حل تمرین درس یادگیری ماشین

سری ۱

نام و نام خانوادگی دانشجو:

همایون حیدرزاده (۹۵۱۳۱۰۷۰)

نام استاد: دکتر ناظرفرد

آبان ۹۶

-۱

یادگیری با نظارت:

یک دسته از روش‌های یادگیری ماشین است که در آن یادگیری توسط یک معلم انجام می‌شود. معلم در این روش می‌تواند برچسب داده‌ها باشد. در این روش یادگیری ماشین هدف مدل کردن رابطه (خطی یا غیر خطی) بین داده آموزشی ورودی و برچسب‌های آن‌ها به صورتی است که بتوان برچسب داده‌های آزمایشی را تخمین زد.

یادگیری بدون نظارت:

این روش یادگیری بدون معلم صورت می‌گیرد. بنابرین در این روش برچسبی برای داده‌های مسئله وجود ندارد و معمولا هدف یافتن الگوهایی بر اساس شباهت‌های موجود میان داده‌ها، می‌باشد. مثلا در مسئله Kmeans هدف پیدا کردن خوشه‌های موجود در داده‌ها بر اساس معیار شباهت عکس فاصله است.

یادگیری تقویتی:

در یک مسئله یادگیری تقویتی با عاملی روبرو هستیم که از طریق سعی و خطا با محیط تعامل کرده و یاد میگیرد تا عملی بهینه را برای رسیدن به هدف انتخاب نماید. یادگیری تقویتی از اینرو مورد توجه است که راهی برای آموزش عاملها برای انجام یک عمل از طریق دادن پاداش و تنبیه است بدون اینکه لازم باشد نحوه انجام عمل را برای عامل مشخص نمائیم. همچنین یادگیری تقویتی از دو جنبه با یادگیری با ناظر تفاوت دارد:

- داده‌های آموزشی یادگیری بصورت زوج **<ورودی/ خروجی>** مطرح نمی‌شوند. بلکه بعد از اینکه عامل عملی را انجام داد پاداشی را دریافت میکند و به مرحله بعدی میرود. عامل هیچ گونه اطلاعی در مورد اینکه در هر حالت بهترین عمل چیست را ندارد. بلکه این وظیفه عامل است که در طول زمان تجربه کافی در مورد حالتها، عمل های ممکن، انتقال و پاداش جمع آوری نموده و عملکرد بهینه را یاد بگیرد.
- تفاوت دیگر در اینجاست که سیستم باید کارائی آنلاین بالائی داشته باشد. زیرا اغلب ارزیابی سیستم با عمل یادگیری بطور همزمان صورت می پذیرد.

یادگیری برخط:

تفاوت عمده بین روش‌های یادگیری برخط و برون‌خط[1] در نوع ورودی این روش‌ها یعنی ورودی داده ترتیبی در مقابل داده ورودی یکجا است. روش‌های معمول یادگیری ماشین مانند SVM، Bayes و ... روش‌هایی برون‌خط هستن زیرا داده ورودی آن‌ها به صورت یکجا در یک مرحله و یا به صورت دسته‌ای مهیا است. ولی در روش‌های برخط داده ممکن است هنوز تولید نشده باشد مانند پیش‌بینی قیمت سهام. در این روش‌های یادگیری ماشین، الگوریتم‌ها به صورتی هستند که با هر ورودی ترتیبی جدید، یک عملیات یادگیری انجام خواهد شد.

-۲

اولا به طور قطع نمی‌توان گفت که بیش‌برازش اتفاق افتاده است یا خیر، زیرا بیش‌برازش به داده‌های آموزشی وابسته هست و مطمئنا تمام داده‌ها برای یک مسئله خاص قابل دسترس نیست.

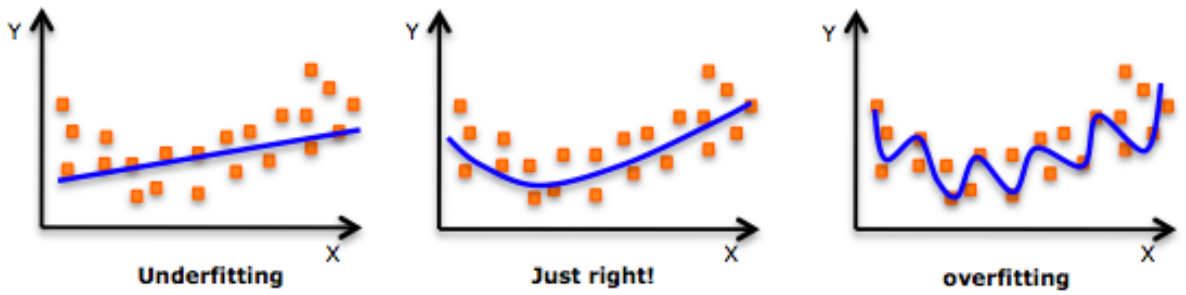
بیش‌برازش زمانی اتفاق می‌افتد که مدل به دست آمده سعی کند که پیچیدگی موجود در داده آموزشی را دنبال کند، در این صورت فرایند یادگیری با حفظ کردن داده‌ها جایگزین خواهد شد. بنابرین چنین مدلی نمی‌تواند داده‌های آزمایشی را به خوبی تخمین بزند زیرا فرایند یادگیری آن به درستی انجام نشده است. همچنین بیش‌برازش به این صورت است که در حین آموزش یک الگوریتم یادگیری تا یک زمانی خطای داده آموزشی و آزمایشی کاهش

---

[1] Offline

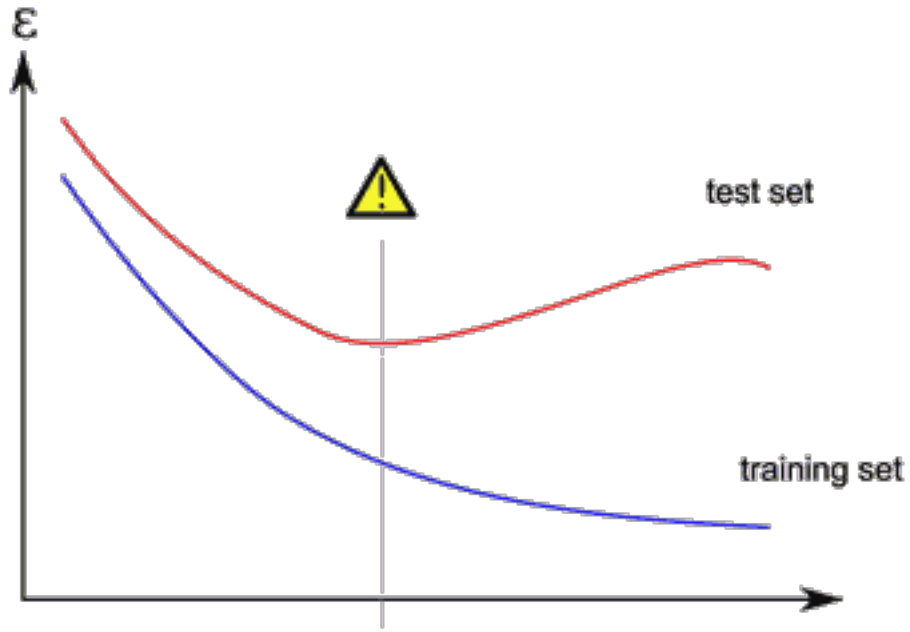
می‌یابد، ولی از یک نقطه‌ای به بعد با کاهش خطای آموزشی، خطای آزمایش افزایش پیدا می‌کند و بیش‌برازش اتفاق می‌افتد. بنابرین در جواب به این سوال می‌توان گفت که در دو حالت ممکن است بیش‌برازش اتفاق افتاده است:

۱- مدل به دست آمده پیچیدگی زیادی دارد و سعی می‌کند پیچیدگی داده را دنبال کند (واریانس بالا مانند شکل یک سمت راست).



شکل ۱.

۲- اگر آموزش به صورت مرحله‌ای باشد، از یک نقطه‌ای به بعد با کاهش خطای آموزشی، خطای آزمایش زیاد شده است (شکل ۲).

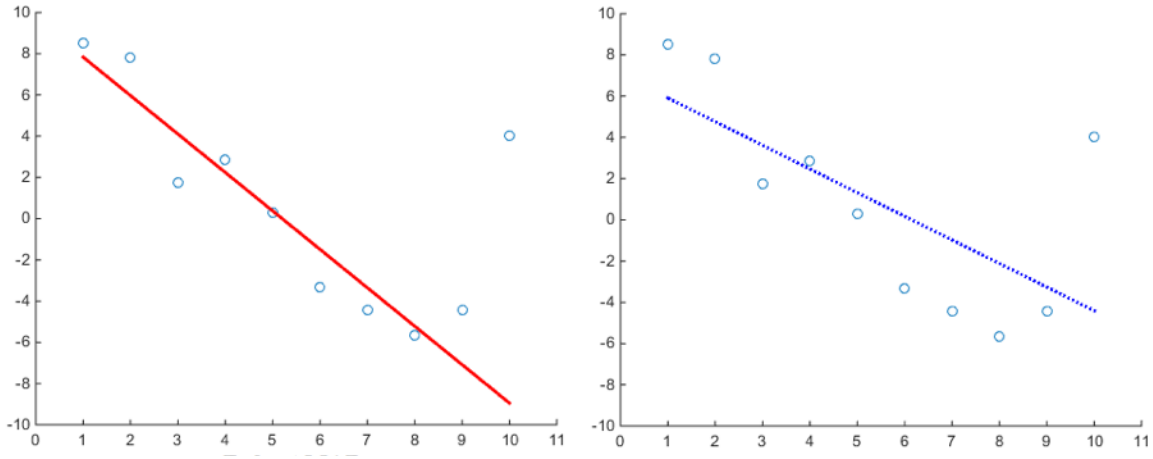


شکل ۲.

۳-

MSE معیاری برای اندازه‌گیری خطای تخمین یک مدل آماری است که به صورت زیر تعریف می‌شود:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$



شکل ۳. سمت چپ RMSE و سمت راست MSE.

MSE مجموع مجذور تمامی خطاهای باقیمانده[۲] مدل آماری است. در معیار MSE اگر داده پرت وجود داشته باشد، چون مقادیر خطا به توان دو می‌رسد، داده‌های پرت مقدار خطا را بسیار بزرگ می‌کنند و مدل به سمت داده نویزی حرکت می‌کند (شکل ۳ سمت راست). برای مثال فرض کنید خطاهای باقیمانده یک مدل آماری به صورت ۱.۵، ۰.۵، ۱.۱، ۰.۴ و ۱۰ باشد. در نهایت مقدار MSE برابر ۱۰.۳۷۷ می‌شود و خطای باقیمانده داده پرت (۱۰) تاثیری بسیار بیشتری بر روی این معیار خواهد داشت. حال این‌که داده پرت معمولا مهم نیست و یک نوع داده نویزی محسوب می‌شود. بنابرین مدلی بهتر است که در مقابل این داده پرت مقاوم‌تر باشد. در این‌جا RMSE دارای این ویژگی است. زیرا RMSE برابر جذر MSE است و برای داده پرت تاثیر کمتری در نظر می‌گیرد. در مثال بالا مقدار RMSE برابر ۳.۲۲ می‌شود.

۴-

در الگوریتم گرادیان نزولی نوسان زیاد، رسیدن به همگرایی را سخت می‌کند. بنابرین می‌توان از تکنیک اثر تکانه استفاده کرد. در این تکنیک الگوریتم با یک ضریبی در جهت تغییرات قبلی وزن حرکت می‌کند و در نتیجه الگوریتم در مقابل نوسان مقاوم‌تر خواهد بود و زودتر همگرا می‌شود. در این روش رابطه‌ای که برای بروزرسانی پارامترها استفاده می‌شود به صورت زیر است که در آن میو ضریب تکانه است.

$$\Delta W_k(i) = -\alpha \frac{\partial E}{\partial W_k} + \mu \Delta W_k(i-1)$$

مزیت استفاده از این تکنیک به صورت زیر است:

- **باعث همگرایی زودتر و پایدارتری می‌شود.**
- **نوسان را کاهش می‌دهد.**

---

[2] Residual

در عمل تکانه باعث می‌شود که برای نقاطی که گرادیان هم‌جهت تغییرات است سرعت الگوریتم افزایش و برای نقاطی که گرادیانی در جهت مخالفی است سرعت کاهش پیدا کند. به عبارت دیگر حرکت در جهت نمونه‌های مرتبط سریع‌تر خواهد شد.

-۵

مقدار θ_0 ضریب ثابت چندجمله‌ای است. بنابرین تاثیری بر شکل نمودار برازش و مدل به دست آمده نخواهد داشت و فقط نمودار به دست آمده را در فضا در جهت عمودی جابجا خواهد کرد و مقادیر خطای باقیمانده و سایر پارامترها مستقل از این پارمترها هستند. بنابرین برای θ_0های مختلف تابع هزینه تغییر نخواهد کرد. همچنین اگر این ثابت را در تابع هزینه در نظر نگیریم بهتر است زیرا این مقدار در شکل نمودار برازش نهایی تاثیری ندارد و تاثیر لاندا را کم خواهد کرد زیرا ممکن است مقداری بسیار بزرگ یا بسیار کوچک داشته باشد و شکل نمودار برازش را به هم بریزد.

-۶

بیش‌برازش زمانی اتفاق می‌افتد که مدل بیش از حد آموزش ببیند و داده‌های آموزشی را اصطلاحا حفظ کند. در نتیجه تخمین مدل برای داده‌های دیده نشده ممکن است بد باشد. زیرا ممکن است داده دیده نشده با داده‌های آموزشی تفاوت زیادی داشته باشد. از طرفی هر چه داده آموزشی افزایش پیدا کند و به طبع آن دانش مدل نسبت به مسئله افزایش یابد. احتمال دیده شدن داده آزمایشی نامرتبط کمتر می‌شود. به عبارت دیگر بیش‌برازش به دلیل دانش کم مدل نسبت به مسئله اتفاق می‌افتد و هر چه دانش مدل افزایش یابد (افزایش داده آموزشی) احتمال بیش‌برازش کمتر می‌شود.

از طرفی در الگوریتم‌هایی که به صورت ترتیبی عمل می‌کنند از یک زمانی به بعد با کاهش خطای آموزشی، خطای داده ارزیابی افزایش می‌یابد و بیش‌برازش اتفاق می‌افتد، این نقطه می‌تواند جواب بهینه مدل باشد. با افزایش تعداد داده‌های آموزشی می‌توان این زمان را بیش‌تر به تعویق انداخت و به مدل‌های بهینه‌تری رسید.

-۷

- **افزایش تعداد داده‌های آموزشی:** هر چه تعداد داده‌های آموزشی بیشتر شود. دانش مدل بیشتر شده و احتمال بیش‌برازش کاهش پیدا می‌کند.
- **regularization:** لحاظ کردن اندازه پارامترهای آموزشی در تابع هزینه می‌تواند حساسیت مدل را نسبت به تغییرات داده‌ها کمتر کند و مدلی هموارتر ایجاد کند.
- **cross-validation:** انتخاب مدل مناسب می‌تواند از بیش‌برازش یا عدم‌برازش جلوگیری کند و مدلی با بهترین قابلیت عمومیت را نتیجه دهد.
- **توقف به موقع:** در الگوریتم‌هایی که به صورت ترتیبی عمل می‌کنند از یک زمانی به بعد با کاهش خطای آموزشی، خطای داده ارزیابی افزایش می‌یابد و بیش‌برازش اتفاق می‌افتد، این نقطه می‌تواند جواب بهینه مدل باشد.

-۸

در تمامی قسمت‌های این تمرین داده‌های مسئله ابتدا نرمال شدند و سپس به صورت تصادفی ۲۰ درصد آن‌ها به عنوان داده آزمایشی در نظر گرفته شد. همچنین برای اجرای کدها از ipython استفاده شد.

الف)

[فایل اجرایی مربوط به این قسمت main_gd.ipynb]

مدل‌ها و پارامترهای استفاده شده به صورت زیر است:

model3: ord=3, alpha=0.003, iter=60000

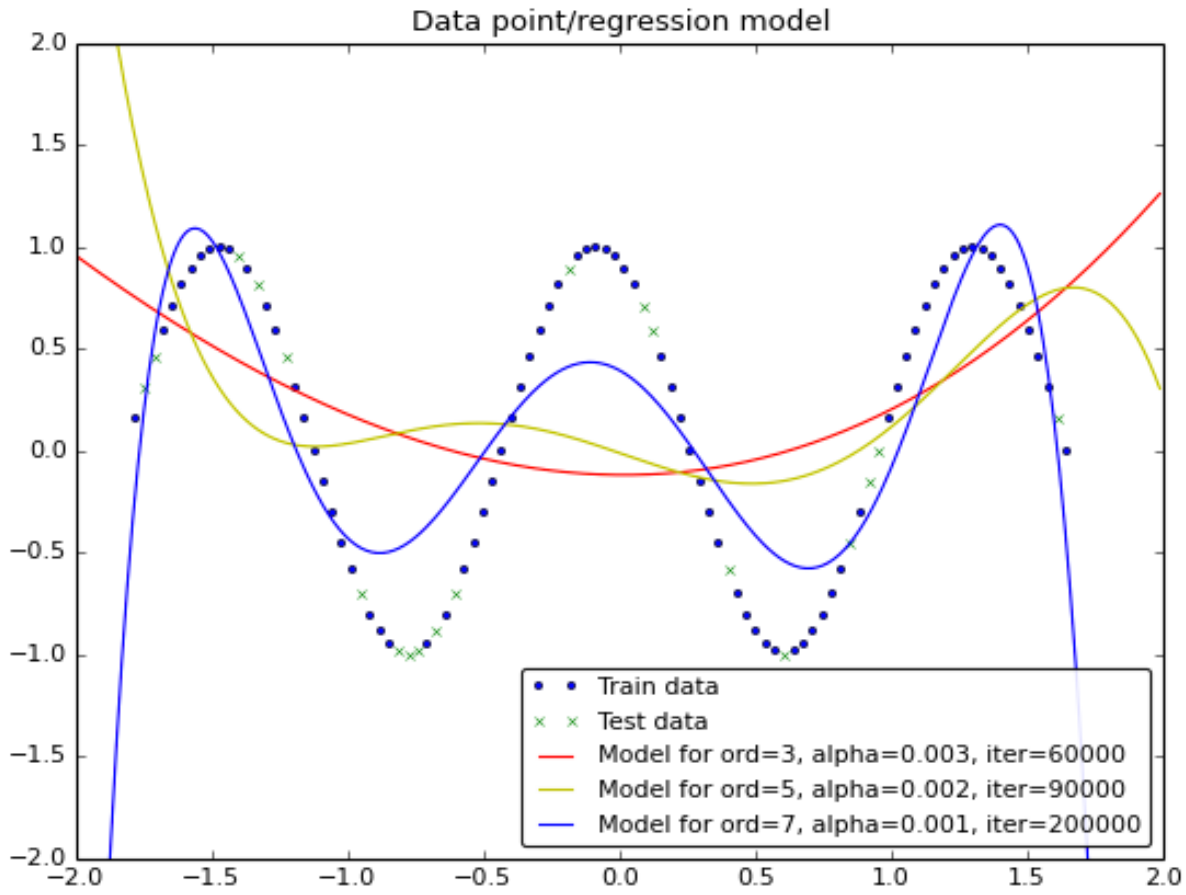
model5: ord=5, alpha=0.002, iter=90000

model7: ord=7, alpha=0.001, iter=200000

که در آن پارامتر alpha بدون استفاده از روش‌های cross-validation و صرفا بر اساس سعی و خطا انتخاب شده است و همچنین تعداد قدم‌ها نیز به صورتی انتخاب شده است که مدل تا مرز بیش‌برازش شدن پیش برود.

همچنین هر چه مدل پیچیده‌تر شود مقدار alpha باید کاهش یابد تا همه پارامترها به درستی آموزش داده شوند.

نمودار داده‌ها و منحنی‌های برازش در شکل زیر آمده است:



همان‌طور که مشاهده شود هر چه مدل پیچیده‌تر شده است، بهتر پیچیدگی موجود در داده را دنبال کرده است. همچنین مدل مرتبه ۳، به خوبی برازش نشده است و دارای پیچیدگی پایینی هست.

معیار MSE برای هر سه مدل به صورت زیر محاسبه شده است:

Train MSE for model3: ord=3, alpha=0.003, iter=60000 is 0.195102132515

Test MSE for model3: ord=3, alpha=0.003, iter=60000 is 0.257593254357

Train MSE for model5: ord=5, alpha=0.002, iter=90000 is 0.180988131464

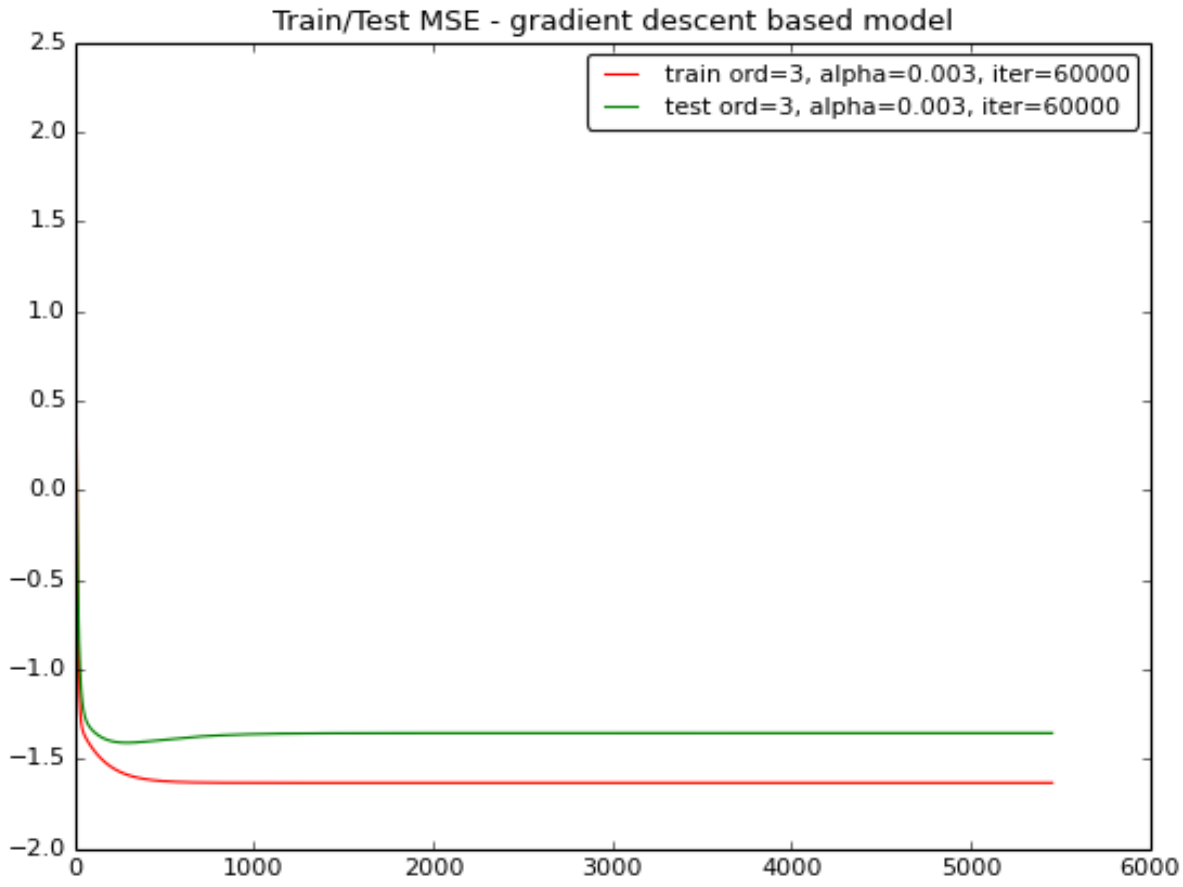
Test MSE for model5: ord=5, alpha=0.002, iter=90000 is 0.244488389941

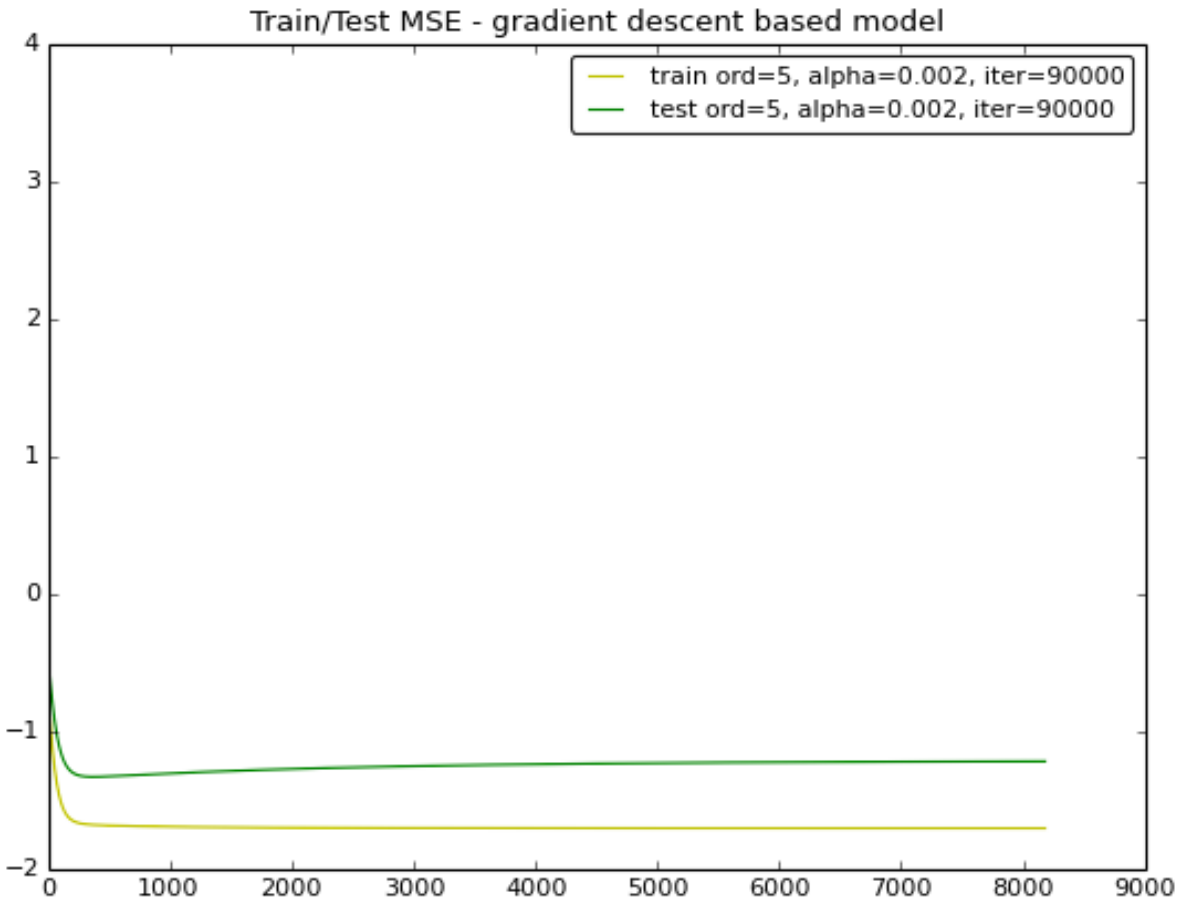
Train MSE for model7: ord=7, alpha=0.001, iter=200000 is 0.0497836994358

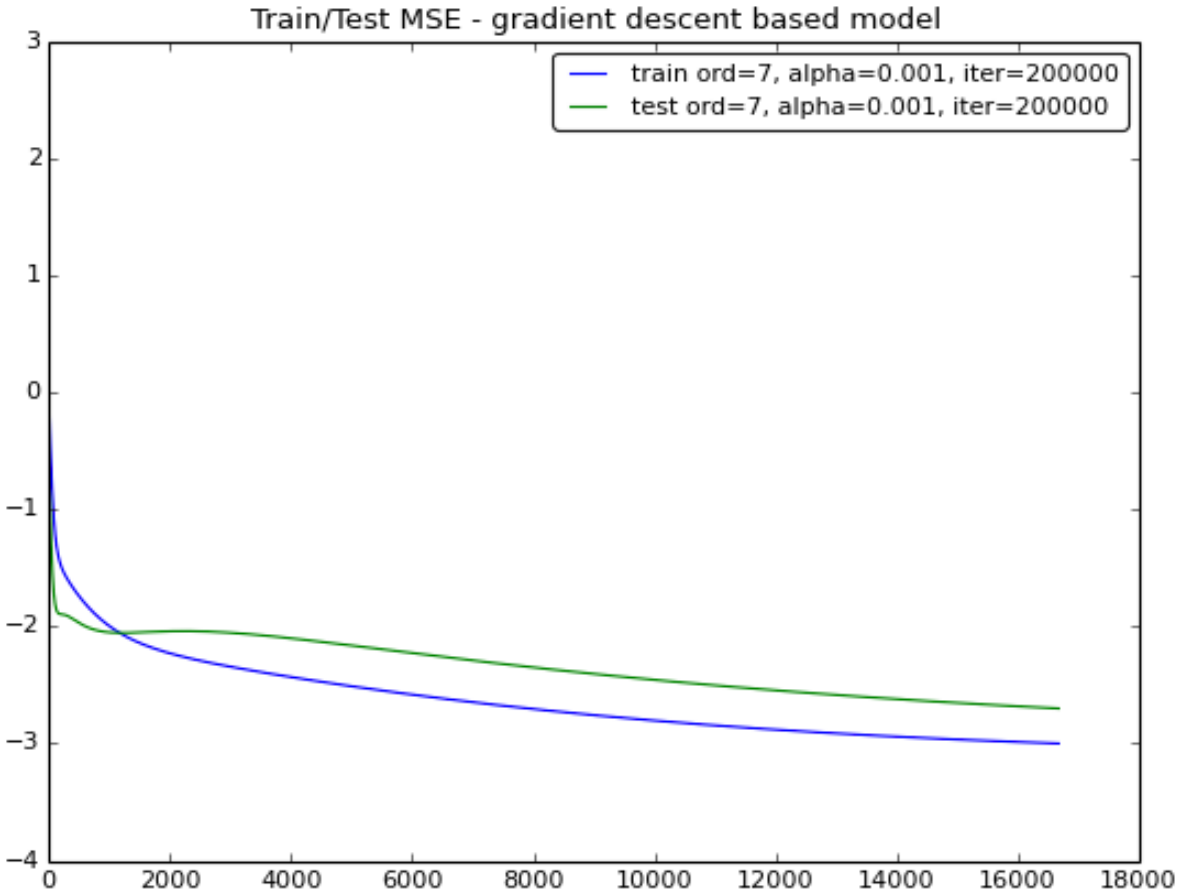
Test MSE for model7: ord=7, alpha=0.001, iter=200000 is 0.0671271417709

با توجه به مقادیر MSE مشاهده می‌شود که هر چه مدل پیچیده‌تر شده است خطای آموزشی و آزمایشی نیز کاهش یافته است.

نمودارهای خطای داده آموزشی و آزمایشی برای هر سه مدل مرتبه ۳، ۵ و ۷ به ترتیب به صورت زیر به دست آورده شد:




Train/Test MSE - gradient descent based model
train ord=5, alpha=0.002, iter=90000
test ord=5, alpha=0.002, iter=90000
4
3
2
1
0
−1
−2
0
1000
2000
3000
4000
5000
6000
7000
8000
9000


Train/Test MSE - gradient descent based model
train ord=7, alpha=0.001, iter=200000
test ord=7, alpha=0.001, iter=200000
3
2
1
0
−1
−2
−3
−4
0
2000
4000
6000
8000
10000
12000
14000
16000
18000

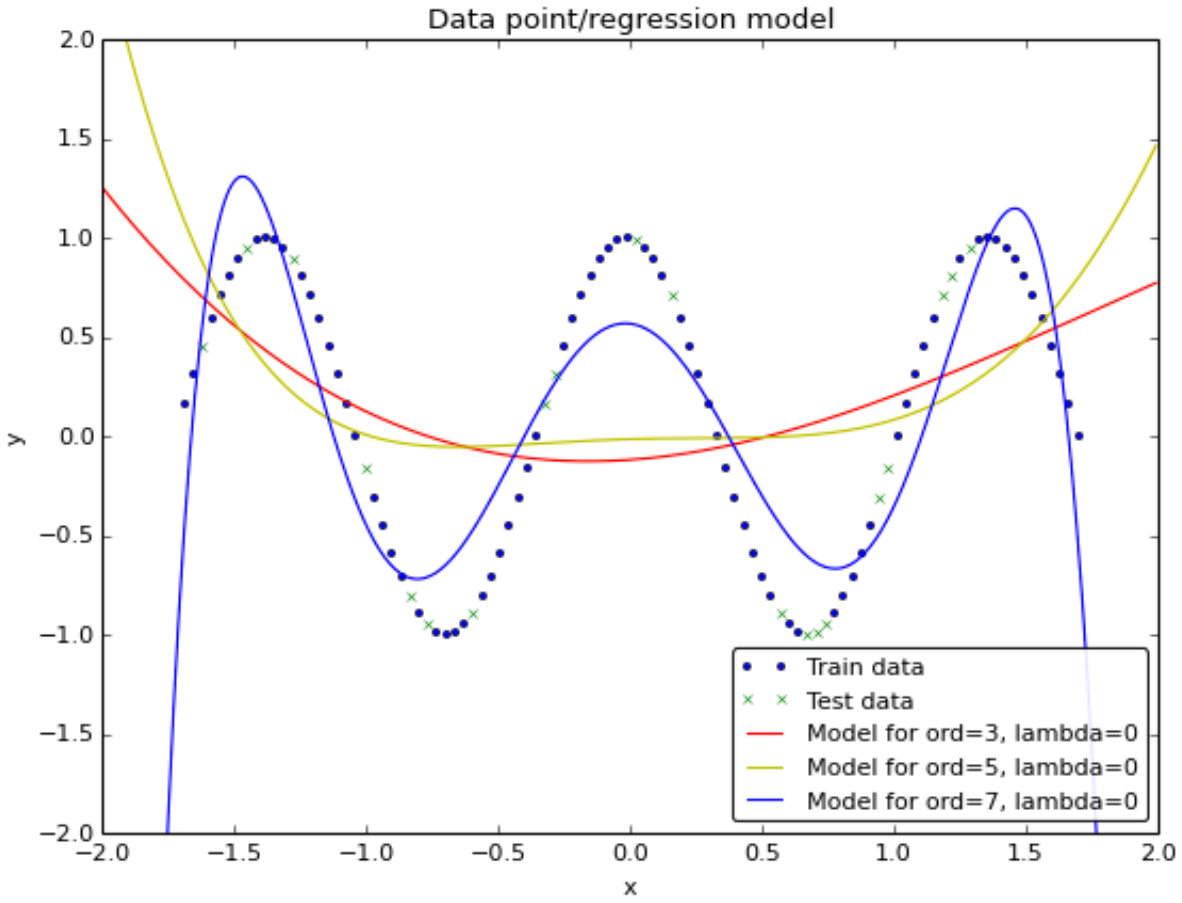
با توجه به نمودار اول خطای اول و دوم، دو مدل اول دچار بیش‌برازش شده اند، زیرا از یک نقطه‌ای به بعد خطای داده آزمایشی افزایش یافته است. ولی مدل مرتبه ۷ با تعداد بیشتری قدم دچار بیش‌برازش نشده است و به نظر میرسد برای این نوع داده آزمایشی از دو مدل دیگر بهتر عمل کرده است.

ب)

[فایل اجرایی مربوط به این قسمت main_closed_form.ipynb]

مدل‌های استفاده شده در این بخش به ترتیب دارای مرتبه ۳، ۵ و ۷ می‌باشند.

نمودارهای برازش به همراه داده‌های مسئله در نمودار زیر آورده شده است.



معیار MSE برای هر سه مدل به صورت زیر محاسبه شده است:

Train MSE for model3: ord=3, lambda=0 is 0.193223513617

Test MSE for model3: ord=3, lambda=0 is 0.257712784999

Train MSE for model5: ord=5, lambda=0 is 0.18844985886

Test MSE for model5: ord=5, lambda=0 is 0.242239918096

Train MSE for model7: ord=7, lambda=0 is 0.0458918677049

Test MSE for model7: ord=7, lambda=0 is 0.0425955414671

با توجه به مقادیر MSE مشاهده شده، مدل‌های بدست آمده مانند بخش قبلی این تمرین به دست آمده شدند. همچنین پارامتر لامبدا صفر در نظر گرفته شده است و با توجه به نمودارهای برازش، مانند قسمت قبل برای دو مدل اول بیش‌برازش اتفاق افتاده است.

ج)

[فایل اجرایی مربوط به این قسمت **main_closed_form_lambda.ipynb**]

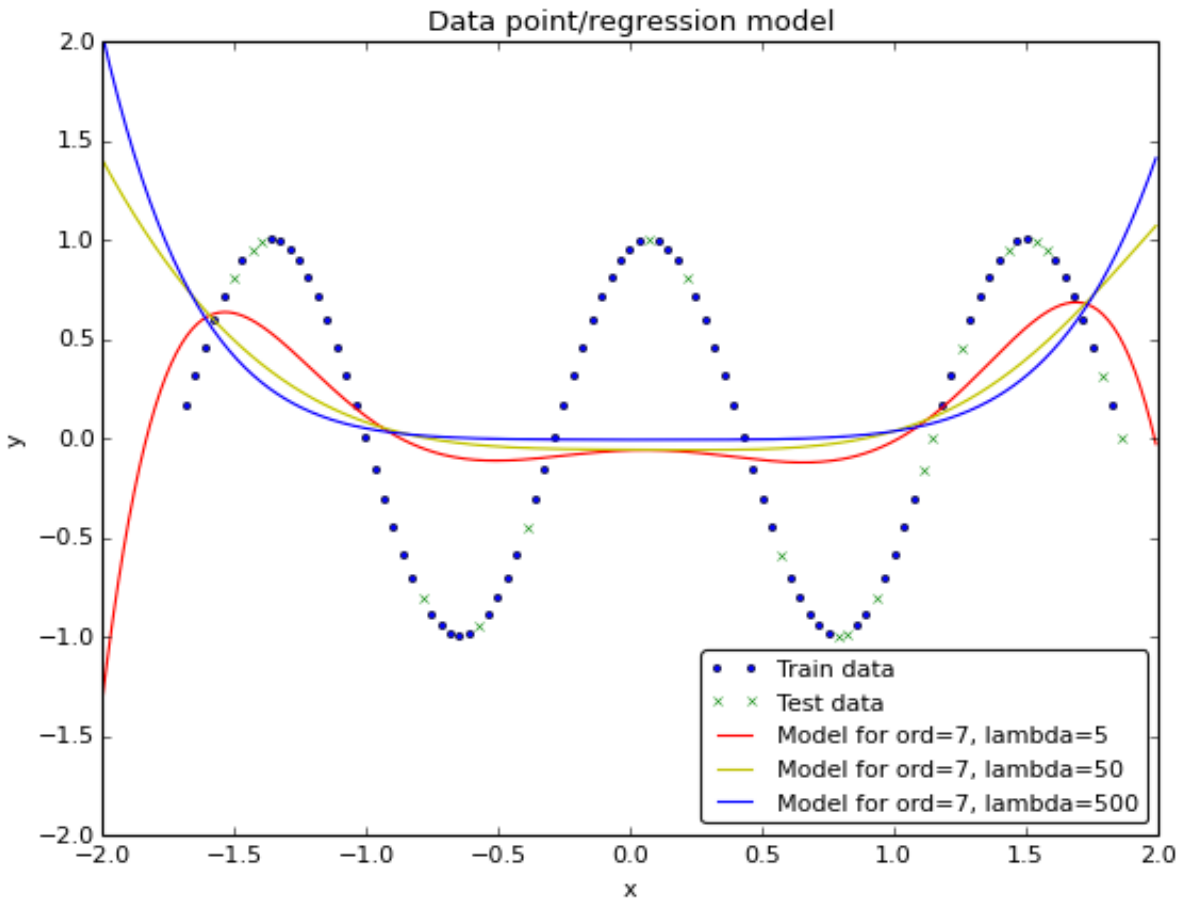
مدل‌های استفاده شده در این قسمت به صورت زیر بودند:

model1: ord=7, lambda=5

model2: ord=7, lambda=50

model3: ord=7, lambda=500

نمودارهای برازش و داده به صورت زیر می‌باشند:



در این قسمت مقدار MSE به صورت زیر محاسبه شد:

Train MSE for model ord=7, lambda=5 is 0.15862876039

Test MSE for model ord=7, lambda=5 is 0.173159674515

Train MSE for model ord=7, lambda=50 is 0.189787215528

Test MSE for model ord=7, lambda=50 is 0.229956341927

Train MSE for model ord=7, lambda=500 is 0.198485285483

Test MSE for model ord=7, lambda=500 is 0.253168286946

همچنین بردار تتا و اندازه آن به صورت زیر مشاهده گردید:

Model1: theta= [-0.06210842 0.03902292 -0.34025495 -0.15959605 0.58361951 0.06103426 -0.1346369 -0.00110254] , |theta|= 0.713505286945

Model2: theta= [-0.0589147 -0.00775498 0.00801579 -0.0221506 0.11717755 0.01068621 -0.00946226 -0.00240541] , |theta|= 0.134261230677

Model3: theta= [-0.00915592 -0.00173743 0.00623322 -0.00310414 0.02289565 0.00010443 0.02122232 -0.00208679] , |theta|= 0.0333811547582

مشاهده می‌شود که با افزایش ضریب لامبدا اندازه بردار تتا کاهش می‌یابد زیرا با افزایش ضریب لامبدا تاثیر اندازه بردار تتا افزایش یافته و در نتیجه اندازه بردار تتا کاهش بیشتری می‌یابد. البته این کاهش اندازه تتا در اینجا تاثیر عکس داشته است و هر چه مقدار آن بیشتر شده است، مقدار خطا افزایش و مدل بیشتر دچار عدم برازش شده است.