

به نام خدا  
دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)  
دانشکده مهندسی کامپیووتر



دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)

تمرین اول درس یادگیری ماشین

گزارش سوال‌های پیاده سازی

استاد درس: دکتر احسان ناظرفرد

دانشجو: فاطمه غلامزاده

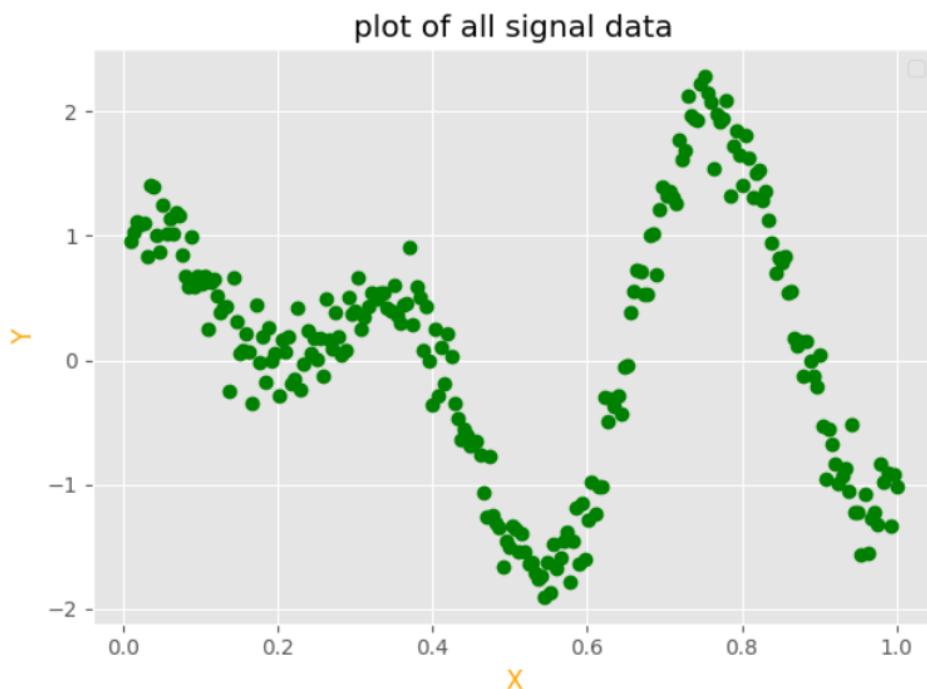
۹۹۱۳۱۰۰۳

نیم سال دوم ۱۴۰۰-۱۳۹۹

## سوال‌های پیاده‌سازی

(مسئله ۱)

الف) رسم داده‌ها :



ب) برای انجام این قسمت نیاز به شافل کردن داده‌ها بود زیرا داده‌ها به صورت مرتب شده بودند و این ویژگی دقیق یادگیری را پایین می‌آورد به همین دلیل در ابتدا داده‌ها را به صورت رندوم جابه‌جا می‌کنیم تا ترتیب خاصی در آن‌ها نباشد.

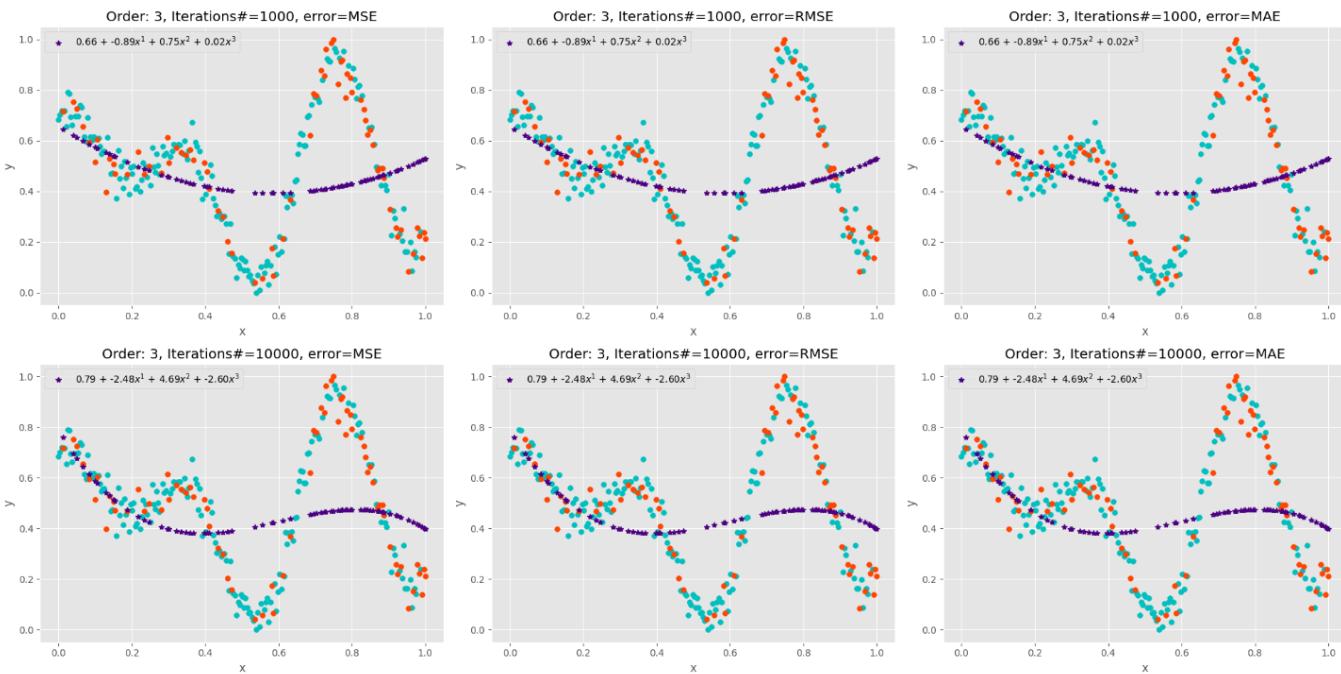
همچنین برای این قسمت داده‌ها بعد از شافل کردن، داده‌ها را نرمال کردم تا در محدوده مشخصی قرار داشته باشند و بازه‌ی خطاهای گزارش شده بهتر مشخص باشد.

در جدول زیر مقدار خطاهای آموزش و آزمون برای حالت‌های مختلف از درجه ۳ گزارش شده است:

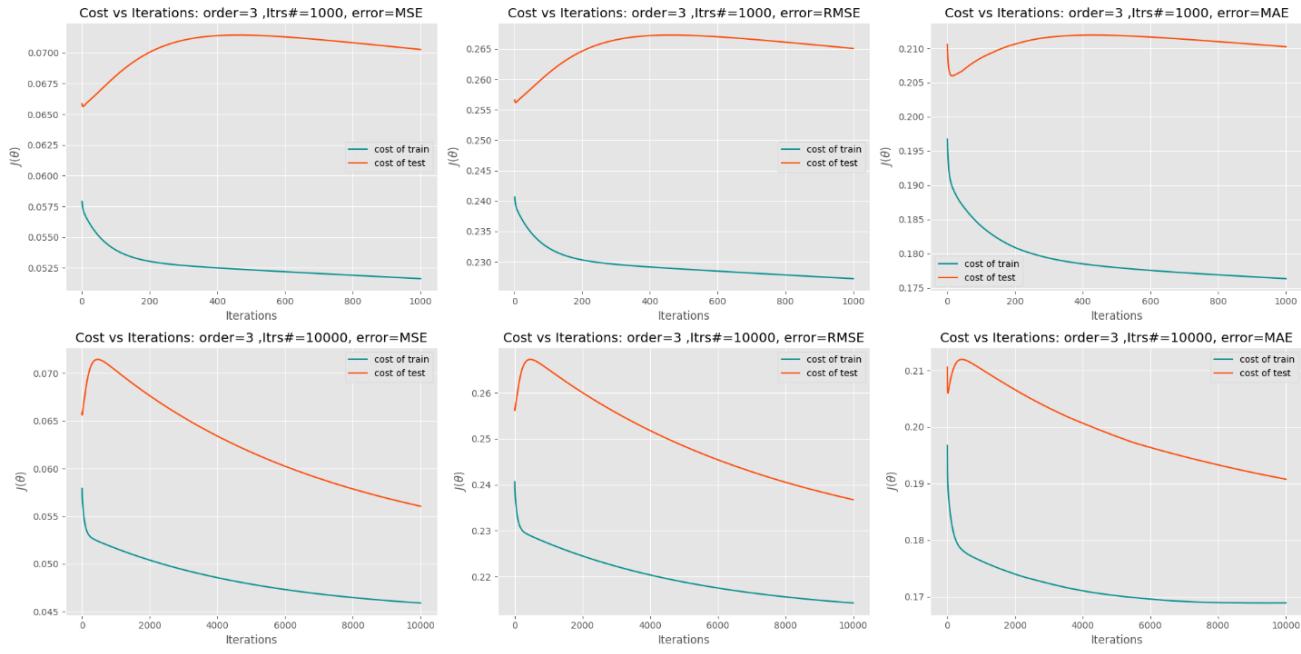
Order= 3	MSE(Train)	MSE(Test)	RMSE(Train)	RMSE(Test)	MAE(Train)	MAE(Test)
1000	0,0516	0,0702	0,2272	0,2650	0,1763	0,2102

10000	0,0459	0,0560	0,2142	0,2367	0,1689	0,1907
-------	--------	--------	--------	--------	--------	--------

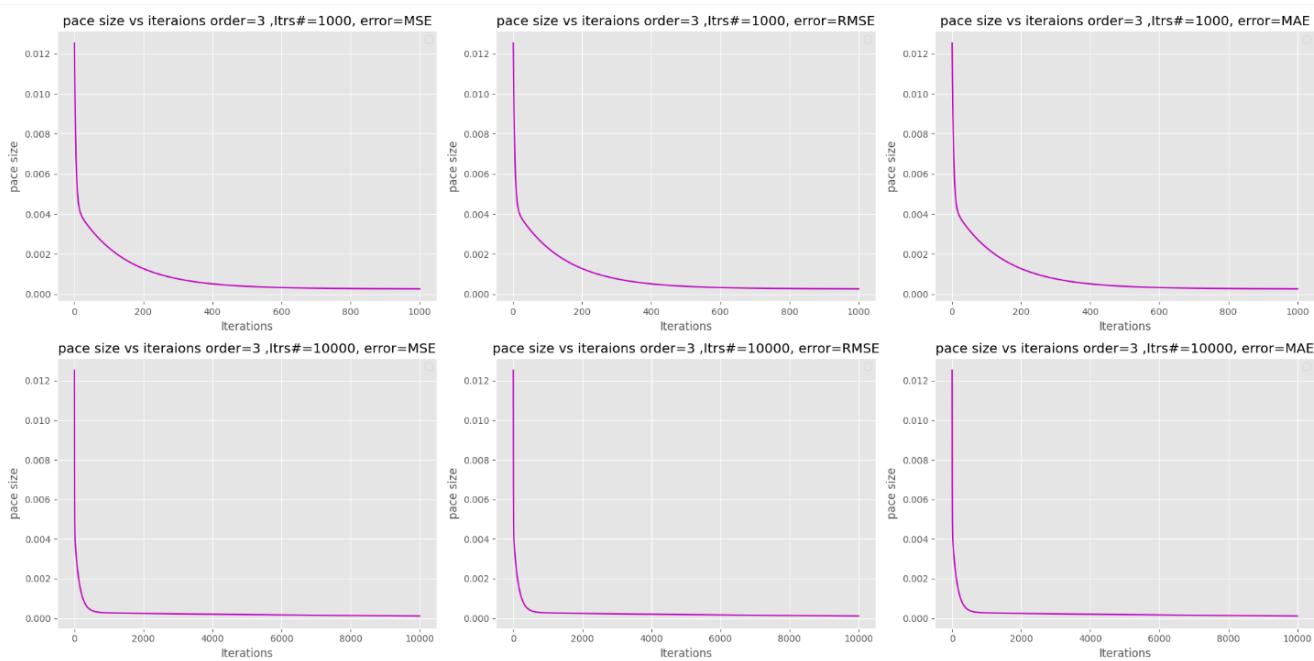
شکل‌های زیر نمودارهای فیت شده بر روی داده‌ها برای حالات مختلف از درجه ۳ را نشان می‌دهد. در این نمودارها نقاط آبی رنگ داده‌های آموزش، نقاط قرمز داده‌های تست و نقاط بنفش مقدار پیش بینی شده برای داده‌های تست می‌باشد.



نمودارهای خطای آموزش و آزمون به ازای تکرارهای مختلف برای درجه ۳ :



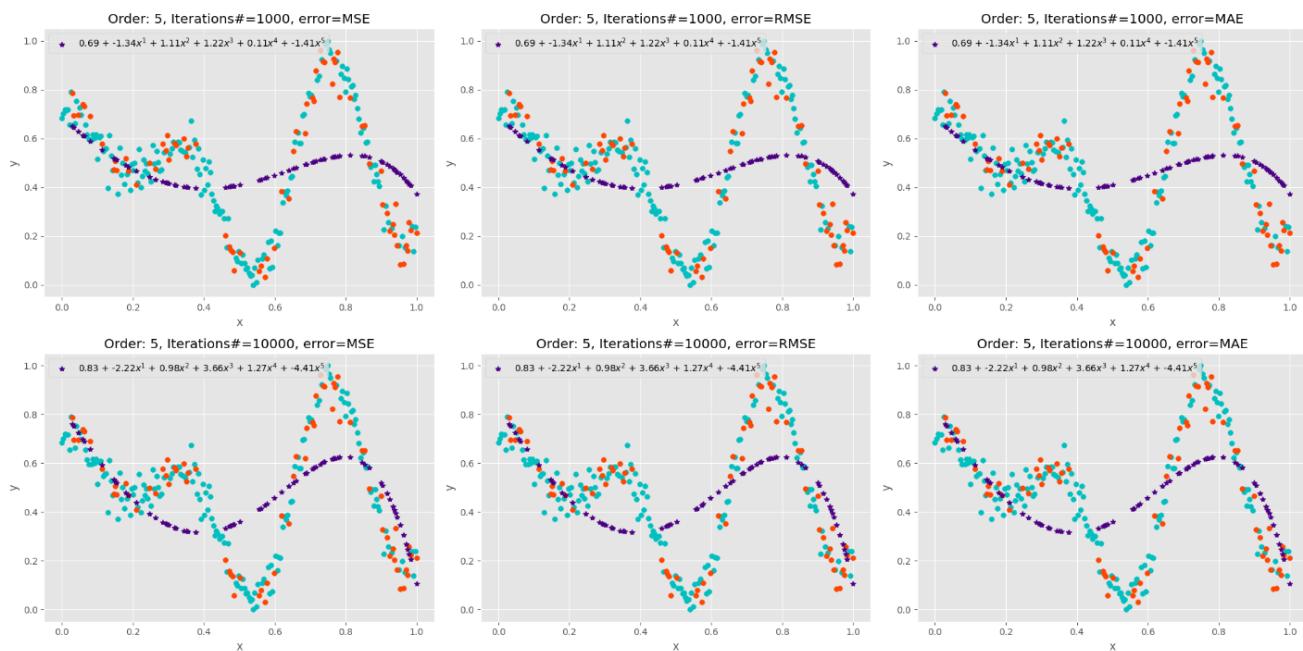
نمودارهای اندازه قدم به ازای تکرارهای مختلف برای درجه ۳ در شکل زیر رسم شده است. لازم به ذکر است که برای رسم تغییرات اندازه قدم یک بُعد از بردار تنا در نظر گرفته شده است.



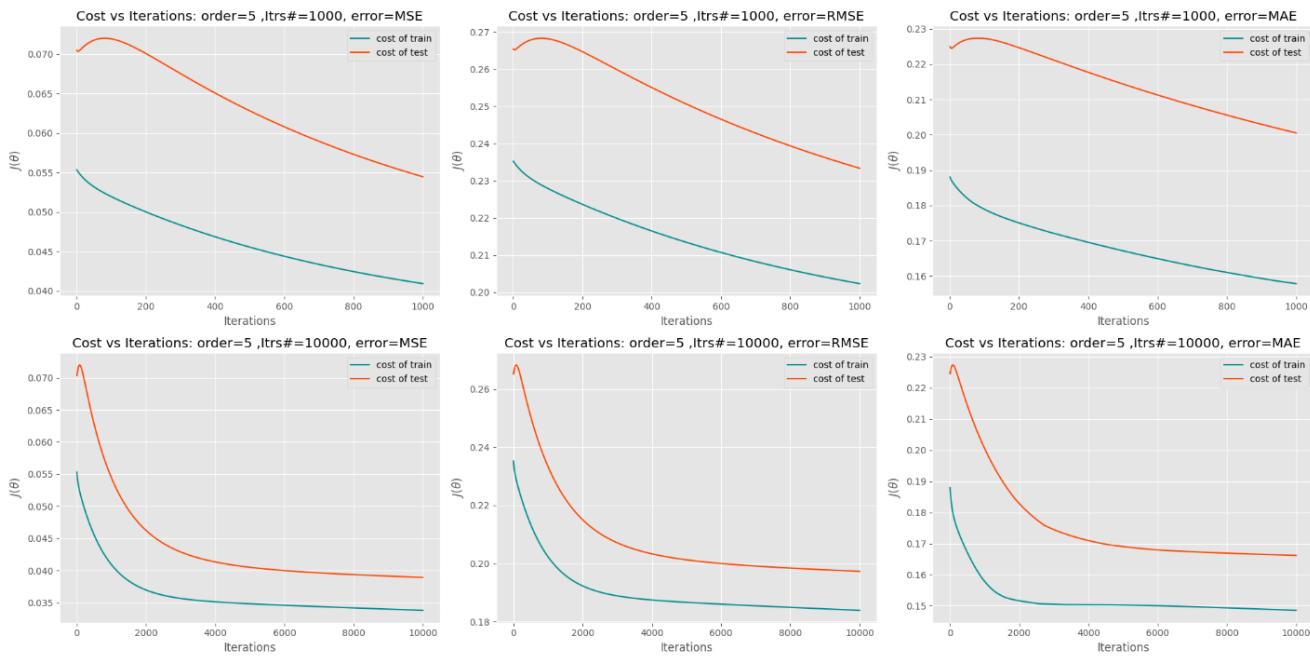
## جدول مقدار خطاهای آموزش و آزمون برای حالت‌های مختلف از درجه ۵ :

Order= 5	MSE(Train)	MSE(Test)	RMSE(Train)	RMSE(Test)	MAE(Train)	MAE(Test)
<b>1000</b>	<b>۰,۰۴۰۹</b>	<b>۰,۰۵۴۴</b>	<b>۰,۲۰۲۳</b>	<b>۰,۲۳۳۳</b>	<b>۰,۱۵۷۸</b>	<b>۰,۲۰۰۵</b>
<b>10000</b>	<b>۰,۰۳۳۸</b>	<b>۰,۰۳۸۹</b>	<b>۰,۱۸۲۸</b>	<b>۰,۱۹۷۳</b>	<b>۰,۱۴۸۵</b>	<b>۰,۱۶۶۲</b>

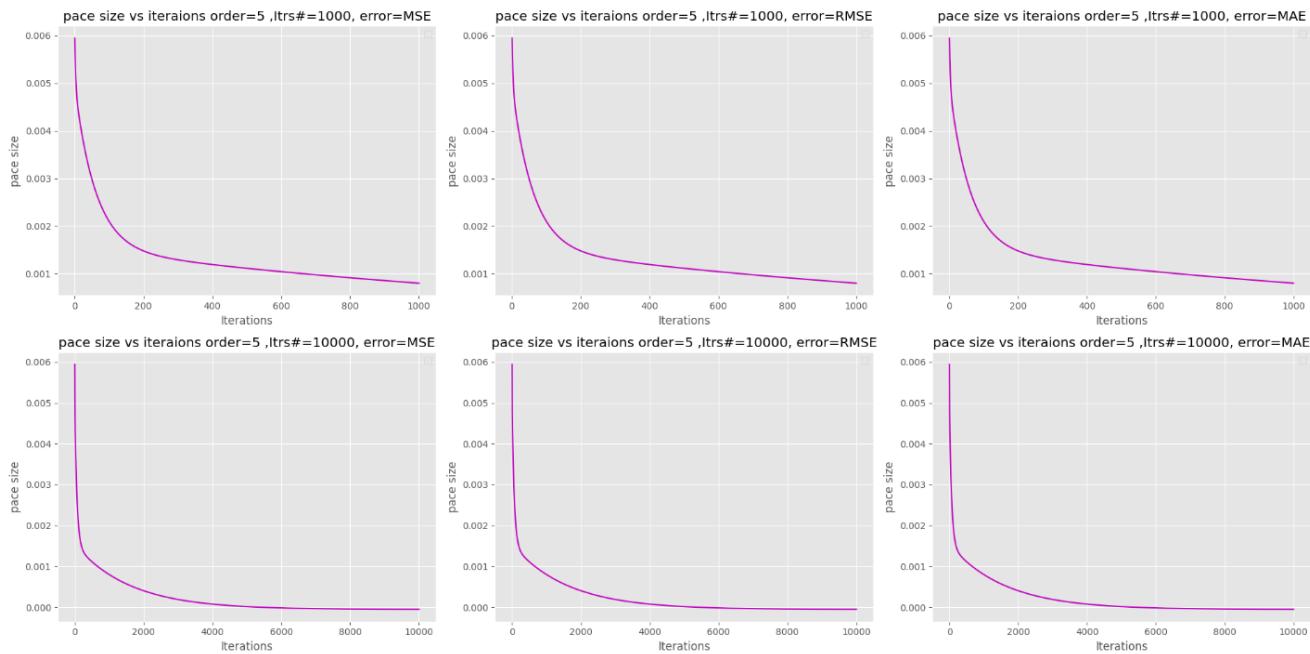
نمودارهای درجه ۵ فیت شده بر روی داده ها:



نمودارهای خطای آموزش و آزمون به ازای تکرارهای مختلف برای درجه ۵ :



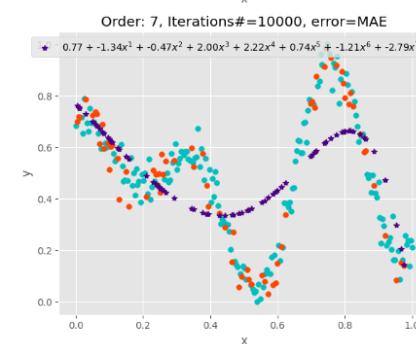
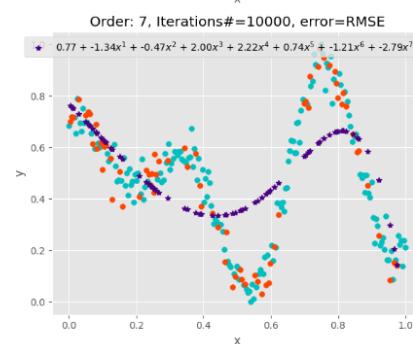
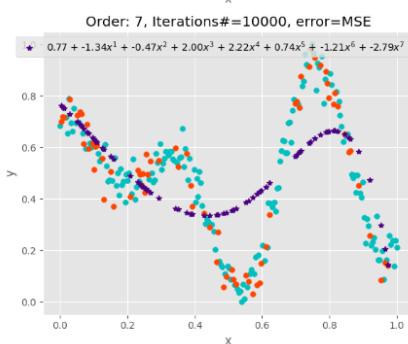
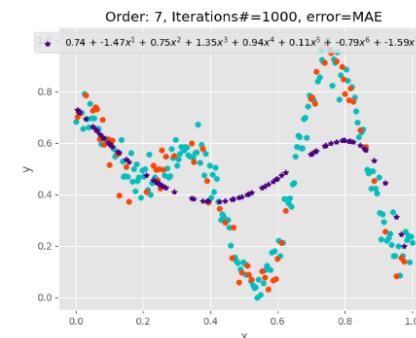
نمودارهای اندازه قدم به ازای تکرارهای مختلف برای درجه ۵



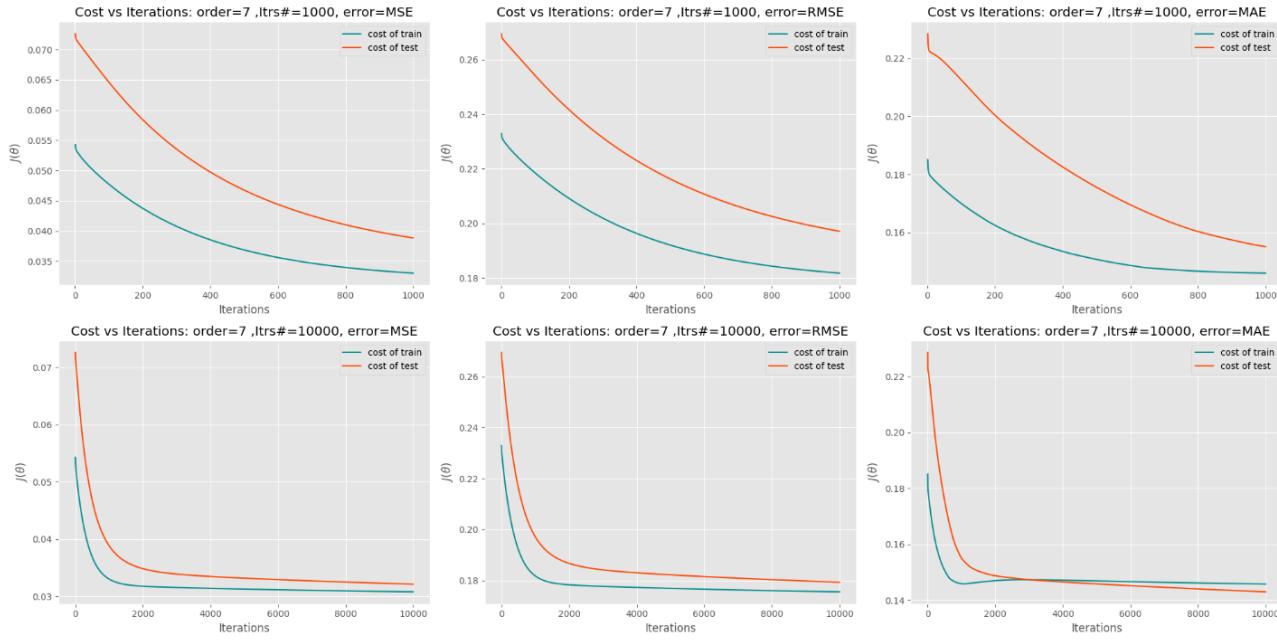
## جدول مقدار خطاهای آموزش و آزمون برای حالت‌های مختلف از درجه ۷:

Order= 7	MSE(Train)	MSE(Test)	RMSE(Train)	RMSE(Test)	MAE(Train)	MAE(Test)
1000	۰,۰۳۳۰	۰,۰۳۸۸	۰,۱۸۱۷	۰,۱۹۷۰	۰,۱۴۵۸	۰,۱۵۵۰
10000	۰,۰۳۰۸	۰,۰۳۲۱	۰,۱۷۵۵	۰,۱۷۹۲	۰,۱۴۵۷	۰,۱۴۲۹

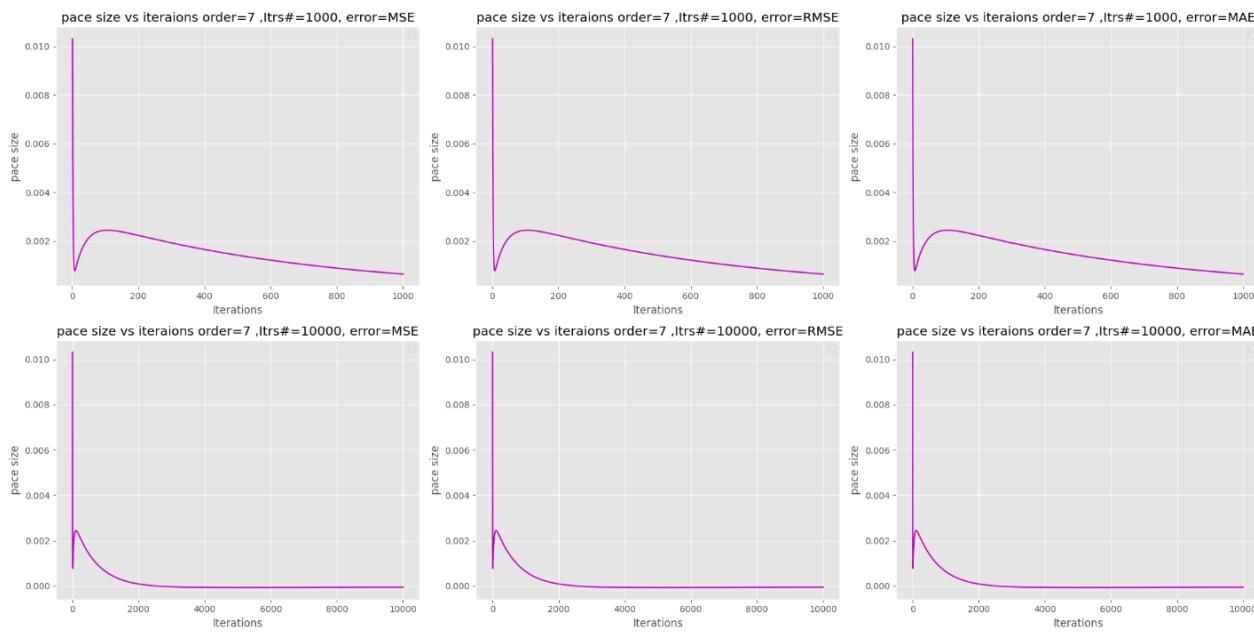
نمودارهای درجه ۷ فیت شده بر روی داده ها:



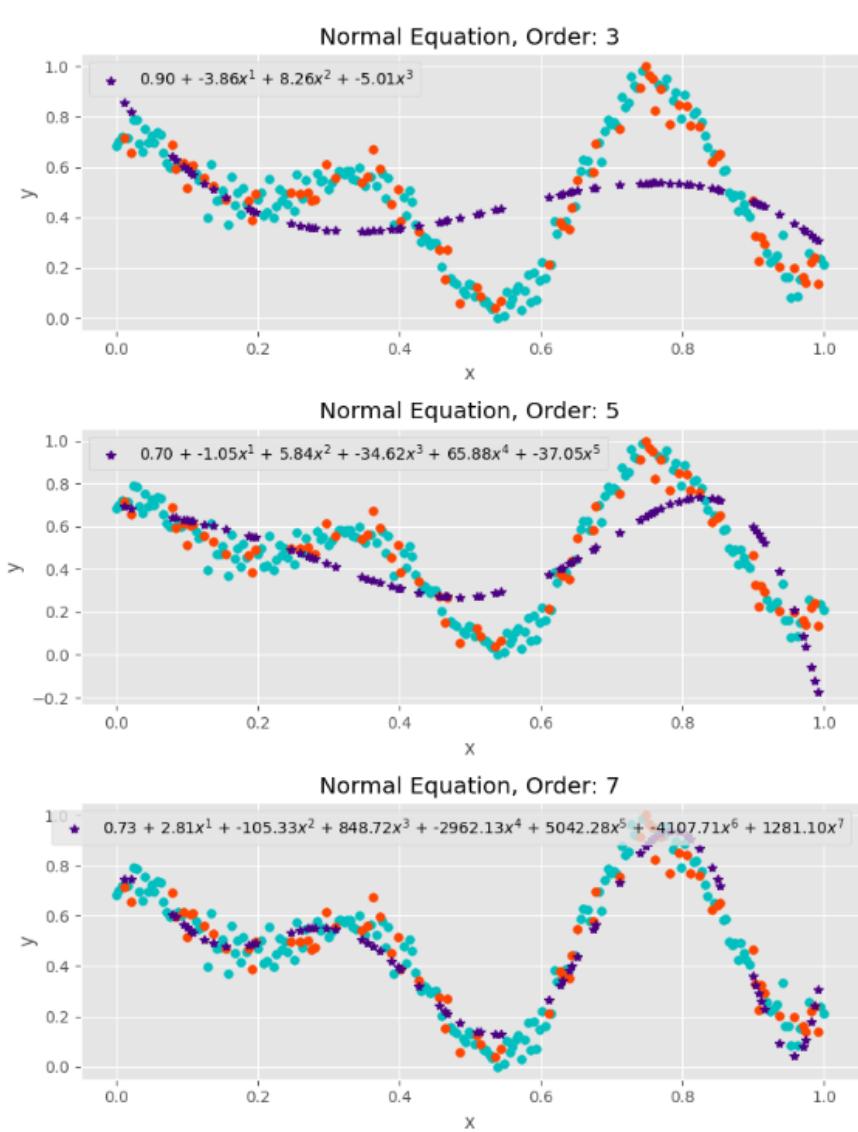
## نمودارهای خطای آموزش و آزمون به ازای تکرارهای مختلف برای درجه ۷ :



## نمودارهای اندازه قدم به ازای تکرارهای مختلف برای درجه ۷ :

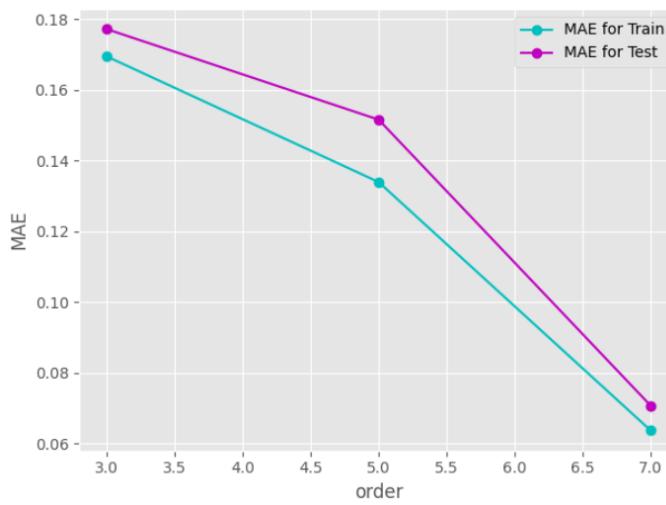
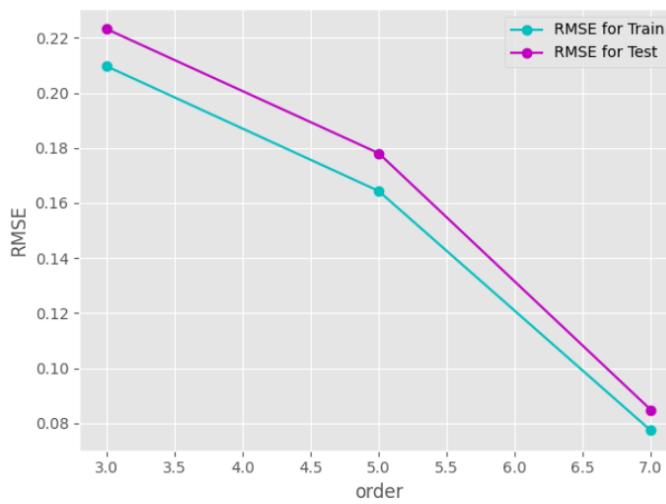
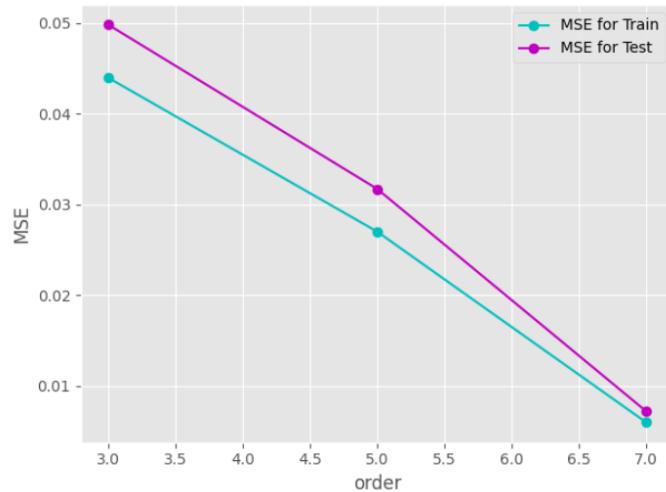


قسمت ج) در این قسمت از معادله نرمال بدون ضریب  $\lambda$  استفاده شده است. نمودارهای فیت شده برای درجه های ۳ و ۵ و ۷ به شکل زیر است:



همانطور که در نمودارها مشاهده می‌شود برای روش معادله نرمال درجه ۷ بیش برآش داریم.

نمودار خطاهای MSE و RMSE و MAE به ازای درجه های ۳ و ۵ و ۷ با استفاده از روش معادله نرمال ۷ بدون ضریب  $\lambda$ :

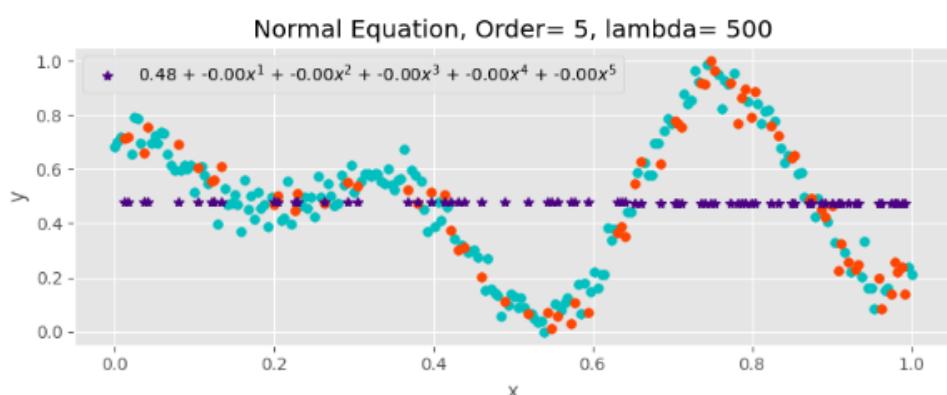
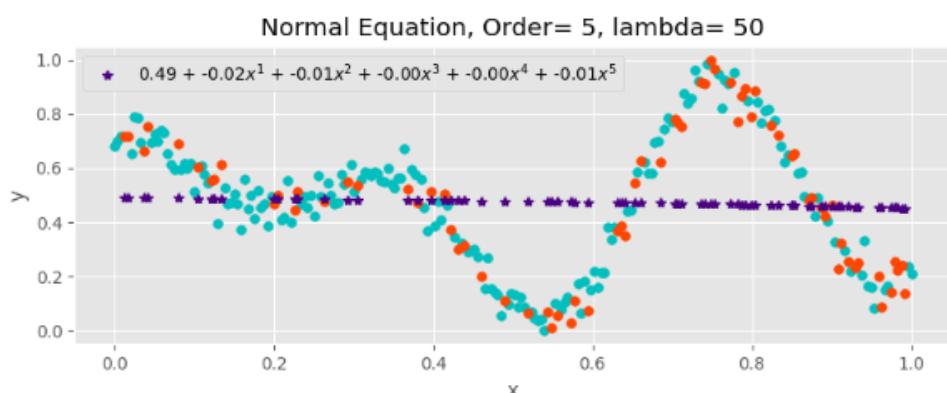
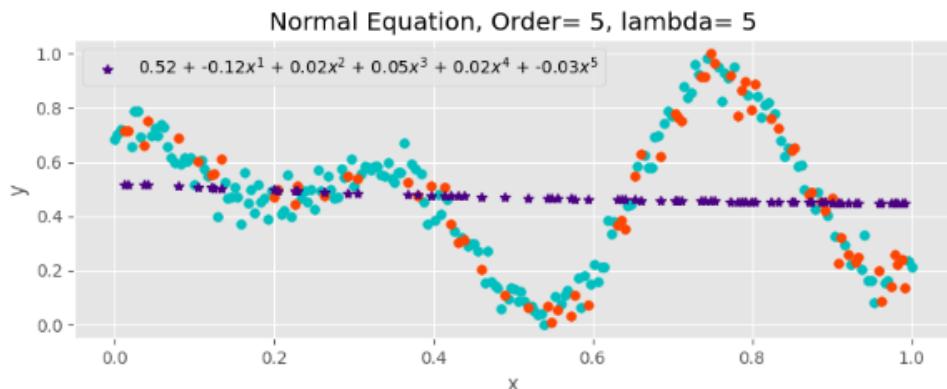


جدول زیر مقادیر خطاهای MSE و RMSE و MAE به ازای درجه های ۳ و ۵ و ۷ با استفاده از روش معادله نرمال بدون ضریب  $\lambda$  را نشان می دهد.

	MSE	RMSE	MAE
Order = 3	۰,۰۴۷۱	۰,۲۱۷۱	۰,۱۷۹۴
Order = 5	۰,۰۲۹۹	۰,۱۷۲۹	۰,۱۴۶۶
Order = 7	۰,۰۰۶۶	۰,۰۸۱۴	۰,۰۶۴۱

**مقایسه با قسمت قبل:** با توجه به خطاهای به دست آمده، روش معادله نرمال برای هر سه درجه خطای کمتری دارد. برای درجه های ۳ و ۵ در روش معادله نرمال مقادیر خطا مقدار کمی نسبت به روش گرادیان نزولی کاهش پیدا کرده است اما برای درجه ۷ روش معادله نرمال کاهش چشمگیری در میزان خطا داشته است که البته این مقدار کاهش خطا نشانه خوبی نیست زیرا همانطور که در نمودارها مشاهده می شود برای درجه ۷ در روش معادله نرمال مدل دچار بیش برازش (overfit) شده و داده ها را حفظ کرده است.

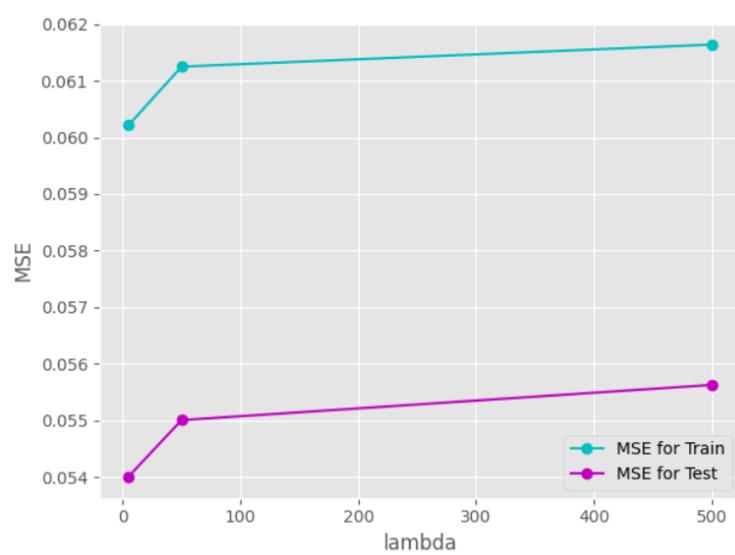
د) نمودار برازش شده بر روی داده ها به ازای درجه ۵ و مقادیر  $\lambda$  برابر ۵ و ۵۰ و ۵۰۰ رسم شده است:



خطای MSE برای داده‌های آموزش و آزمون و برای مقادیر  $\lambda$  مختلف در جدول زیر آورده شده است:

	Lambda = 5	Lambda = 50	Lambda = 500
MSE (Train)	0.0541	0.0552	0.0556
MSE (Test)	0.0686	0.0690	0.0696

نمودار خطای MSE به ازای مقادیر مختلف لامبدا و برای داده‌های آموزش و آزمون :

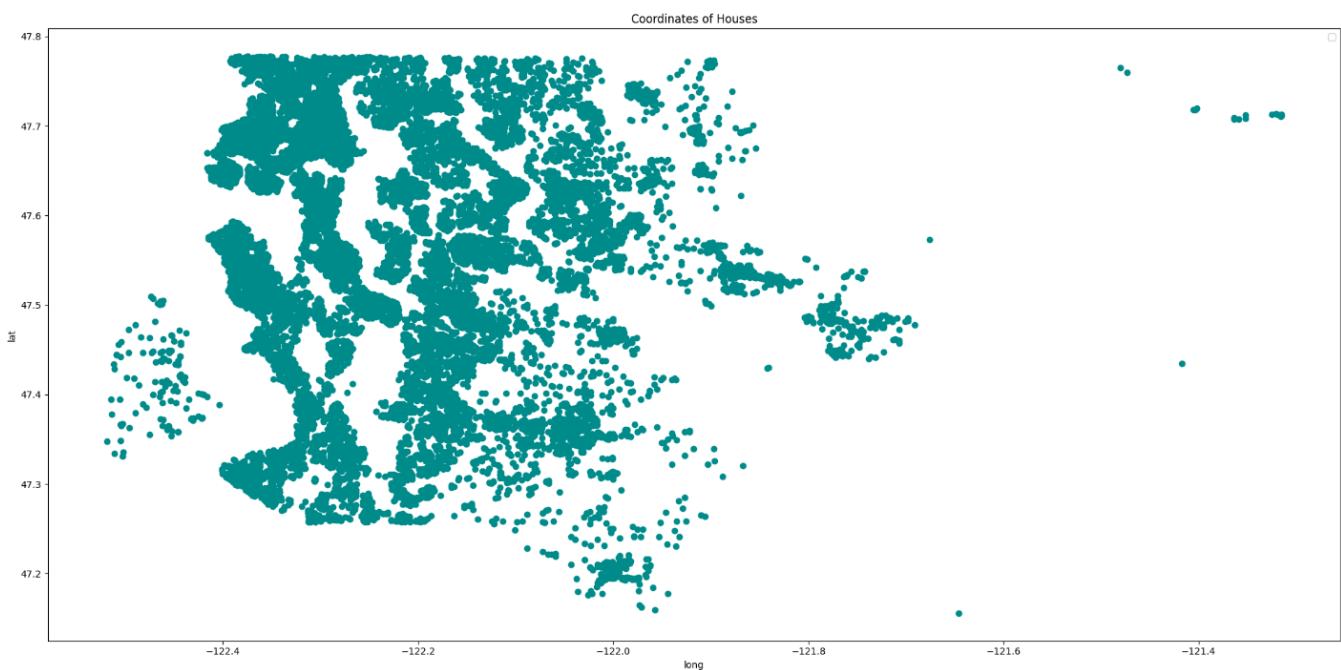


### بررسی تاثیر ضریب $\lambda$ برای بردار ضرایب تنا :

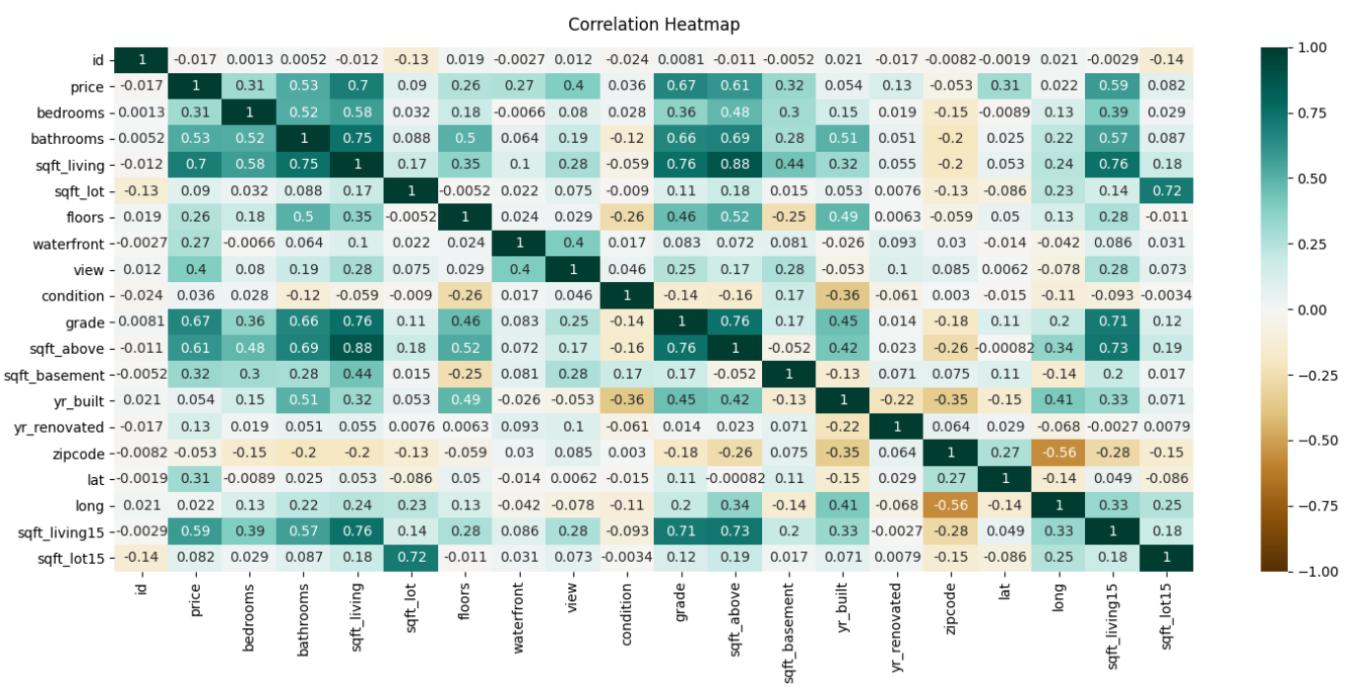
همانطور که از معادله‌ی منحنی رسم شده در نمودارها مشخص است، وقتی روش معادله نرمال را با ضریب  $\lambda$  به کار می‌بریم مقادیر بردار ضرایب تنا کوچک می‌شوند به طوری که برای  $\lambda = 500$  تمامی ضرایب در بردار تنا صفر شده‌اند و فقط ضریب جمله‌ی  $X$  به توان ۰ (یعنی جمله ثابت) صفر نیست یعنی معادله یک خط موازی با محور  $X$  به دست آمده است. علت این امر این است که ضریب  $\lambda$  در واقع سعی دارد که پیچیدگی چندجمله‌ای را کاهش داده و از این طریق مقدار بایاس را افزایش دهد به همین ضرایب جمله‌هایی از چندجمله‌ای که درجه بالاتری دارند را کاهش می‌دهد و هر چقدر مقدار  $\lambda$  بیشتر باشد این کاهش نیز بیشتر است.

### ( ۲ ) مسئله ۲

الف) رسم مختصات داده ها :



ب) نمودار همبستگی بین ویژگی‌ها :



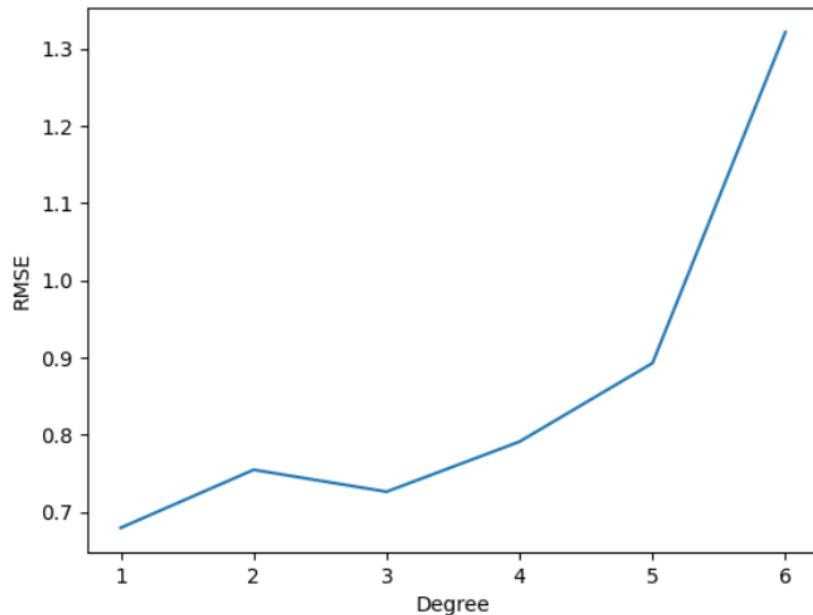
ج) با توجه به نمودار می‌توانیم یک **threshold** در نظر بگیریم و فیچرهایی که در سطر یا ستون آن‌ها مقدار بیشتر از این **threshold** باشد را حذف کنیم. در واقع با این عمل داریم فیچرهایی که همبستگی

زیادی با هم دارند را حذف می کنیم. با در نظر گرفتن مقدار threshold = 0.7 فیچرهایی که باید حذف شوند :

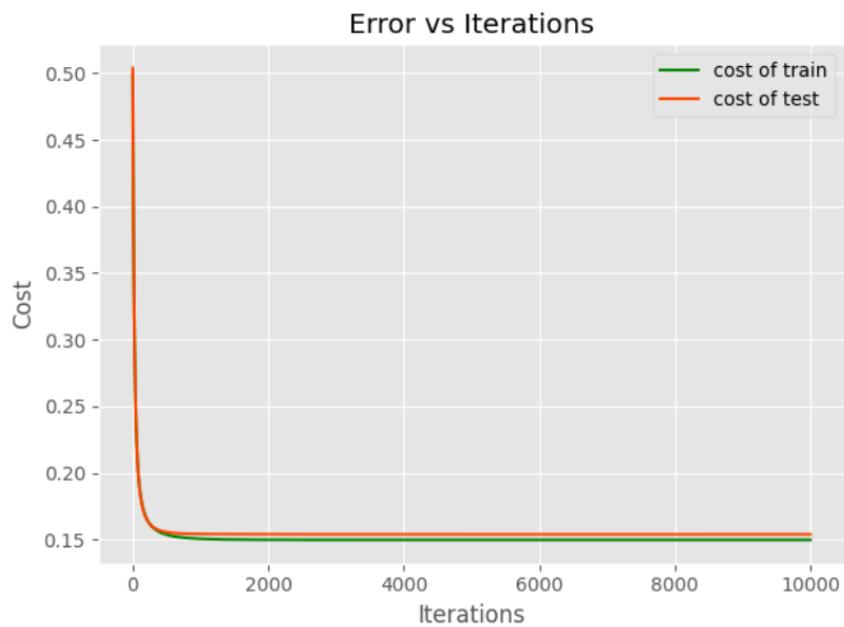
'sqft\_living', 'grade', 'sqft\_above', 'sqft\_living15', 'sqft\_lot15'

(d)

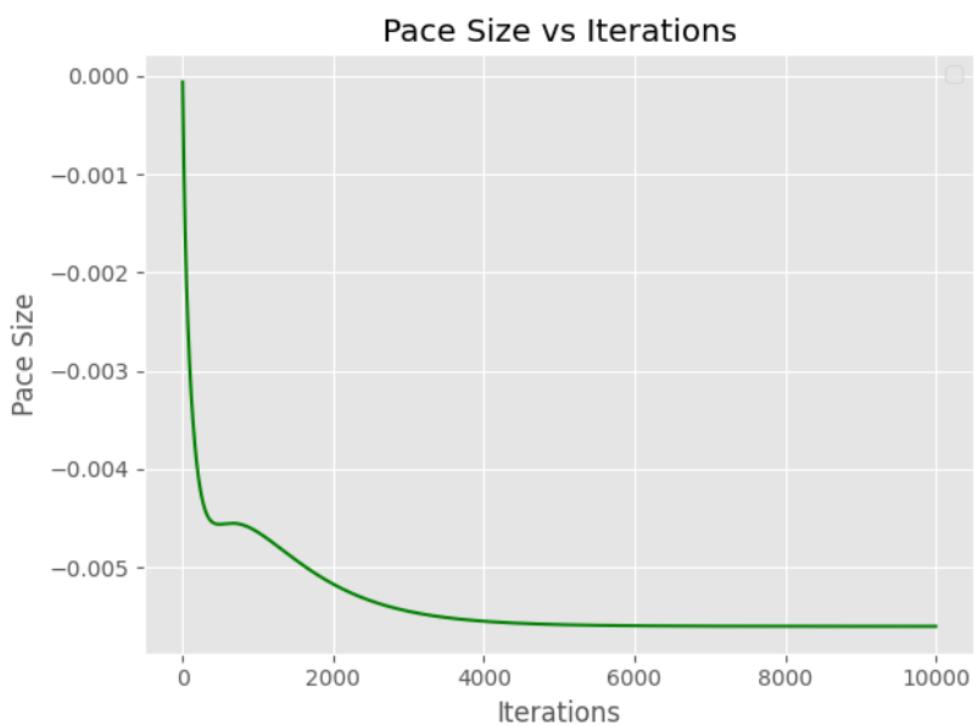
برای بدست آوردن درجه بهینه (d) مقدار خطای RMSE به ازای درجه های مختلف (از 1 تا 6) اندازه گیری شد و نمودار آن در زیر آورده شده است. همانطور که مشاهده می شود میزان خطای برای درجه 1 از سایر درجات کمتر است بنابراین تصمیم گرفته شد که یک منحنی درجه 1 (خط) بر روی داده ها فیت شود.



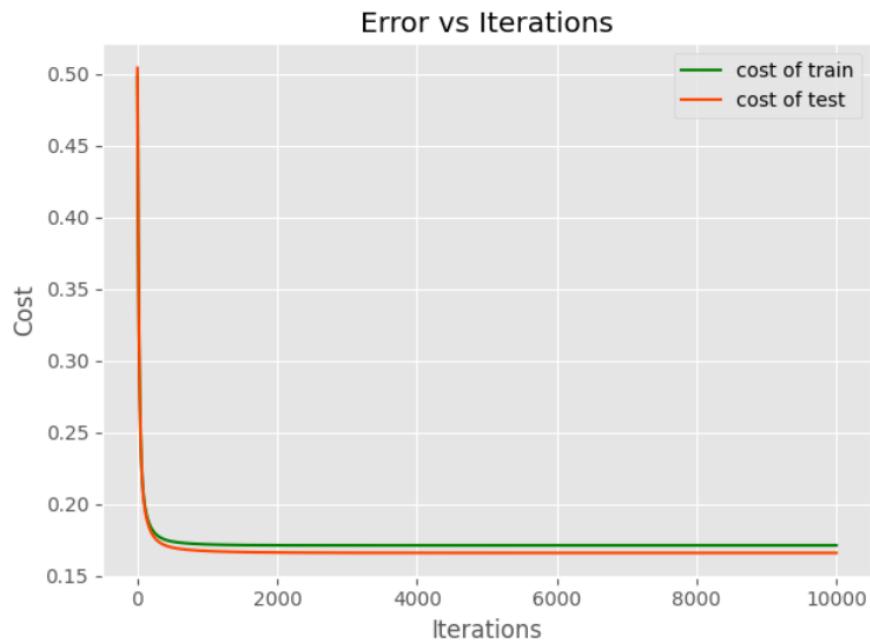
نمودار خطای آموزش و آزمون با استفاده از کل ویژگی ها:



نمودار اندازه قدم با استفاده از کل ویژگی ها :



نمودار خطای آموزش و آزمون با حذف ویژگی های گفته شده:



نمودار اندازه قدم با حذف ویژگی‌ها:

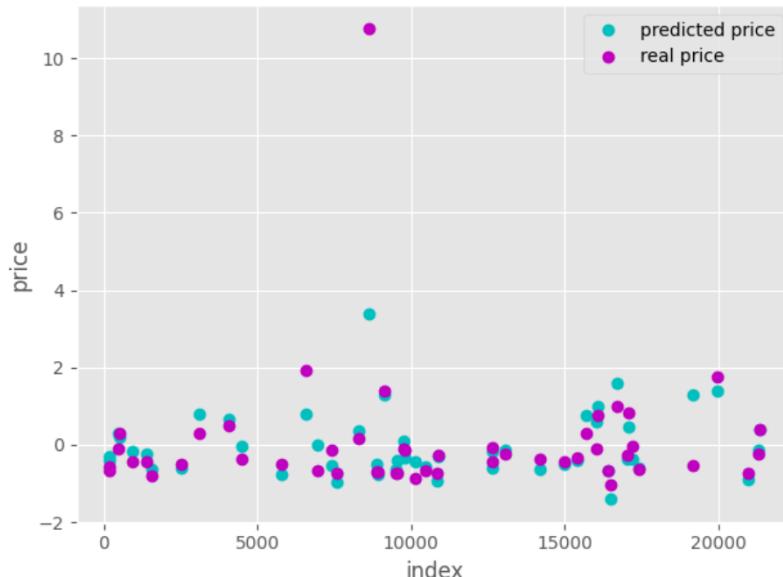


مقدار خطای MSE برای آموزش و آزمون در دو روش در جدول زیر آورده شده :

	With all features	With selected features
MSE (Train)	۰,۱۴۹۷	۰,۱۳۱۰
MSE (Test)	۰,۱۵۳۸	۰,۱۴۵۷

همانطور که مشاهده می شود با استفاده از ویژگی های منتخب، خطای کاهش یافته است.

در نمودار زیر ۵۰ داده تصادفی انتخاب شده و قیمت واقعی و قیمت پیش بینی شده برای آن ها با استفاده از روش گرادیان نزولی رسم شده است. محور افقی ایندکس داده انتخاب شده را نشان می دهد.

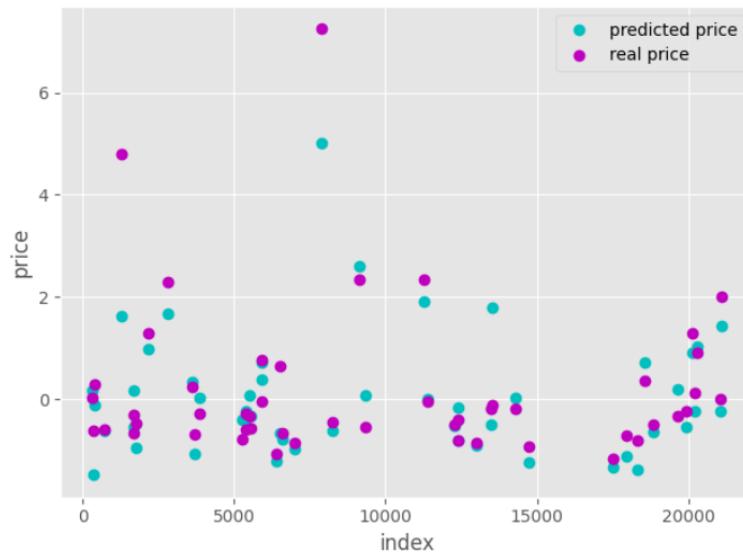


۵) مقدار خطای MSE برای آموزش و آزمون (معادله نرمال) با استفاده از کل ویژگی و استفاده از ویژگی های منتخب در جدول زیر آورده شده :

	With all features	With selected features
MSE (Train)	۰,۱۷۸۵	۰,۱۶۰۵۶
MSE (Test)	۰,۱۷۷۰	۰,۱۶۲۷

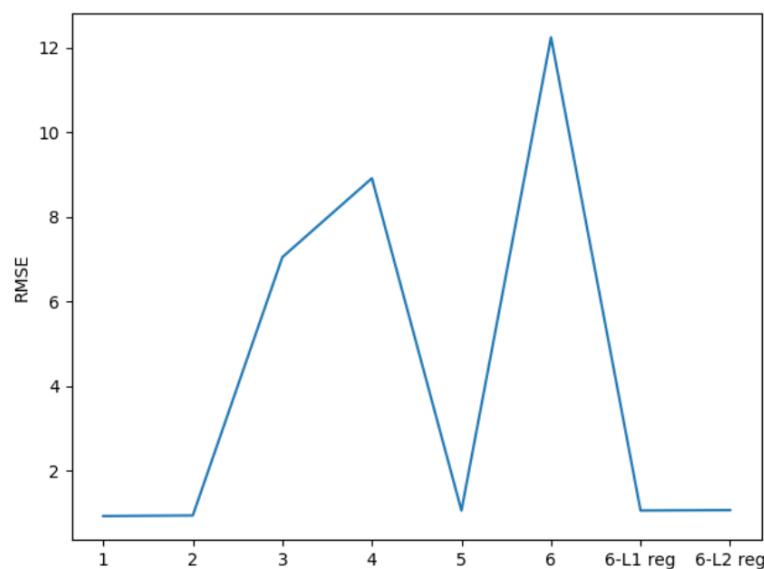
همانطور که مشاهده می شود با استفاده از ویژگی های منتخب، خطای کاهش یافته است.

در نمودار زیر ۵۰ داده تصادفی انتخاب شده و قیمت واقعی و قیمت پیش بینی شده برای آن ها با استفاده از روش معادله نرمال رسم شده است. محور افقی ایندکس داده انتخاب شده را نشان می دهد.



### مسئله (۳)

الف) در این سوال ابتدا برای یافتن درجه بهینه منحنی فیت شده خطای rmse به ازای درجه های ۱ تا ۶ محاسبه شد و خطای برای درجه ۱ مینیمم شد پس یک خط بر روی داده ها فیت می کنیم.

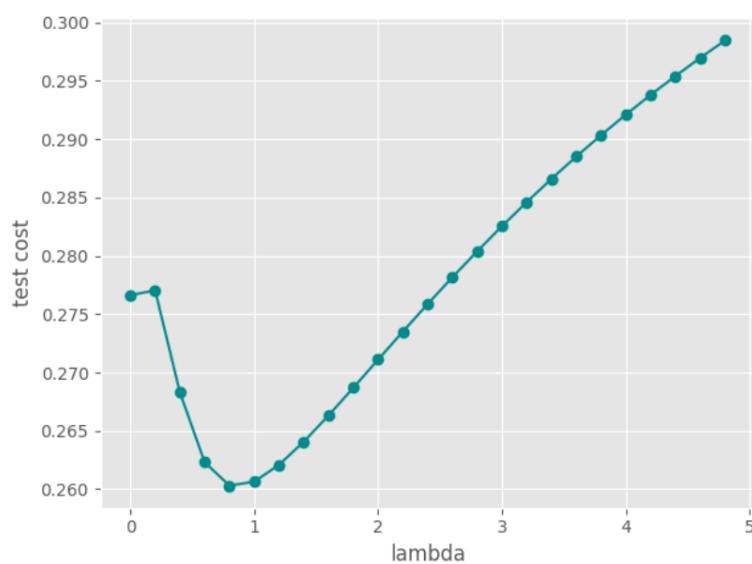


نمودار خطای آموزش و آزمون به ازای گام های مختلف (در هر گام تعداد داده آموزش ۱۰ تا افزایش یافته است) :



همانطور که مشاهده می شود خطای تغییرات زیادی دارد اما به طور کلی روند خطای آموزش رو به افزایش است و روند خطای تست رو به کاهش است.

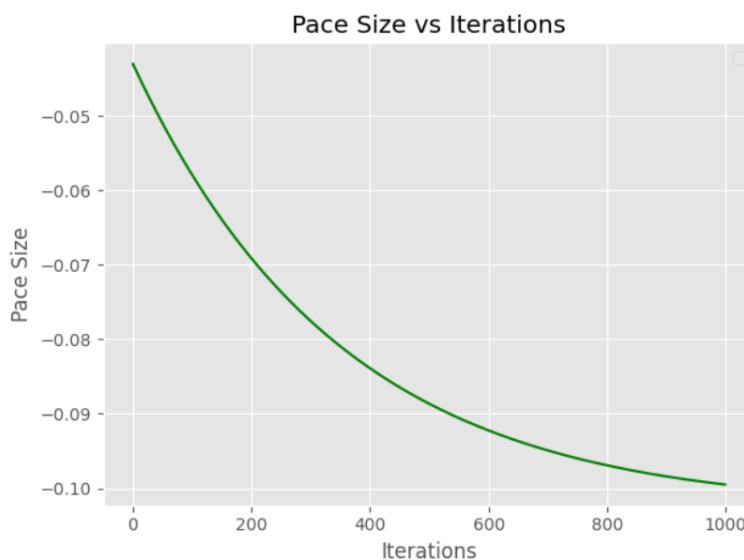
ب) نمودار مقدار خطای تست برای مقادیر مختلف لامبدا رسم شده است :



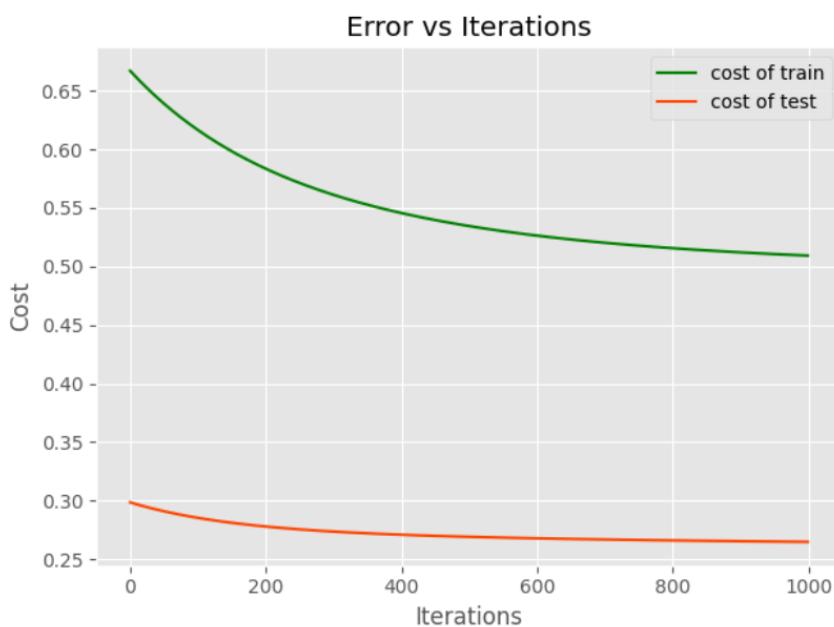
با توجه به این نمودار مقدار بهینه لامبدا که خطای کمینه در آن رخ می دهد برابر ۰.۸ می باشد.

مقدار خطای MSE روی داده های تست : ۰,۲۶۰۳

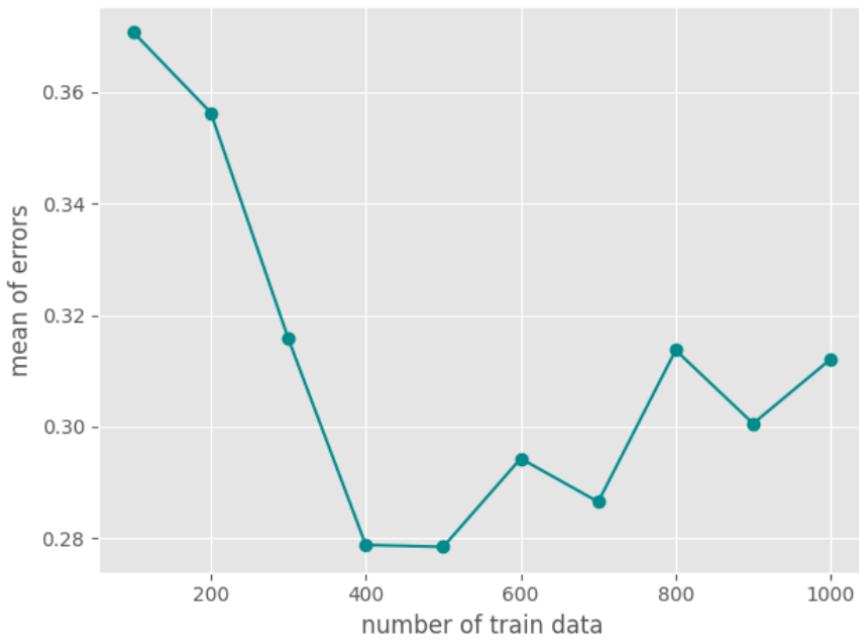
نمودار اندازه گام به ازای تکرار های مختلف :



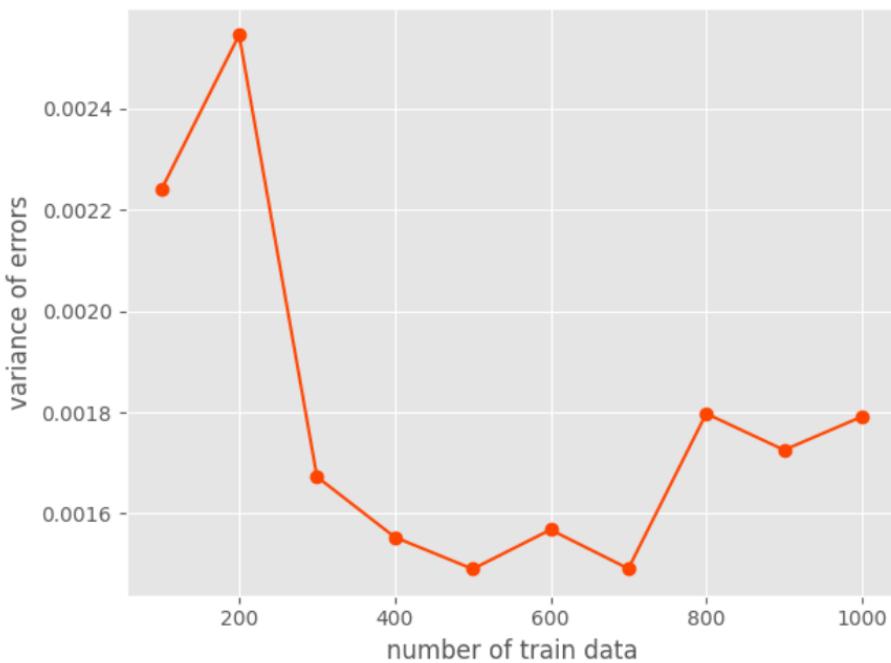
نمودار خطای آموزش و آزمون به ازای تکرارهای مختلف:



ج) نمودار تغییرات میانگین داده ها با افزایش تعداد داده های آموزش:



نمودار تغییرات واریانس داده‌ها با افزایش تعداد داده‌های آموزش:



رونده کلی واریانس و میانگین خطاهای آموزش به این صورت است که ابتدا کاهش می‌یابند و سپس دوباره افزایش می‌یابند و وقتی تعداد داده آموزشی حدود ۴۰۰ است کمترین میزان واریانس و میانگین را در خطاهای داریم.