e-02 1.35162756e-01]

For k = 8

error: 18.44278767319284
error: 17.604294637546293
error: 36.852459095004676
error: 16.149064055011184
error: 45.46670734967984
average error of f1 is26.903062562086962

error: 18.370067746685987
error: 18.70695343382857
error: 36.31576322696834
error: 15.81619973691514
error: 45.56596590176624
average error of f2 is26.95499000923286

error: 18.56271337036138
error: 18.054409756840638
error: 36.571708962085765
error: 16.448961347574603
error: 46.25211115628357
average error of f3 is27.177980918629192

error: 17.27469296408002
error: 17.296023677866764
error: 34.47850924175904
error: 17.707911619106046
error: 49.35851121917838
average error of f4 is27.22312974439805

error: 18.617844652423116
error: 18.527799960374725
error: 36.407992907076604
error: 16.47682123107051
error: 46.433878108587045
average error of f7 is27.292867371906397

error: 18.378208350131803
error: 16.960755003083758
error: 35.875089527236824

error: 16.573268779601495
error: 45.44940936853648
average error of f10 is26.647346205718073

Smallest cross-validation error is 26.647346205718073
The test error is 19.27505630594479
S8: [13, 6, 11, 8, 5, 12, 9, 10]
W parameters: [3.46225325e+01 -4.96337482e-01 4.45430100e+00 -1.11427647e+00 -1.21208359e+00 -1.96247290e+01 1.13220816e-02 2.88960229e-01 -1.07676784e-02]

For k = 9

error: 18.1308741316447
error: 16.431876121228683
error: 36.26966649440663
error: 16.26205428401992
error: 44.89975638956379
average error of f1 is26.398845484172746

error: 18.11874204780021
error: 17.33734364142002
error: 35.559242400979215
error: 15.520958844236665
error: 44.82866561636056
average error of f2 is26.27299051015933

error: 18.620134997512793
error: 17.080528609463087
error: 35.75438920979561
error: 16.578298358746682
error: 45.33589416693292
average error of f3 is26.673849068490217

error: 17.149600571538844
error: 16.31439484127489
error: 34.189172236446495
error: 17.694320322613954
error: 48.81190933165675
average error of f4 is26.831879460706187

error: 18.40613400133334
error: 17.372783029080725
error: 35.79720157138501
error: 16.617324321970045
error: 45.93967800992005

average error of f7 is26.82662418673783

Smallest cross-validation error is 26.27299051015933
The test error is 19.013426564771507
S9: [13, 6, 11, 8, 5, 12, 9, 10, 2]
W parameters: [3.35925212e+01 -5.02626795e-01 4.23640458e+00 -9.48995030e-01
-1.51479498e+00 -1.79905123e+01 1.13335465e-02 2.95721691e-01
-1.30092044e-02 4.53579176e-02]

For k = 10

error: 17.792215886133075
error: 16.845460584900284
error: 35.96325418165597
error: 15.049187599855436
error: 44.13697760946196
average error of f1 is25.95741917240134

error: 18.438540264496883
error: 17.54600080277295
error: 35.352536867539946
error: 15.49189096686744
error: 44.61527538738355
average error of f3 is26.288848857812148

error: 16.76249229954875
error: 16.545891181701904
error: 33.848573932813274
error: 16.857537659734145
error: 48.27678897331309
average error of f4 is26.45825680942223

error: 18.186816446053925
error: 17.62708201894164
error: 35.60805333718975
error: 15.689347487863028
error: 45.14553210135307
average error of f7 is26.451366278280283

Smallest cross-validation error is 25.95741917240134
The test error is 17.604229121891038
S10: [13, 6, 11, 8, 5, 12, 9, 10, 2, 1]
W parameters: [3.54957044e+01 -4.83051005e-01 4.12920009e+00 -9.60663217e-01
-1.58334004e+00 -1.89352018e+01 9.87709314e-03 3.51052893e-01
-1.34686180e-02 4.93585690e-02 -9.84436382e-02]

For k = 11

error: 18.11493735514093
error: 17.018622014927075
error: 35.80675569492645
error: 15.032215512538825
error: 43.96716732398697
average error of f3 is25.98793958030405

error: 16.546340227765743
error: 16.01168012580616
error: 34.423926710689656
error: 16.449495671662493
error: 47.51006234384239
average error of f4 is26.18830101595329

error: 17.86797422518968
error: 17.119002736751742
error: 36.02179280921575
error: 15.264589961661342
error: 44.45126473187465
average error of f7 is26.144924892938633

Smallest cross-validation error is 25.98793958030405
The test error is 17.79068661998166
S11: [13, 6, 11, 8, 5, 12, 9, 10, 2, 1, 3]
W parameters: [3.58815613e+01 -4.86358180e-01 4.17156027e+00 -9.80791172e-01
-1.53638232e+00 -2.01631268e+01 9.98895660e-03 3.72079886e-01
-1.53553447e-02 5.07243776e-02 -9.64546396e-02 6.96545879e-02]

For k = 12

error: 16.78926967325437
error: 16.20084272454754
error: 34.35336404967953
error: 16.431334567174254
error: 47.49311708862403
average error of f4 is26.25358562065594

error: 18.1992109267974
error: 17.32850297579159
error: 35.86014616819409
error: 15.241035448841693
error: 44.2810697321631
average error of f7 is26.181993050357573

Smallest cross-validation error is 26.181993050357573
The test error is 17.956475727290446
S12: [13, 6, 11, 8, 5, 12, 9, 10, 2, 1, 3, 7]
W parameters: [3.57886448e+01 -4.78465248e-01 4.21285307e+00 -9.77075858e-01
-1.56830846e+00 -1.97391435e+01 1.00487701e-02 3.70551289e-01
-1.53562226e-02 4.98404326e-02 -9.60278024e-02 7.13025269e-02
-6.75682072e-03]

For k = 13

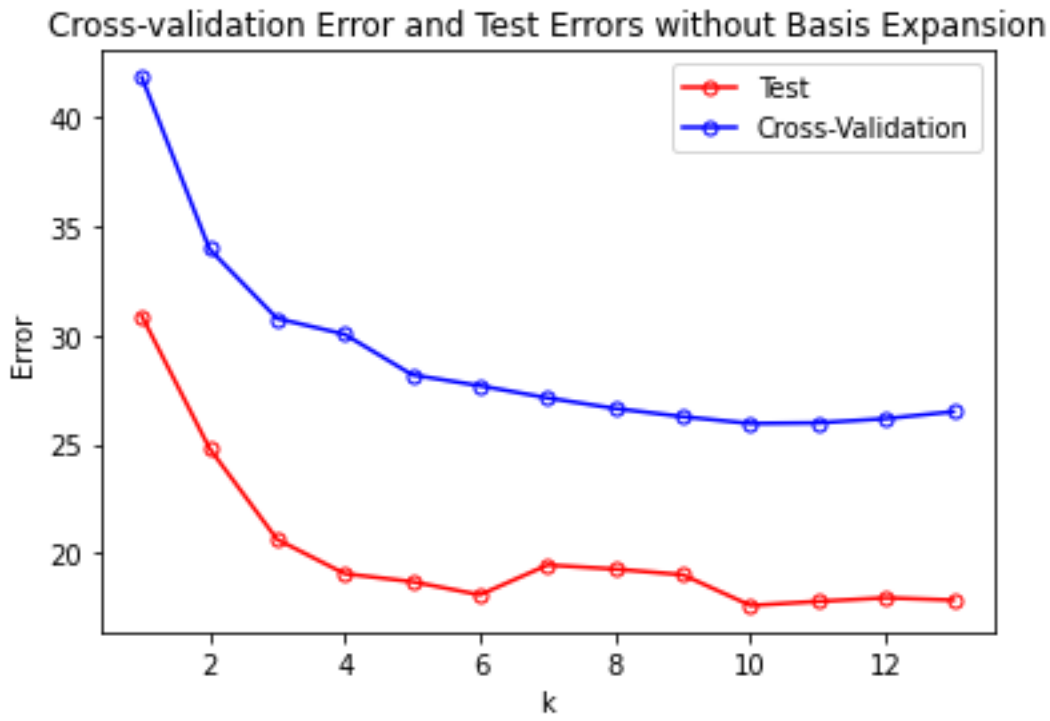
error: 16.84357567211874
error: 16.49876172452921
error: 34.390832102531995
error: 16.515215418822876
error: 48.26332032818037
average error of f4 is26.50234104923664

Smallest cross-validation error is 26.50234104923664
The test error is 17.851545229812043
S13: [13, 6, 11, 8, 5, 12, 9, 10, 2, 1, 3, 7, 4]
W parameters: [3.52261381e+01 -4.72526949e-01 4.14752789e+00 -9.35676890e-01
-1.54405339e+00 -1.94441657e+01 9.20702264e-03 3.36074155e-01
-1.35997433e-02 4.90887388e-02 -8.85287108e-02 4.66160457e-02
-9.21149682e-03 2.97385386e+00]

As a result, the final subset *S is*:

$$S = [f13, f6, f11, f8, f5, f12, f9, f10, f2, f1, f3, f4]$$

Below shows the plot of each cross-validation error and test error of the 13 models created.



As a result, the cross-validation error is larger than the test error for all models. This relation is consistent all throughout. The smallest cross-validation error is at k=10 where the test error is also at its lowest.

In the next part of the assignment, basis expansion has been used to improve the performance of the models. Several functions have been used in attempt to achieve a cross-validation error small than the models without basis expansion. The models used were f(x) = ln(x), f(x) = x^2, f(x) = √x, and f(x) = x_i*x_m.

Below show the results of the errors with the models that computes the least errors for each subset *S*.

For S1: [f13]

Model 1: f(x) = ln(x)
error: 23.06022264286463
error: 22.839376180613854
error: 43.33102591300475
error: 25.575216888601915
error: 37.976136890792134
Cross validation error is: 30.556395703175458
Test error is: 21.622483552740817

Model 2: f(x) = x^2
error: 26.39161546036841
error: 25.213412850048993
error: 50.328974674734226
error: 26.94656964023581
error: 39.63843444013041
Cross validation error is: 33.70380141310358
Test error is: 23.31883639046594

For S2: [f13, f6]

Model 1: f(x) = ln(x)
error: 15.312812675781897
error: 15.895209598192656
error: 28.996488210289442
error: 12.638153303875384
error: 44.12599001884355
Cross validation error is: 23.393730761396586
Test error is: 16.89169121039898

Model 2: f(x) = x^2
error: 15.812904534323723
error: 15.128900868358178
error: 29.291708827843717
error: 11.857873455456001
error: 47.280248972497574
Cross validation error is: 23.874327331695838

Test error is: 15.98102605691785

For S3: [f13, f6, f11]

Model 1: f(x) = ln(x)
error: 14.565800592361494
error: 14.152781869102856
error: 28.749902382460085
error: 10.516918034135724
error: 41.904167757152734
Cross validation error is: 21.977914127042578
Test error is: 14.373960236761604

Model 2: f(x) = x^2
error: 15.376592028117882
error: 13.56287683302528
error: 29.170721353791404
error: 9.899205690749366
error: 44.99987811008633
Cross validation error is: 22.60185480315405
Test error is: 13.636721881100552

For S4: [f13, f6, f11, f8]

Model 1: f(x) = ln(x)
error: 14.744301390240476
error: 13.995268111146665
error: 27.43434467681183
error: 11.222793869839622
error: 40.84909681619695
Cross validation error is: 21.649160972847106
Test error is: 13.952247986844915

Model 2: f(x) = x^2
error: 15.439226402082074
error: 13.48937942439099
error: 27.485096187742133
error: 10.599912955315594
error: 43.6438738446385
Cross validation error is: 22.131497762833856
Test error is: 13.014659885965951

For S5: [f13, f6, f11, f8, f5]

Model 1: f(x) = ln(x)
error: 13.59252982511499

error: 14.788026422031
error: 25.311137273800185
error: 10.89095655205269
error: 36.64313100264981
Cross validation error is: 20.245156215129732
Test error is: 12.93734164848139

Model 2: f(x) = x^2
error: 14.132387275906433
error: 13.658528982071005
error: 26.227895073923275
error: 9.968303950553079
error: 40.43503924584386
Cross validation error is: 20.88443090565953
Test error is: 12.131621291187152

For S6: [f13, f6, f11, f8, f5, f12]

Model 1: f(x) = ln(x)
error: 12.874986370352833
error: 13.788969166420081
error: 24.431368216091887
error: 10.97398170929492
error: 36.589867407105054
Cross validation error is: 19.731834573852957
Test error is: 12.47521635210354

Model 2: f(x) = x^2
error: 13.294972889475927
error: 12.980846020167833
error: 25.29302507488836
error: 10.487816490559561
error: 40.52591768262951
Cross validation error is: 20.51651563154424
Test error is: 11.347642014053813

For S7: [f13, f6, f11, f8, f5, f12, f9]

Model 1: f(x) = ln(x)
error: 13.525401210535154
error: 14.954152947764708
error: 23.641212913608964
error: 10.577396160121664
error: 35.57237306222644
Cross validation error is: 19.65410725885139
Test error is: 13.250903859364193

Model 2: f(x) = x^2
error: 13.884472855678839
error: 14.480651414991758
error: 24.286338777818916
error: 10.272780250042976
error: 39.47788544238008
Cross validation error is: 20.480425748182515
Test error is: 12.81181973627022

For S8: [f13, f6, f11, f8, f5, f12, f9, f10]

Model 1: f(x) = ln(x)
error: 13.381455377408727
error: 13.410590916992925
error: 22.412659759364747
error: 11.377845499478598
error: 34.47653999114146
Cross validation error is: 19.01181830887729
Test error is: 12.919713545638984

Model 2: f(x) = x^2
error: 13.480058322492011
error: 13.15909199964338
error: 23.14520106229419
error: 11.514602486791102
error: 38.508205692387016
Cross validation error is: 19.961431912721544
Test error is: 12.73309695708915

For S9: [f13, f6, f11, f8, f5, f12, f9, f10, f2]

Model 2: f(x) = x^2
error: 13.390148515892363
error: 13.238693536555642
error: 22.78933218614701
error: 11.859371242848631
error: 38.14864090471163
Cross validation error is: 19.885237277231056
Test error is: 12.733791879852783

Model 3: f(x) = sqrt(x)
error: 13.236032313965254
error: 13.317384356207674
error: 22.36190967018442
error: 11.970729127120753

error: 35.73093328812466
Cross validation error is: 19.323397751120552
Test error is: 12.851308288615481

For S10: [f13, f6, f11, f8, f5, f12, f9, f10, f2, f1]

Model 2: f(x) = x^2
error: 11.462547649288895
error: 11.799691919550131
error: 21.8195060162936
error: 10.692422975888265
error: 36.16476578340432
Cross validation error is: 18.38778686888504
Test error is: 12.19171891165555

Model 3: f(x) = sqrt(x)
error: 11.862625124147945
error: 14.815302421854797
error: 21.25762062308581
error: 10.360458041446861
error: 33.47801753572096
Cross validation error is: 18.354804749251276
Test error is: 11.646538764123877

For S11: [f13, f6, f11, f8, f5, f12, f9, f10, f2, f1, f3]

Model 2: f(x) = x^2
error: 11.629990128379166
error: 11.963235071532738
error: 21.66590249959679
error: 10.717985027676448
error: 36.12155314057926
Cross validation error is: 18.419733173552878
Test error is: 12.300936039296653

Model 3: f(x) = sqrt(x)
error: 12.197873263415882
error: 14.630471508486043
error: 21.140033446655888
error: 10.26490554054414
error: 33.43225320018659
Cross validation error is: 18.33310739185771
Test error is: 11.732106197515666

For S12: [f13, f6, f11, f8, f5, f12, f9, f10, f2, f1, f3, f7]

Model 2: f(x) = x^2
error: 11.77328131745768
error: 12.268083975086125
error: 21.79260827934486
error: 10.954987535090396
error: 37.10494277526869
Cross validation error is: 18.778780776449548
Test error is: 12.38134223212504

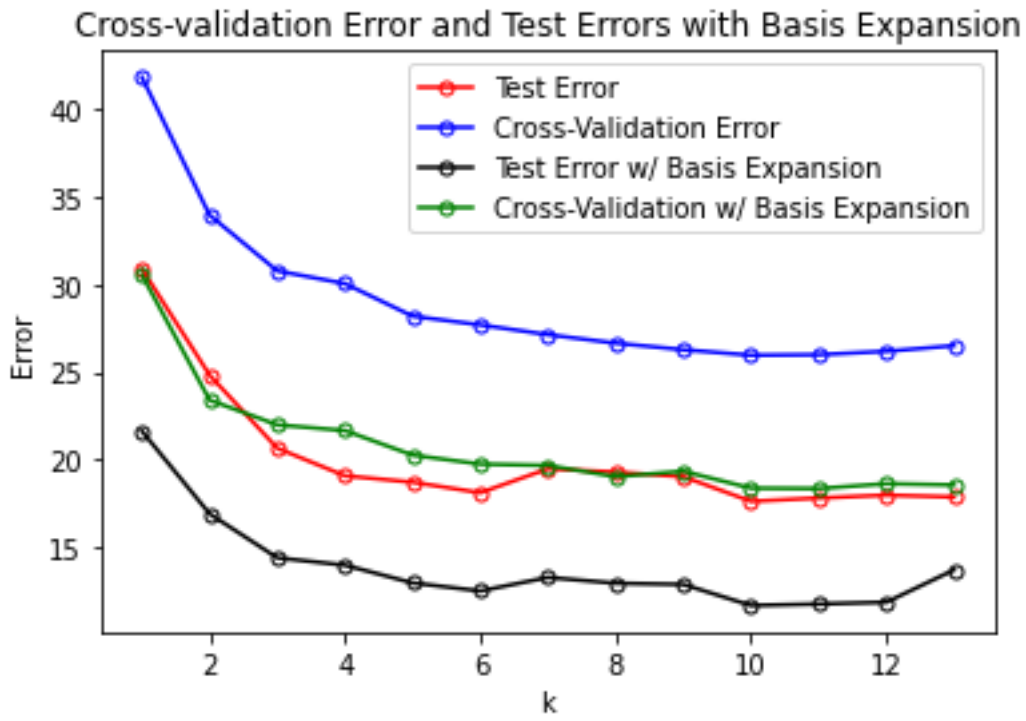
Model 3: f(x) = sqrt(x)
error: 12.34745055884592
error: 14.859878779239507
error: 21.25515065584801
error: 10.464306662240157
error: 34.12592569507589
Cross validation error is: 18.610542470249896
Test error is: 11.833814947778327

For S13: [f13, f6, f11, f8, f5, f12, f9, f10, f2, f1, f3, f7, f4]

Model 2: f(x) = x^2
error: 723.3233732589167
error: 9.43890176635739e+25
error: 349.1394105215879
error: 8.03819153061025e+27
error: 403.9521665616662
Cross validation error is: 1.6265161096547648e+27
Test error is: 1.3072689611181637e+27

Model 4: f(x) = xi * xm
error: 8.954880992597673
error: 13.898144802646907
error: 19.57177395031335
error: 8.932287422718806
error: 41.34040175357545
Cross validation error is: 18.53949778437044
Test error is: 13.680884810255584

Below shows the plot of the cross-validation errors and of the test errors of the 26 models built, versus *k*.



As a result of the models with basis expansion, it is evident that there are improvements in the error. However, the relation stayed fairly the same between cross-validation and test errors compared to the models without basis expansion. The plots and errors still resulted in cross-validation error being larger than the test errors and it stayed consistent throughout all models. The same models have the same smallest test error which is at k=10.

All the widely used basis functions listed in the lectures slides has been used for my trials. This experiment required more than 2 models with basis functions as some functions compute errors in the data set, for example with model 1, f(x) = ln(x), this model could not be used for every feature as some data sets contain 0 and the natural logarithm of zero is undefined. For model 3 with the square root function, some values produce “nan”, meaning “not a number”, which is also undefined and unpresentable. But model 1 and model 3 was kept into the trials because it would produce the smallest errors in some of the other features.

In conclusion, linear regression, basis expansion and feature selection has been successfully experimented in this assignment. The goal of the assignment has been achieved.

Sources:

https://scikit-learn.org/stable/modules/cross_validation.html
https://machinelearningmastery.com/k-fold-cross-validation/
https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html