یادگیری ماشین - تمرین دوم - بخش عملی

دانیال ملک محمد - 94100092

سوال 6)

در فایل HW2_94100092 در ابتدا تعدادی تابع کمکی نوشته شده . سپس هر قسمت سوال به صورت تابعی با نام آن قسمت نوشته شده که می توانید برای بررسی صحت عملکرد آن تابع را فقط در بدنه فایل صدا بزنید.

W=[age^2 , gender , bmi , children , smoke , southwest , southeast , northwest , northeast , constant]

کلیه خطا ها برحسب MSE بیان شده اند.

سوال 1-6)

الف)

از فرمول بسته ی $W = (X^\top X)^{-1} X^T Y$ استفاده کردم .
تابع numpy.linalg.inv(X)، $(X^\top X)^{-1}$ را دقیق بدست نمی آورد و هنگام ضرب در دیگر ماتریس های رابطه، خطایش انتشار یافته و W بدست آمده به حدی اشتباه بود که برای تعداد قابل توجهی داده های آموزش ، مقدار منفی پیشبینی می کرد!!!
به منظور کاهش این خطا از تابع numpy.linalg.pinv(X) استفاده کردم که
Pseudo Inverse ماتریس X را خروجی می دهد : $pinv(X) = (X^\top X)^{-1} X^\top$

W = [3.30752409 , -283.18596 , 337.031996 , 553.892111 , 23882.3566 , -2181.96141 , -2203.08940 , -1361.03029 , -889.808859 , -6635.88997]

Mean Square Loss on Train : 34673667.87030672
Mean Square Loss on Test : 43731033.440142736

لازم به ذکر است که رگرسیون خطی با در نظر گرفتن عدد کانستنت و ثابت (w0) با حالت بدون آن تفاوتی نداشت و عدد خطای مشابهی داشتند.

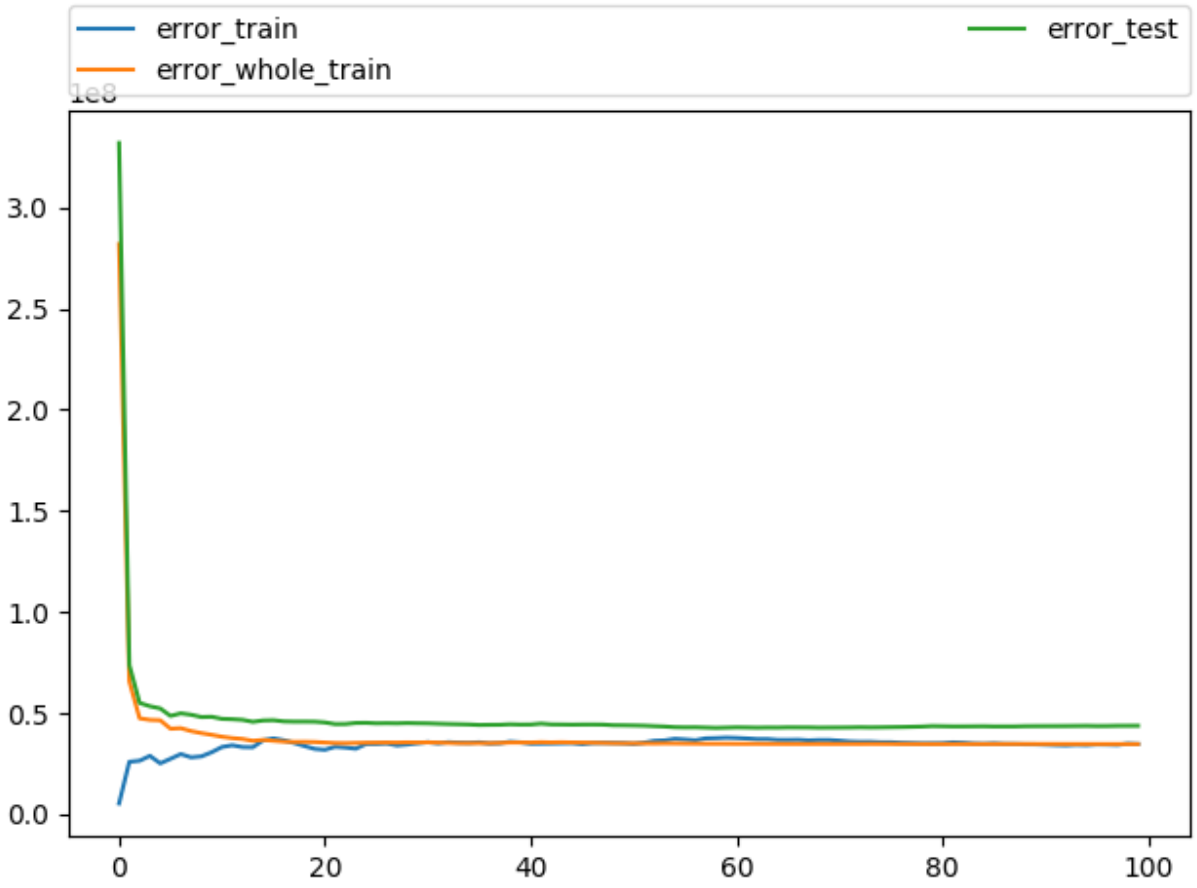
ب)

حاصل نمودار زیر شد، خط آبی خطا روی آن تعداد داده ها از train است که برای آموزش انتخاب شدند.
خط نارنجی خطا روی کل داده های train است .
خط سبز خطا روی کل داده های تست است.

در ابتدا چون داده هایی که برای آموزش انتخاب می شوند، کم هستند مدل به راحتی رویشان Overfit می شود و لذا خطا روی خط آبی کم است. اما به مرور زمان که تعداد داده های انتخابی برای آموزش زیاد می شود ، مدل دیگر نمی تواند به آن ها overfitشود و خطا روی آن ها افزایش می یابد و خط آبی به خط نارنجی میل می کند.
اینکه در ابتدا خط نارنجی هم زیاد است منطقی است، مدل Overfit روی تعداد محدود انتخابی از داده ای Train مدل خوبی برای کل داده ها نیست و لذا خطایش روی واقیت زیاد است پس در ابتدا هم خط نارنجی و هم خط سبز باید زیاد باشند و به مرور زمان که مدل دقیقتر می شود این خطا کاهش می یابد.

اینکه خط آبی به خط نارنجی میل کند و در آخر یکی شود هم منطقی است چون در نهایت خط آبی کل داده های آموزش را برای آموزش انتخاب می کند و خطایی که حساب می کند دقیقا خطای کل داده های آموزش است یعنی همان خط نارنجی.

و نکته ی حائز اهمیت اینکه خط آبی و نارنجی به هم میل می کنند و یکی می شوند اما خط سبز ، در مقدار خطای کمی بیشتر ،ساکن می شود. وخوب معمولا خطای تست از آموزش بیشتر است. می توان گفت در کل آموزش نهاییمان مقدار خیلی overfitکمی شده.



سوال2-6)

(Batch Gradient Descend

از فرمول $W_{t+1} = W_t + \eta X^\top (Y - X W_t)$ برای آپدیت W استفاده کردیم .
البت قسمتی از محاسبات این فرمول در هر گام تکراری و مستقل از W است که آن را به فرم زیر نوشتیم تا سریعتر انجام شود.

$$W_{t+1} = W_t + \eta(A - BW_t) \quad : \quad A = X^\top Y \quad , \quad B = X^\top X$$

A , Bکافی است یکبار محاسبه شود .

من ضریب اتا را برابر با 3×10^{-10} گذاشتم و تا 743 000 000 ایتریشن جلو رفتم.

کدی که زدم به این صورت کار می کند که هر یک صد هزار ایتیریشن یکبار W جدید را ثبت می کند . همچنین خطایMSE را در MSE_BGD ذخیره می کند.
دو فایل مذکور ضمیمه شده اند و هریک شامل ۷۴۰۰ واحد زمانی داده هستند.

خطا روی داده ی train :

BGD_MSE_train = 34673667.96640654
Closed_Form_MSE_train = 34673667.87030672

در ایتیریشن نهایی، تفاوت خطا روی داده های train در GDB و فرمول بسته، 0.1 بود.

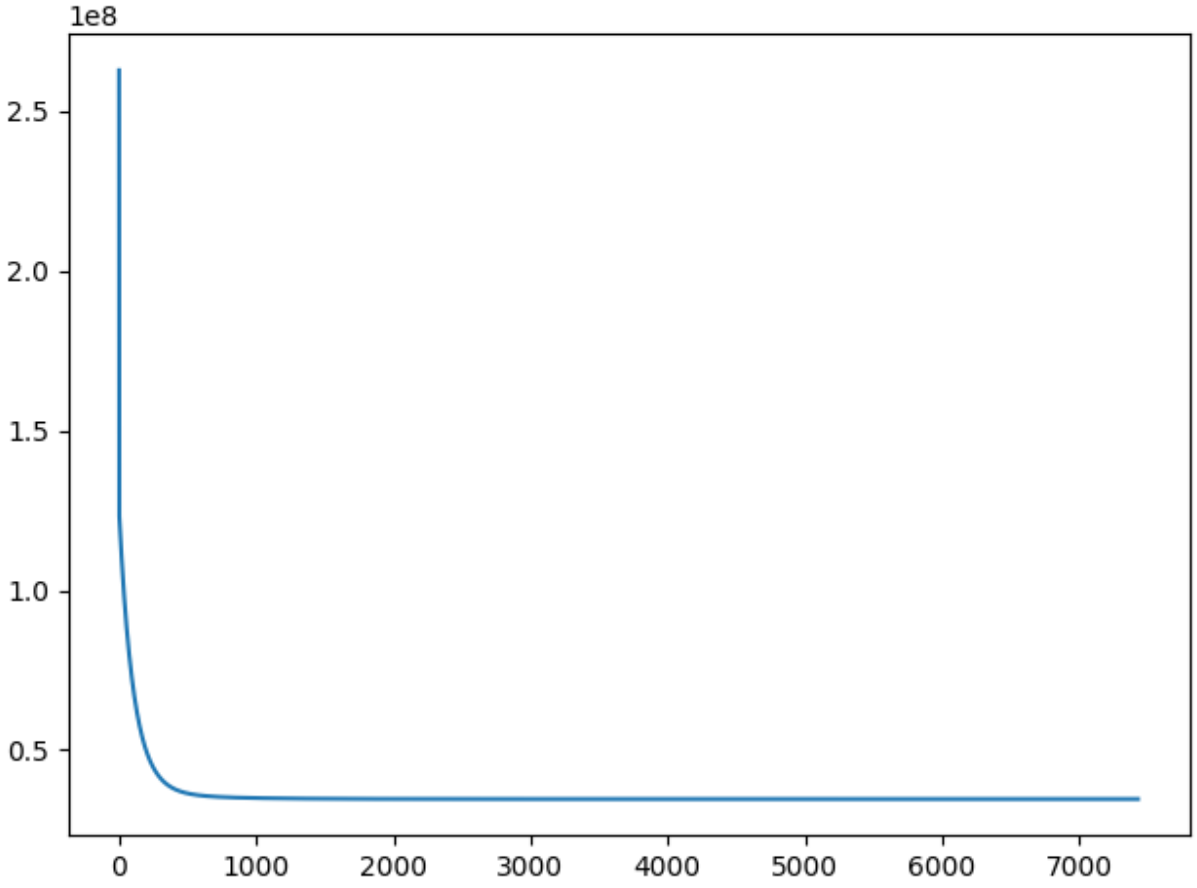
خطا روی داده ی تست:

BGD MSE on Test = 43730783.16194436
Closed Form MSE on Test = 43731033.440142736

مقایسه جواب ها:

W_BGD = [3.30747897 , -283.299432 , 336.983502 , 553.858615 , 23882.2529 , -2181.00849, -2202.00228e , -1360.16349 ,-888.924736 , -6635.09899]

W_Closed_Form = [3.30752409 , -283.18596 , 337.031996 , 553.892111 , 23882.3566 , -2181.96141 , -2203.08940 , -1361.03029 , -889.808859 , -6635.88997]



نمودار خطای روی داده ی train بر حسب تعداد ایتریشن batch . هر واحد محور X ، 100 000 تا iteration است.

سوال2-2-6)

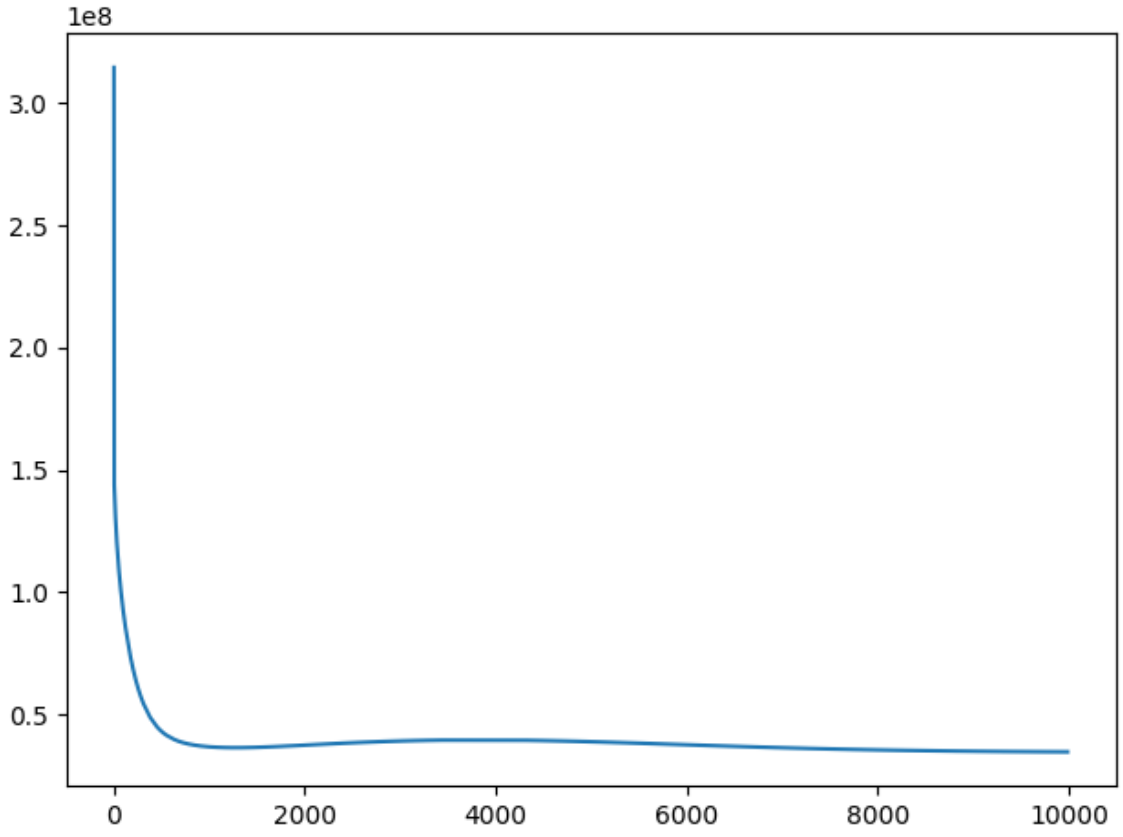
من ضریب اتا را برابر با 2×10^{-7} گذاشتم و کمی کمتر از 000 000 1000 ایتریشن جلو رفتم.
کد دقیقا مشابه قسمت قبل است و هر 000 100 تا ایتریشن سمپل، یک بار W و MSE ثبت می شد.
این داده ها در فایل های W_SGD.txt و MSE_SGD.txt ثبت شدند که ضمیمه گزارش شده اند.
خطا ها به شرح زیر هستند :

Sample Mode Gradient Descend MSE on Train : 34759037.02944656
Closed Form MSE on Train : 34673667.87030672

Sample Mode Gradient Descend MSE on Test : 44090248.3435503
Closed Form MSE on Test : 43731033.440142736

W_SGD = [3.188822 , -446.7202517 , 322.679422 , 503.886951 , 23873.914343 , -1985.751250, -2015.411643, -1309.186604, -862.541683, -6172.891183]

W_Closed_Form = [3.30752409 , -283.18596 , 337.031996 , 553.892111 , 23882.3566 , -2181.96141 , -2203.08940 , -1361.03029 , -889.808859 , -6635.88997]



نمودار خطای روی داده ی train بر حسب تعداد ایتریشن batch . هر واحد محور X ، 000 100 تا iteration است

سوال3-6)

فرمول بسته ی جواب برای فرم L2 به صورت زیر است :

$$(X^\top X + \lambda I)^{-1} X^\top Y$$

که به دلیلی که در بخش اول تمرین عملی در مورد پایین بودن دقت تابع ماتریس معکوس numpy گفته شد، نیازمند محاسبه ی sudo inverse هستیم اما مشکل اینجاست که این فرم، فرم sudo inverse از ماتریسی را برای ما تداعی نمی کند.
به این منظور trick یی را به کار بردیم .
به راحتی می توان بررسی کرد که جواب بسته مسئله ی رگرسیون خطی عادی برای

$$X' = X^\top X + \lambda I \quad , \quad Y' = X^\top Y$$

معادل رگرسیون L2 نرم برای X,Y است . به بیان دیگر :

$$OLS(X^\top X + \lambda I \quad , \quad X^\top Y) = L2\_OLS(X, Y)$$

برای اثباتش کافی است جواب بسته ی سمت چپ را ساده کنید تا به فرمول بسته ی L2_norm برسید.

برای مقادیر مختلف لاندا حاصل به صورت زیر شد:

(10)^-4 : 35398556.17517044
(10)^-3 : 35398547.81262727
(10)^-2 : 35398465.00268036
(10)^-1 : 35397718.20304868
(10)^0 : 35398110.4206911
(10)^1 : 35990987.80218144
(10)^2 : 53972880.321190156
(10)^3 : 106399032.68167143
(10)^4 : 123126461.34155416

(روی یک دسته ی ۸۰۰ تایی آموزش دید و روی دسته ۲۰۰ تایی validation ، mse اش به دست آمد و این کار را به ۵ حالت مختلف دسته بندی به این شکل انجام داد و میانگین گرفت)

سپس مشخص شد که لاندا 0.1 مناسب ترین مقدار است و با این مقدار لاندا آموزش روی کل داده ی آموزش و validation (۱۰۰۰ تا داده) انجام شد و خطای MSE روی داده های train و test به صورت زیر شد:

MSE on whole train : 34673723.321276926
MSE on test : 43731868.03658344