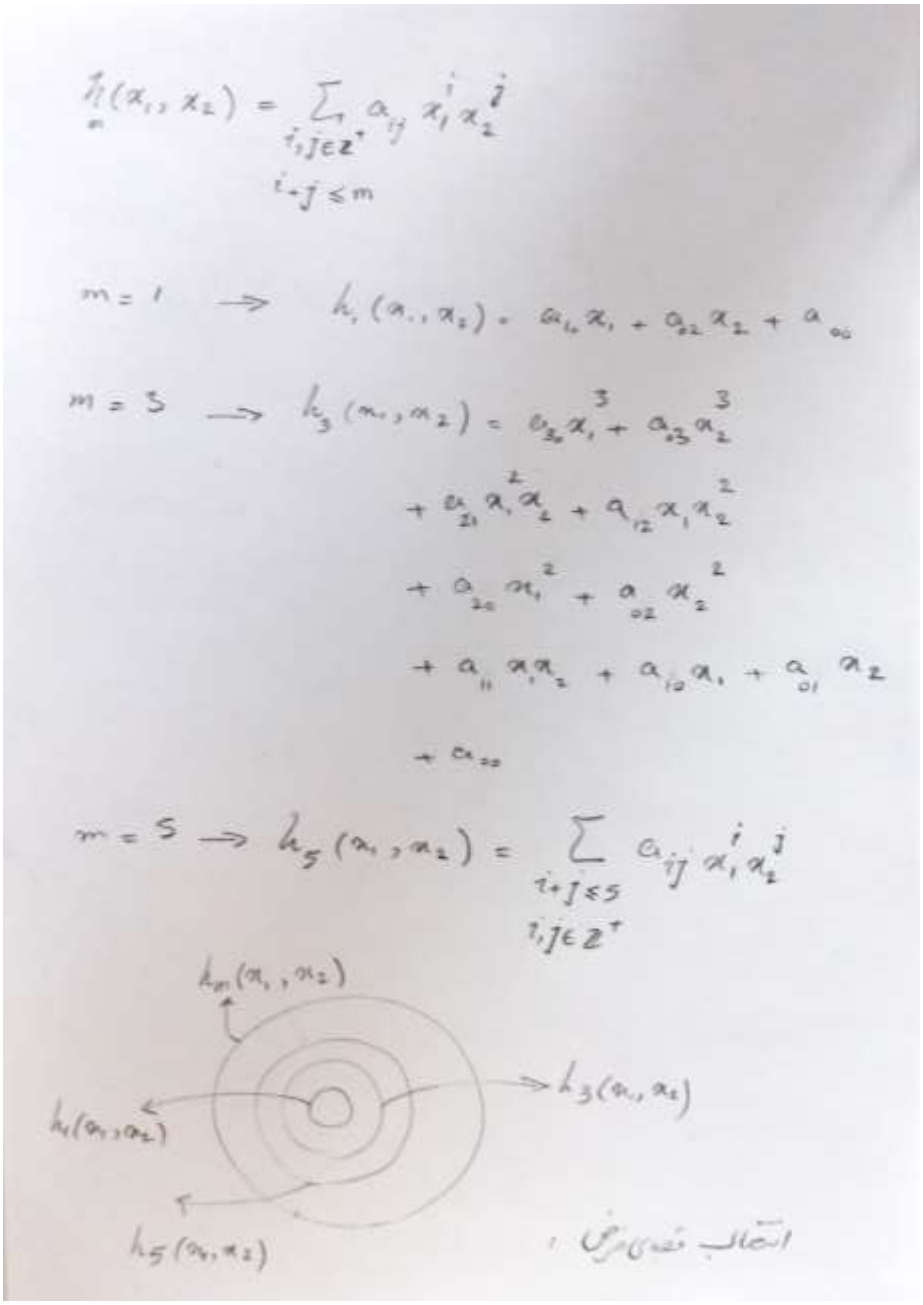
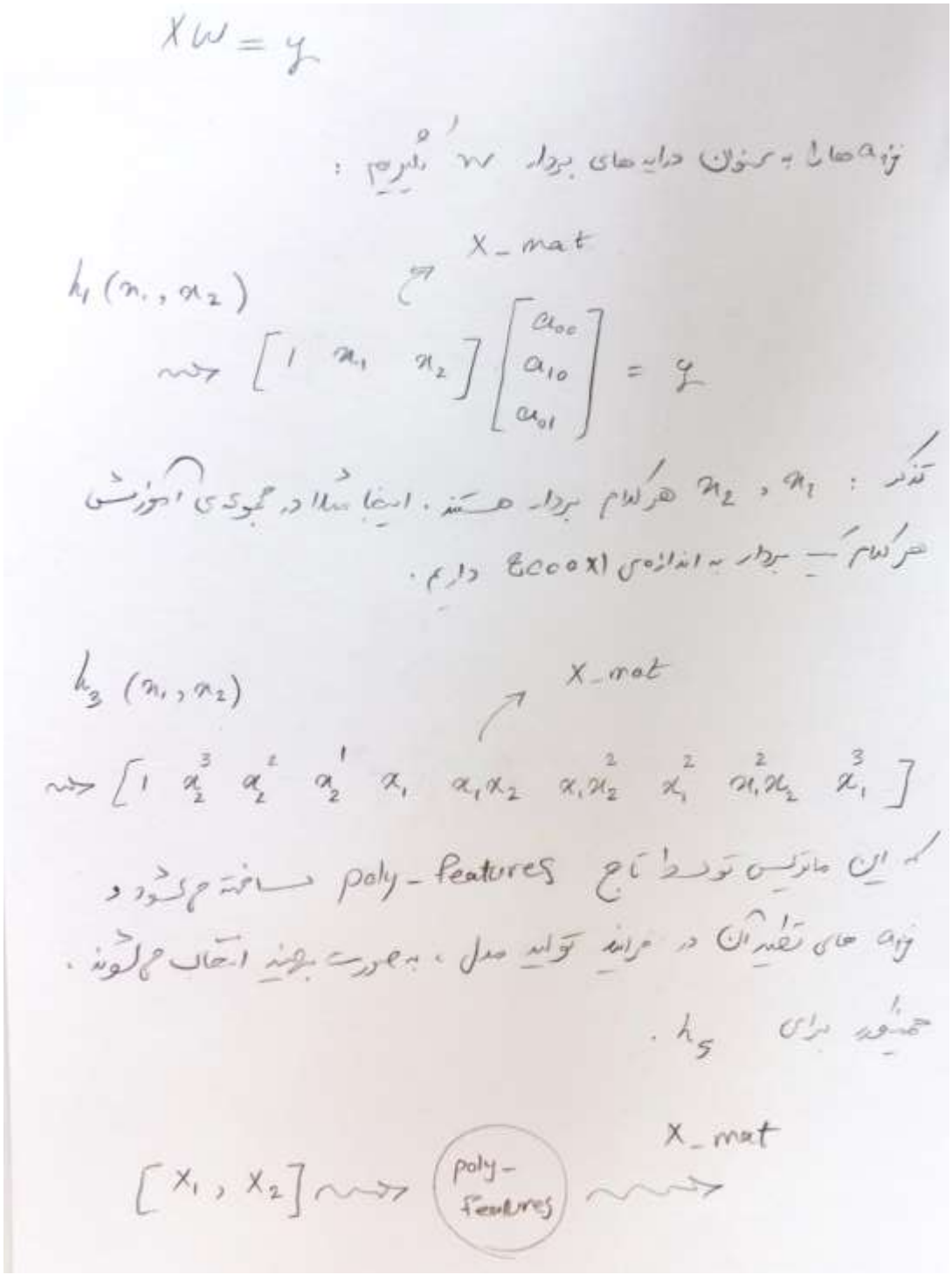
درس یادگیري ماشین
گزارش تکلیف برنامه نویسی: رگرسیون
کدها و اعداد به دست آمده: فایل ضمیمه HW1.ipynb

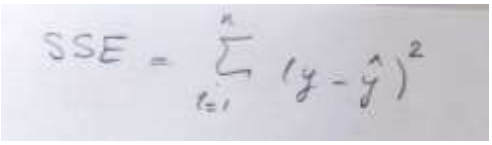
قبل از ورود به بخش برازش ها، لازم اسـت درباره ی تابع poly_features و نحوه ی تولید ماتریس داده ها توضـیح بدهم. قرار است مدل رگرسیون چندجمله ای تولید کنیم. یعنی مجموعه ی توابع فرض ما به فضای توابع دو متغیره ی h است که در هر قسمت خطی، درجه ۳ و ۵ به صورت زیر محدود شده.



مدل نهایی، همان مقادیر ضرایب مجهول است. در مجموعه داده ها دو تا ستون x1 و x2 وجود دارند ولی مدل علاوه بر این دو مقدار به **حاصل ضرب های از درجه حداکثر درجه ی فضای فرض از این دو متغیر** نیاز دارد. در poly_features به ازای درجه ی فضای توابع فرض، این حاصل ضرب ها روی دو بردار ورودی x1 و x2 محاسبه و به صورت ماتریس x آماده برای استفاده در معادله ی رگرسیون چندجمله ای در اختیار قرار داده می‌شود. در تمام بخش ها هم برای برازش مدل روی داده های آموزشی و هم برای براورد داده های تست و اموزش از همین ماتریس استفاده می‌شود.

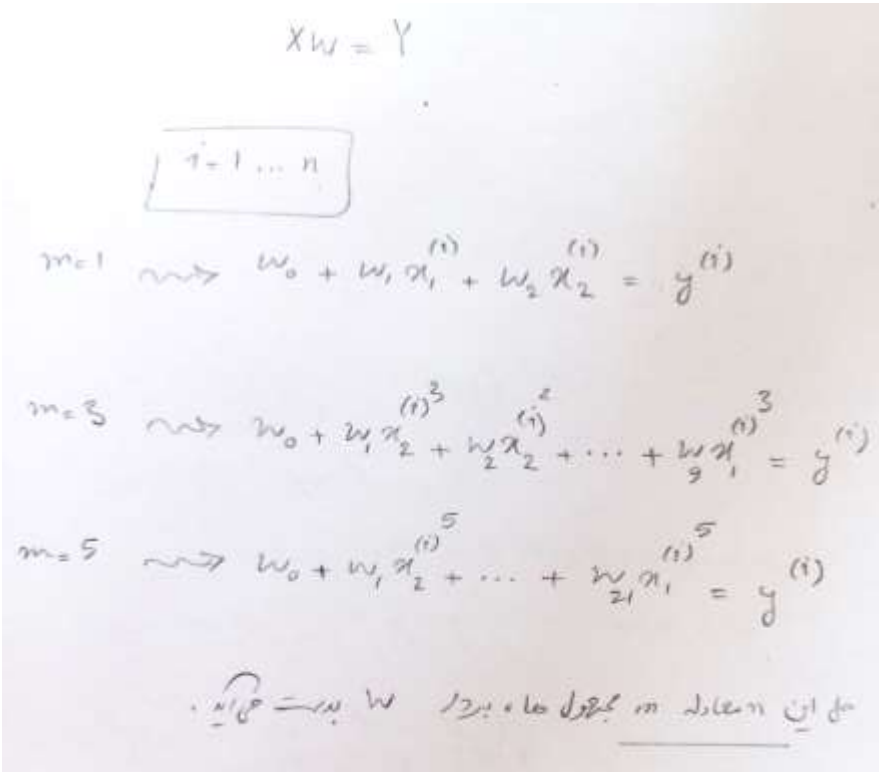


برای به د ست آوردن معیار خطا که تابع SSE انتخاب شده، لازم ا ست مقادیر y را به ازای داده های آموزشی و ت ست را هر دسته به صورت جداگانه با مدلی که به دست آمد پیش بینی کنم. در نهایت با استفاده از این مقادیر و مقادیر هدف داده شده در داده ها، در تابع sum_squere_error قابل محاسبه ست و در تمام بخش های همین خطا گزارش شده.

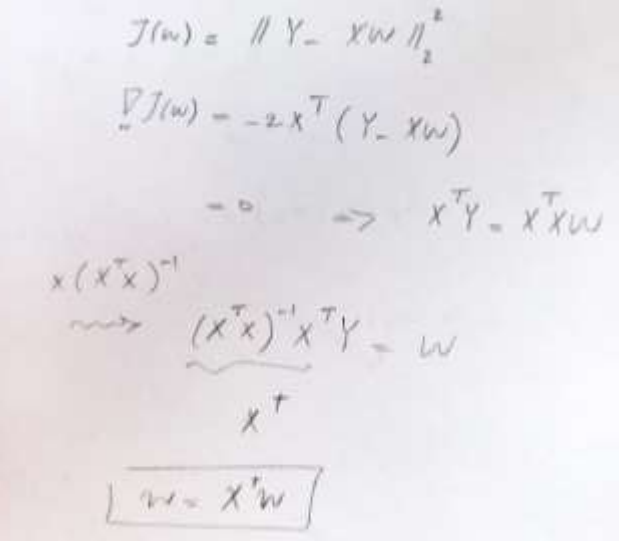


a) به دست آوردن معادله رگرسیون از فرمول بسته

معادله ی رگرسیون یک د ستگاه خطی به فرم Xw = y ه ست. در هر مورد خطی، درجه ۳ و درجه ۵ معادله ها و مجهول ها به صورت زیر هستند.



برای حل این دستگاه ها، در حالت کلی که ماتریس ضرایب دستگاه، مربعی نباشد، با استفاده از فرمول شبه معکوس، مقدار مجهول قابل محاسبه ست. در استفاده از روش فرمول بسته، از همین فرمول استفاده می‌کنم. تابع multivar_poly_regression این کار را انجام می‌دهد.



مقادیر زیر به ازای هر کدام از مدل ها به دست آمده.

	SSEtrain	SSEtest
رگرسیون خطی	18317759690.368008	**12944759955.772043**
رگرسیون چندجمله ای درجه ۳	1.6877332437997384e-18	**7.636451083577081e-19**
رگرسیون چندجمله ای درجه ۵	7.823644442756832e-17	**6.206512135546437e-17**

مدل خطی که به هیچ وجه قابل ا ستفاده نی ست و مقدار خطای زیاد آن ن شان دهنده ی این ا ست که این داده ها رابطه ی خطی ندارند و پیچیدگی فضـــای فرض بســـیار کم اســـت. روی چندجمله ای های درجه ۳ و ۵ مقادیر خطا هر دو قابل قبول ه ستند که البته در درجه ۵ مقدار بهتر هم ه ست. از روی نزدیکی مقدارهای ت ست و اموزش هردوی اینها می توان فهمید که حداقل تا درجه ی ۵ مدل پیچیدگی مناسب دارد و با این تعداد داده مشکل بیش برازش نداریم که البته در این شرایط انتخاب کمترین پیچیدگی یعنی همان درجه ۳ مناسب تر است.

ضرایب به دست آمده در هر روش :

```
reg1
```

```
Out[7]: array([-4226.07005149,   446.63582426,   537.30127365])
```

رگرسیون خطی

```
reg2
```

```
Out[10]: array([ 1.00000000e+00,  3.62376795e-13,  2.00000000e+00, -4.12170298e-15,
                 3.00000000e+00,  6.03961325e-14,  4.00000000e+00,  6.75015599e-14,
                -7.66053887e-15,  2.55351296e-15])
```

رگرسیون چندجمله ای درجه۳

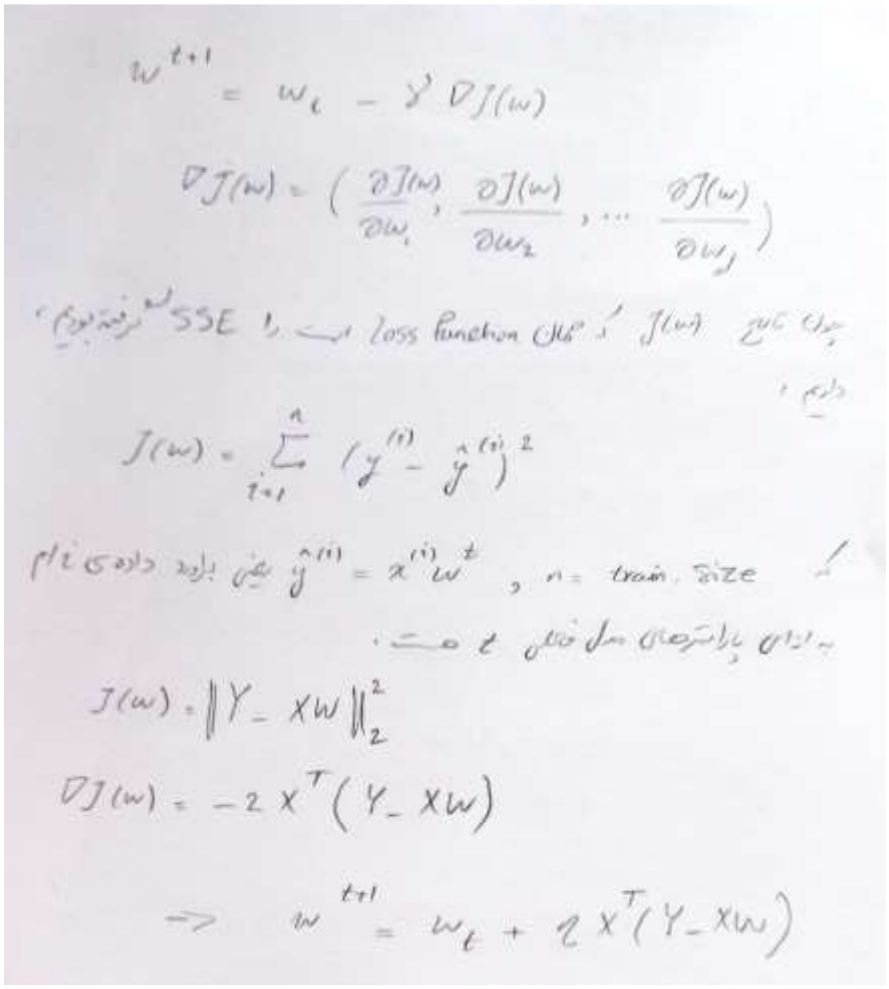
```
reg3
```

```
Out[13]: array([ 1.00000000e+00, -3.41628947e-11,  2.00000000e+00,  1.11910481e-13,
                -8.71525074e-15,  2.24646690e-16,  3.00000000e+00,  8.77520279e-12,
                 4.00000000e+00, -5.55111512e-16,  2.51534904e-17,  3.38218342e-12,
                -1.56319402e-13,  1.16573418e-14, -9.04658293e-16, -2.06057393e-13,
                -9.76996262e-15,  5.75928194e-16,  4.26325641e-14,  8.29197822e-16,
                -1.24900090e-15])
```

رگرسیون چندجمله ای درجه۵

b) به دست آوردن معادله رگرسیون از روش گرادیان کاهشی

روش گرادیان کاه شی یک روش iterative ه ست که در این روش ها، به جای حل معادله ی رگر سیون از یک روش تکراری برای بازتولید یک بردار ضرایب استفاده می‌شود که این مدل جدید هر بار از مقدار قبلی ، خطای کمتری روی داده های آموزش داشته و در نهایت تقریب خوبی از همان جواب اصلی معادله ی رگرسیون خواهد بود با این تفاوت که حجم محاسباتی که لازم است تا به این تقریب قابل قبول رسید، از حجم محاسبات حل معادله کمتر است. اینجا از گرادیان کاهشی برای کمینه سازی خطای MSE ا ستفاده کردم که هر بار وزن ها را به انداره ی gamma در جهت عکس مقدار گرادیان محا سبه می‌کند. چون فضای فرض را می دانم، مستقیم از رابطه ی زیر برای بهینه سازی استفاده می‌کنم.



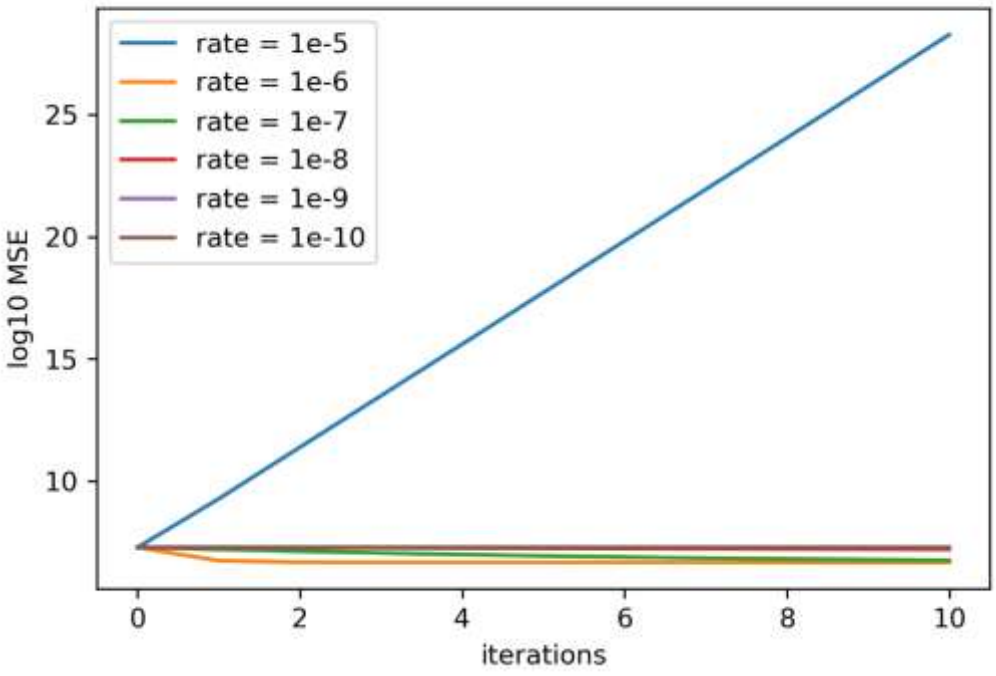
برای قابل درک تر بودن خطا روی هر داده، برای نمودارها از MSE که از فرمول زیر محاسبه می‌شود استفاده کردم. روش محاسبات هم هربار استفاده از کل داده های آموزش بوده یعنی محاسبه ی دسته ای (batch mode) پس اینجا MSE همواره ضریب ثابتی از همان SSE هست و تفاوتی در کار ندارد.

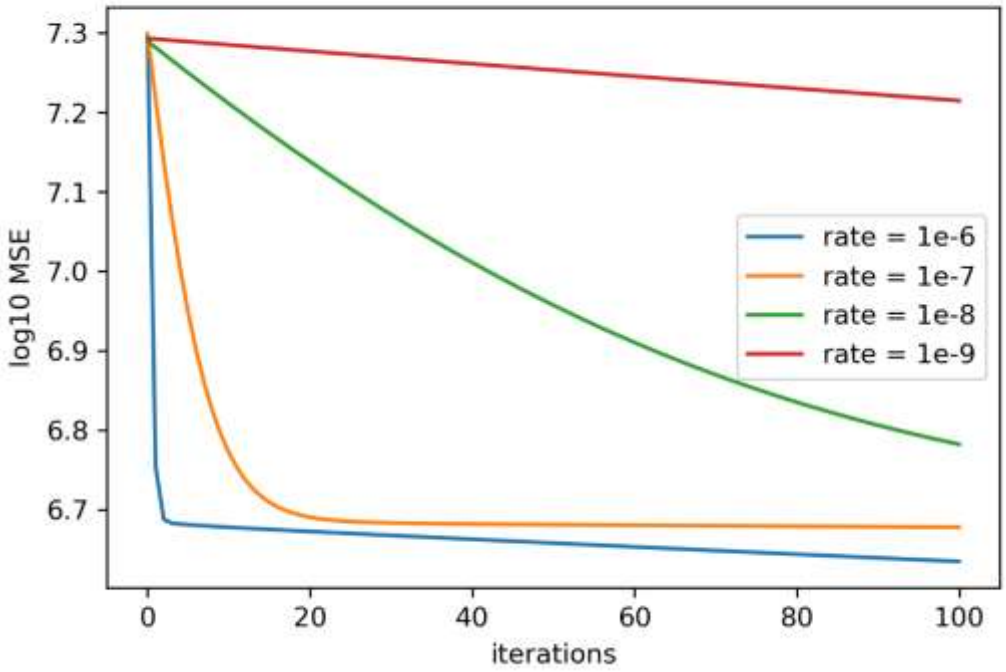
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2 = \frac{1}{n}SSE$$

ماتریس داده های استفاده شده در این روش هم همان ماتریسی است که در روش اول با تابع poly_features تولید و استفاده شد.

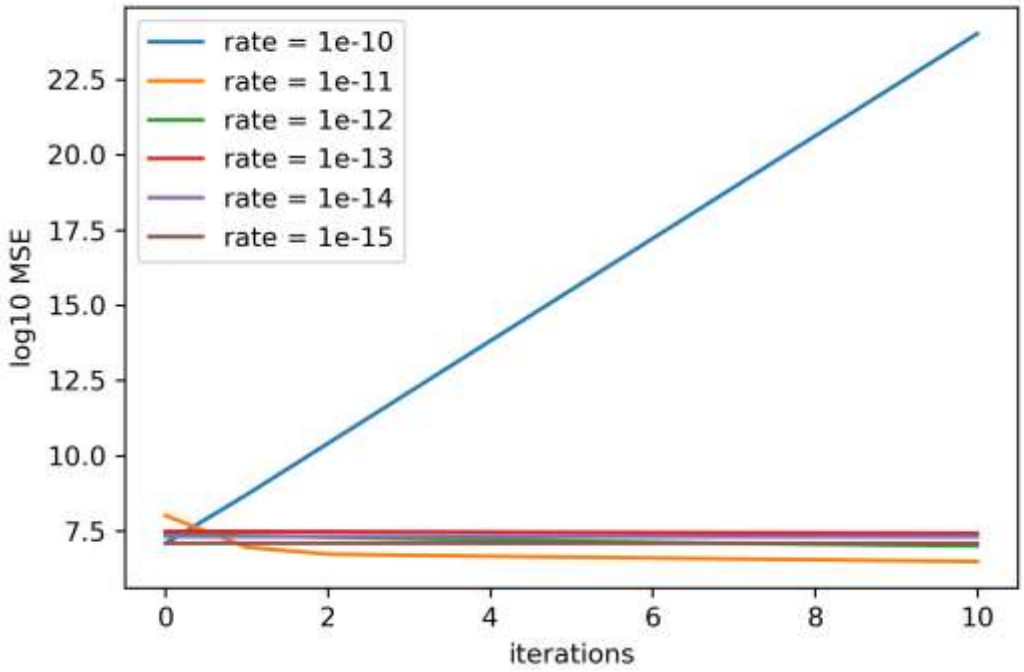
تمام بخش های تو ضیح داده شده در تابع grad_descent پیاده سازی شده. این تابع ۳ پارامتر برمیگرداند که از بین انها فقط weight پارامتر مدل است. از دوتای دیگر فقط برای سنجش gamma ی مناسب و انتخاب آن در مدل نهایی استفاده کردم.

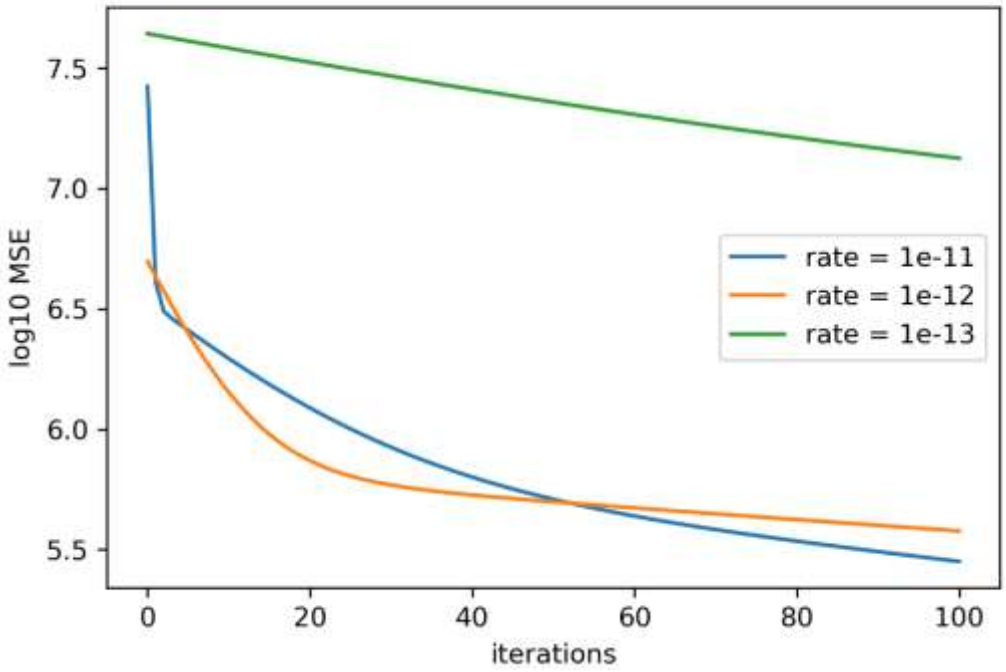
برای اینکه مقدار مناسبی برای gamma انتخاب بشود، قبل از تولید هر یک از سه مدل، اول چند تا مدل با اندازه ی gamma های مختلف انتخاب کردم و با توجه به منحنی کاهش MSE هرکدام از gamma ها (که در لگاریتم ۱۰ رسم شده) یکی را انتخاب کردم.



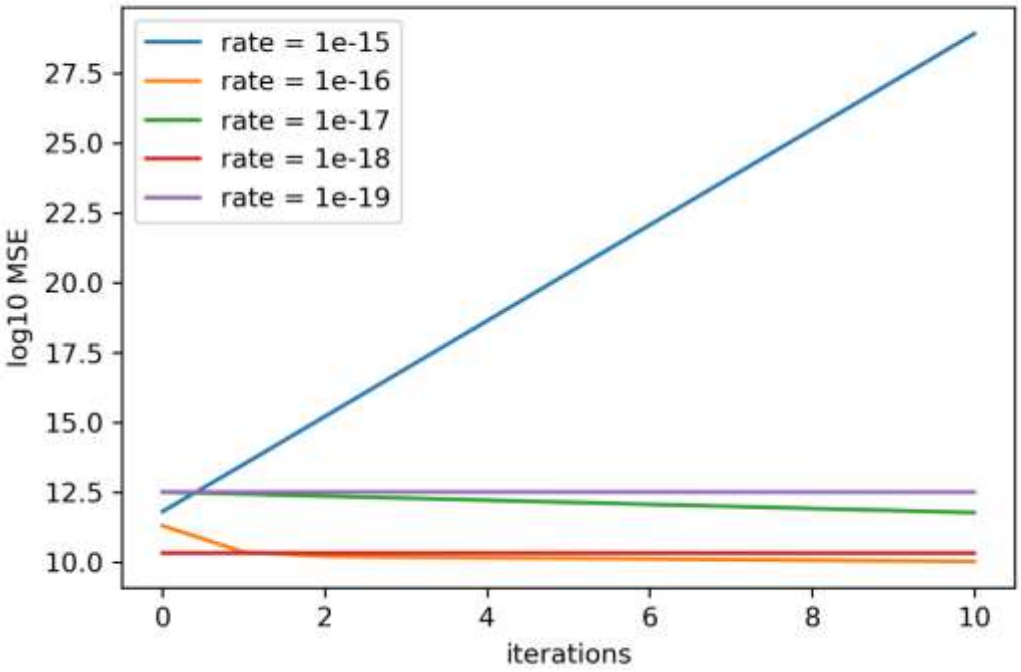

7.3
7.2
7.1
7.0
6.9
6.8
6.7
log10 MSE
0
20
40
60
80
100
iterations
rate = 1e-6
rate = 1e-7
rate = 1e-8
rate = 1e-9

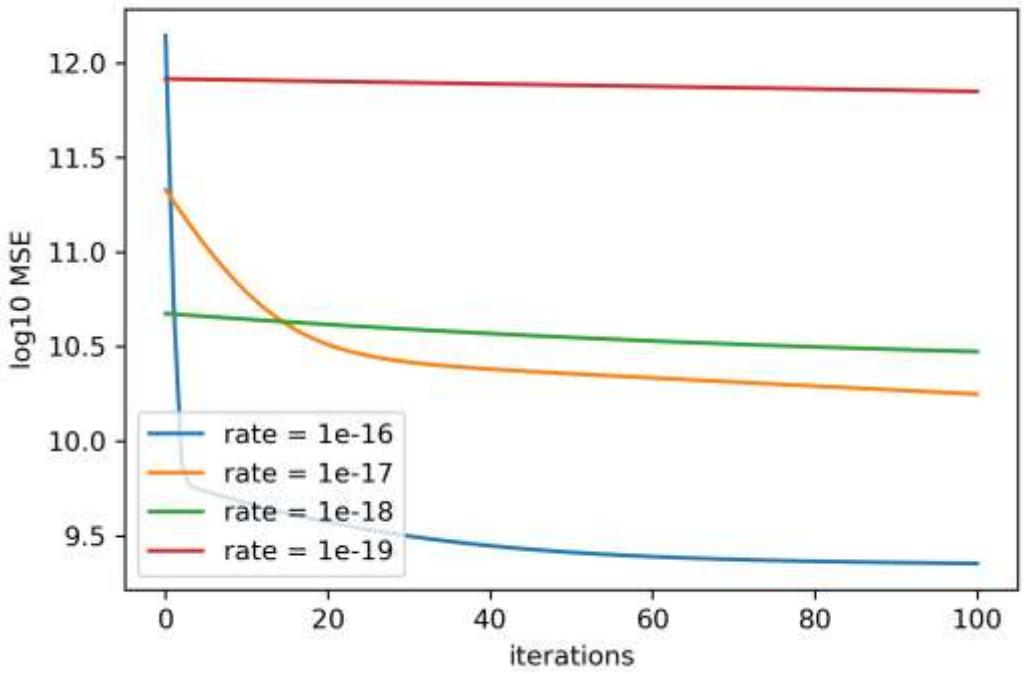
رگرسیون خطی


22.5
20.0
17.5
15.0
12.5
10.0
7.5
log10 MSE
0
2
4
6
8
10
iterations
rate = 1e-10
rate = 1e-11
rate = 1e-12
rate = 1e-13
rate = 1e-14
rate = 1e-15



رگرسیون چندحمله ای درجه ۳





رگرسیون چندحمله ای درجه ۵

روی نمودارها دقت شود که به ازای مقادیر بزرگتر gamma مدل واگرا می‌شود. در نهایت برای سه مدل به ترتیب مقادیر گاما را برابر 1e+7 و 1e+11 و 1e+16 انتخاب کردم.

شرط توقف اموزش در همه ی موارد یکسان و برابر ۱۰۰ بار اجرای چرخه ی بازتولید مدل است. این مقدار را می‌توان متفاوت در نظر گرفت یا می‌شود تعداد تکرارها را وابسته به خطای به دست آمده کرد ولی اینجا چون خواستم همه ی ۳ مدل را یکسان تولید و مقایسه کنم و از طرفی با ازمایش های مختلف مقدار gamma حدود ضرایب و خطای به دست آمده به سرت همگرا می‌شدند، به نظرم آمد همین تعداد تکرارها برای شرط توقف کافی باشد.

مقدارخطای SSE داده های اموزش و تست برای هر سه مدل در زیر آورده شده.

	SSEtrainGD	SSEtestGD
رگرسیون خطی	38097270716.17465	**25775186283.36409**
رگرسیون چندجمله ای درجه ۳	2872831801.8607516	**1882719031.7481675**
رگرسیون چندجمله ای درجه ۵	9628195119367.512	**12180370540265.457**

همینطور که بعد از دیدن نتایج بخش اول انتظار میرفت مدل خطی غیرقابل استفاده است. دو مدل درجه ۳ و ۵ کارکرد خوبی دارند و فرقی نمی کند شیوه ی حل مساله به لحاظ محاسباتی، روش تحلیلی یا شد یا تقریب و تکرار، در نهایت مدلی که از فضای فرض به عنوان جواب بهینه انتخاب می شود، به لحاظ خطا در یک حدود مشخص قرار می‌گیرد. (با فرض اینکه روش

تکراری دچار مشـکلاتی از قبیل واگرایی یا اکسـترمم های موضـعی نشـده باشـد.) اما اینجا با اینکه رگرسـیون درجه ۳ و ۵ در تکرارها با گامای انتخاب شده همگرا شده، ولی مقدار خطا به حدی زیاد است که بازهم مدل ها عملا قابل استفاده نیستند. علت اصلی این مساله تفاوت مرتبه‌ی داده ها و بزرگی نسبی آنهاست. روش های تکراری و مخصوصا گرادیان کاهشی به این فاکتورها حسـاسـیت دارند و برای نتیجه گیری در این روش ها باید از پیش پردازش ها به خصـوص مقیاس بندی داده ها استفاده کرد که البته اینجا خواسته نشده و همین جا نتایج را رها می‌کنم. شاید بشود این مساله را بدون پیش پردازش و با هربار کوچک تر کردن گاما در چرخه تا حدی جبران کرد که نیازمند بررسی های بیشتر هست.

ضرایب نهایی به دست آمده برای مدل ها :

```
w1
```

```
Out[20]: array([-22.08027839, 274.66577143, 154.63077927])
```

رگرسیون خطی

```
w2
```

```
Out[25]: array([-0.64149439, -0.47342055,  0.09590134,  0.41992228, -2.04442711,
                 1.48988695,  2.09382559, -1.65682143,  1.20523848,  1.91695605])
```

رگرسیون چندجمله ای درجه۳

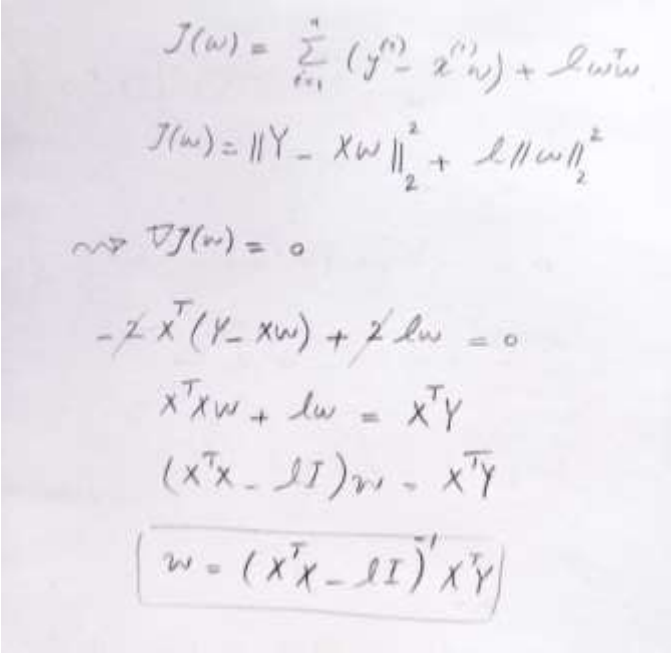
```
w3
```

```
Out[30]: array([ 0.46466906,  0.04814824, -0.21163878, -0.47689156, -1.1021772 ,
                 0.05243408, -0.51706551, -0.15634142,  0.81922125, -1.62995483,
                 0.13774146,  0.22129815,  0.63924855, -0.88840527,  0.59265066,
                 2.61643856, -0.03705765, -1.44369884,  0.64483909,  0.74695394,
                -0.97686382])
```

رگرسیون چندجمله ای درجه۵

c) به دسـت آوردن معادله رگرسـیون با اسـتفاده از جمله‌ی منظم‌سـاز انتخاب‌شـده با روش k-fold cross-validation

اینجا از دو تابع جدید استفاده می‌کنم. اول reg_multivar_poly_regression که فرمول رگرسیون را همراه با یک جمله‌ی منظم ساز l2 برای ماتریس داده های x و مقادیر هدف y پیاده سازی و از روش محاسبه ی فرمول بسته، ضرایب رگرسیون را پیدا کرده و برمی‌گرداند. ماتریس های داده ی استفاده شده در این بخش هم عینا مشابه بخش های قبلی هست که از تابع poly_features به دست آمده بودند.



برای انتخاب مقدار λ مناسب، مدل خواسته شده را به ازای ضرایب مختلف محاسبه می‌کنم. برای این که این مدل ها را باهم مقایسه کنم همان طور که در صورت سوال آمده از روش k-fold cross-validation با مقدار k=5 استفاده میکنم. این عمل در cross_val پیاده سازی شده. در این روش هر بار ۸۰۰۰ داده ی تست را یک بار بر میزند. سپس به ۵ دسته تقسیم می‌کند. هر بار مدل را روی تمام داده های آموزشی به جز دسته ی *i* ام یعنی روی ۶۴۰۰ داده با تابع reg_multivar_poly_regression به دست می‌آورد و در نهایت از داده های دسته ی کنار گذاشته شده برای اندازه گیری خطای اعتبارسنجی استفاده می‌کند. در مجموع ۵ مدل ساخته می‌شود و ۵ تا خطای اعتبارسنحی و ۵ تا خطای اموزشی محاسبه می‌شود. نتیجه ی گزارش شده میانگین این خطاهاست.

در آخر به ازای نتایج گزارش شده از روش k-fold cross-validation به ازای تک تک λ های مجموعه ی زیر، یک نمودار رسم می‌کنم که میزان تاثیر λ های مختلف را روی خطاها ببینم.

{1e-4 ,1e-3, 1e-2 ,1e-1, 1, 1e+1, 1e+2, 1e+3, 1e+4}

برای انتخاب از روی نمودار اولین نقطه‌ی کاهش خطای داده های اعتبارسنجی و افزایش خطای آموزشی در حالتی که بیش برازش داشته باشیم مورد نظر است. در واقع هدف اصلی انتخاب ضریبی است که این دو منحنی را ذوی کمترین مقدار به هم نزدیک کند. ولی در اینجا استفاده از منظم ساز به کار نمی‌اید. در مدل اول که مشکل پیچیدگی کم مدل را دارم که منظم ساز کمکی به حل آن نمیکند. در دو مدل درجه ۳ و ۵ هم اولا که مدل هایی که از بخش a به دست امد کارایی بسیار مطلوب داشتند و مرتبه های خطای تست و اموزش هردو یکسان بودند پس اصلا مشکل بیش برازشی وجود نداشت که نیازی به حل کردن دا شته با شد. دوم هم اینکه اگر یک بار دیگر به ضرایب به د ست آمده ی بخش اول برای درجه ۵ نگاه کنیم، این مدل هرچند درجه ۵ دارد ولی بیشتر ضرایب به دست آمده برای مدل مقدار بسیار کوچک نزدیک صفر دارند و تنها ضرایبی تعین کننده ی مدل هستند که در برای مدل درجه ۳ هم نظیر همان جملات به دست آمده بودند (روی جدول ضرایب مشخص کردم.) مدل درجه ۵ ممکن بود دچار بیش برازش بشود ولی به دلیل وجود تعداد زیاد داده های آموزشی این مشکل برطرف شده و مدل خوب کار می‌کند. در شرایطی که پیچیدگی مدل بیش از نیاز ا ست، کاری که جمله ی منظم ساز می‌دهد همان کوچک کردن ضرایب مدل به ازای جملات درجه های بالاتر است که اینجا به خاطر تعداد کافی داده های آموزشی نیازی به آن نیست.

```
reg2
```

```
Out[10]: array([ 1.00000000e+00,  3.62376795e-13,  2.00000000e+00, -4.12170298e-15,
                 3.00000000e+00,  6.03961325e-14,  4.00000000e+00,  6.75015599e-14,
                -7.66053887e-15,  2.55351296e-15])
```

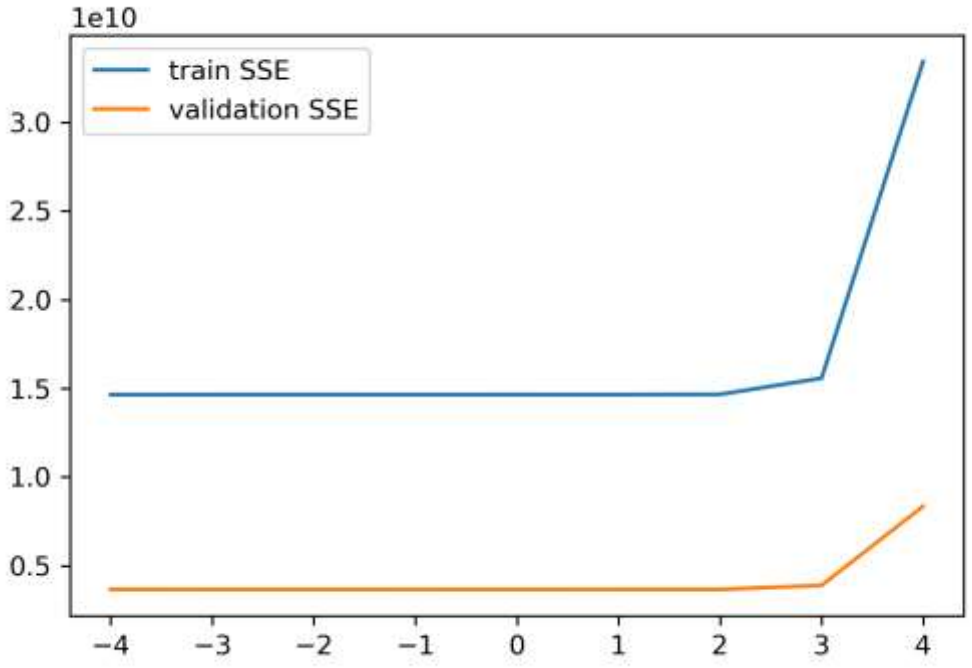
رگرسیون چندجمله ای درجه۳ ازبخش a

```
reg3
```

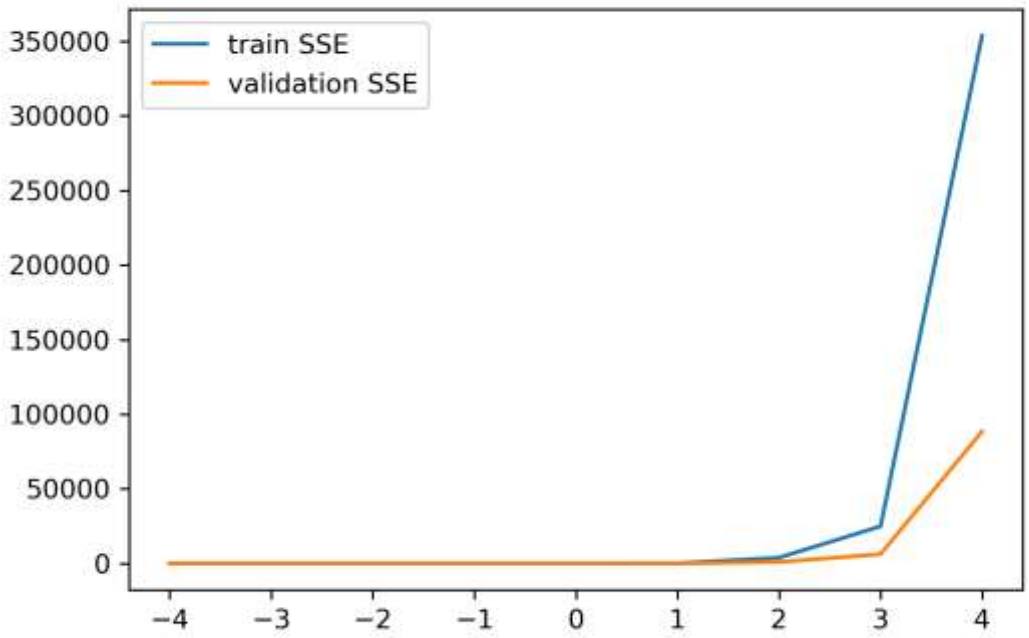
```
Out[13]: array([ 1.00000000e+00, -3.41628947e-11,  2.00000000e+00,  1.11910481e-13,
                -8.71525074e-15,  2.24646690e-16,  3.00000000e+00,  8.77520279e-12,
                 4.00000000e+00, -5.55111512e-16,  2.51524904e-17,  3.38218342e-12,
                -1.56319402e-13,  1.16573418e-14, -9.04658293e-16, -2.06057393e-13,
                -9.76996262e-15,  5.75928194e-16,  4.26325641e-14,  8.29197822e-16,
                -1.24900090e-15])
```

رگرسیون چندجمله ای درجه۵ ازبخش a

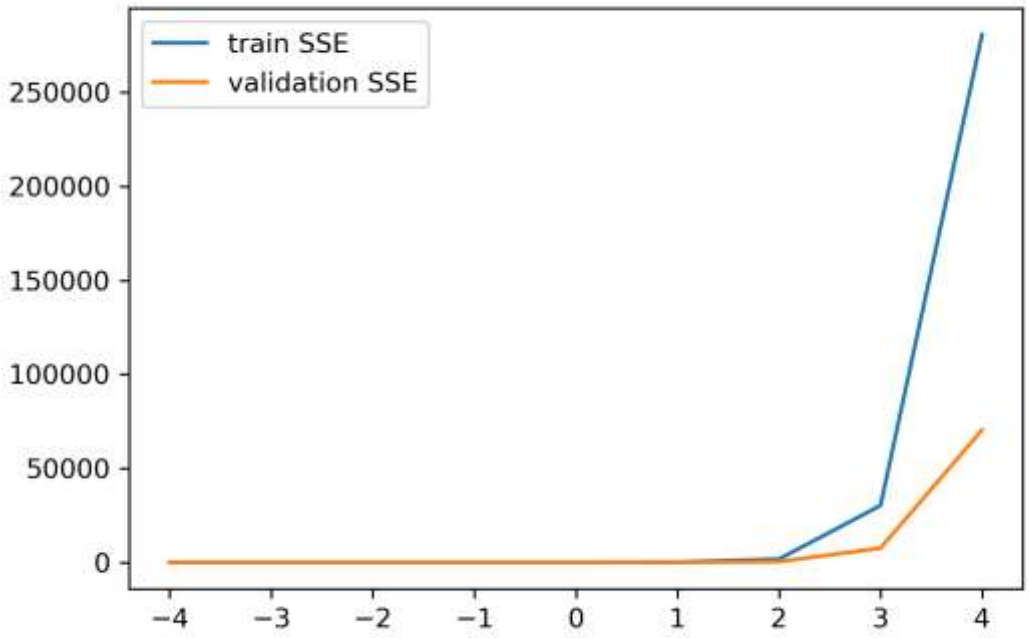
نمودارها به ازای l های مختلف در زیر آورده شده. در هر مورد من l را برابر 1e-4 برای ادامه ی کار در نظر گرفتم.



رگرسیون خطی



رگرسیون چندجمله ای درجه۳



رگرسیون چندجمله ای درجه۵

در آخرین مرحله، مدل نهایی را با تمام داده های آموزشی و و جمله ی منظم ساز انتخاب شده می‌سازم و خطای داده های تست و آموشی را محاسبه می‌کنم که برای هر یک از سه مدل رگرسیون خطی، رگرسیون چندجمله ای درجه ۳ و ۵ به صورت زیر هست. همچنین در جدول زیر مقادیر خطاهای SSE همین مدل که از روش k-fold cross-validation به دست آمده و مورد انتظار بودند هم گزارش شده.

	SSEvalidation	SSEtrainkKfold	SSEtrain	SSEtest
رگرسیون خطی	3670043788.025161	14651324730.91531	18317759690.367973	**12944760110.83464**
رگرسیون چندجمله ای درجه۳	3.1444955092089305e-09	1.2507476793264266e-08	9.9823927225554 69e-09	**9.515612484722473e-09**
رگرسیون چندجمله ای درجه۵	1.6475421309421 08e-06	6.4955715210990 52e-06	1.6958291308018977e-06	**9.571019099293 23e-06**

مدل‌های نهایی نهایی :

```
Out[37]: array([-4226.06995749,   446.63582031,   537.30126523])
```

رگرسیون خطی

```
Out[40]: array([ 1.00000981e+00, -1.67709653e-06,  2.00000010e+00, -2.09853337e-09,
                 2.99999607e+00,  2.76810333e-07,  3.99999999e+00,  5.33582871e-07,
                -1.32669097e-08, -2.36764245e-08])
```

رگرسیون چندجمله ای درجه۳

```
Out[43]: array([ 1.00010918e+00, -8.92428231e-05,  2.00002674e+00, -3.37251546e-06,
                 1.84798952e-07, -3.64004502e-09,  2.99994432e+00,  1.11192821e-05,
                 3.99999894e+00,  4.95376987e-08, -8.83475505e-10,  1.85164678e-05,
                -1.70572619e-06,  7.06013821e-08, -1.14270759e-09, -3.06571614e-06,
                 1.37700855e-07, -2.11165419e-09,  2.43517660e-07, -4.34835093e-09,
                -7.42576395e-09])
```

رگرسیون چندجمله ای درجه۵