درس یادگیری ماشین

گزارش تکلیف Linear Regression

استاد درس:
دکتر افتخاری

نگارش:
امیرحسین ابوالحسنی
شماره دانشجویی: ۴۰۰۴۰۵۰۰۳

پاییز 1403

فهرست مطالب

۱ مقدمه [۱]

رگرسیون خطی یکی از ساده‌ترین و پرکاربردترین روش‌ها در یادگیری نظارت‌شده برای مدل‌سازی رابطه بین یک متغیر وابسته و یک یا چند متغیر مستقل است. با فرض وجود یک رابطه خطی بین ویژگی‌های ورودی و متغیر هدف، رگرسیون خطی تلاش می‌کند بهترین خط را پیدا کند که خطای پیش‌بینی را به حداقل برساند.

هنگامی که تنها یک متغیر مستقل وجود دارد، این روش به عنوان رگرسیون خطی ساده شناخته می‌شود. با این حال، مشکلات دنیای واقعی اغلب شامل چندین عامل مؤثر بر متغیر هدف است. در چنین مواردی از رگرسیون خطی چندگانه استفاده می‌شود، جایی که مدل چندین متغیر مستقل را برای پیش‌بینی متغیر وابسته در نظر می‌گیرد. به صورت ریاضی، این رابطه به صورت زیر مدل‌سازی می‌شود:

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n + \epsilon$$

در اینجا y مقدار پیش‌بینی‌شده، w_0 عرض از مبدأ، $w_1, w_2, \ldots, w_n$ ضرایب متغیرهای مستقل $x_1, x_2, \ldots, x_n$، و ϵ خطای مدل است.

در عمل، رگرسیون خطی می‌تواند با استفاده از تکنیک‌های مختلفی حل شود، مانند حل بسته [۲] که از کمینه‌سازی میانگین مربعات خطا [۳] مشتق شده است یا الگوریتم‌های بهینه‌سازی مانند گرادیان نزولی. علاوه بر این، تکنیک‌های تنظیم [۴] مانند تنظیم L_2 (رگرسیون ریدج) می‌توانند برای جلوگیری از بیش‌برازش با جریمه کردن مقادیر بزرگ ضرایب اعمال شوند. این روش‌ها به‌ویژه در مقابله با چندخطی بودن یا داده‌های پرنویز، پایداری مدل را تضمین می‌کنند.

این گزارش به بررسی رگرسیون خطی ساده و چندگانه می‌پردازد و بر حل بسته در حالت معمولی و با تنظیم L_2، بهینه‌سازی گرادیان نزولی در دو حالت گرادیان نزولی تصادفی [۵] و دسته‌ای [۶] تاکید دارد. . همچنین عملکرد مدل در روش‌های مختلف با استفاده از داده‌های آموزش و آزمایش مقایسه شده است.

۲ رگرسیون خطی تک متغیره

۱.۲ تابع هزینه رگرسیون خطی چیست؟

تابع هزینه رگرسیون خطی با نام میانگین مربع خطاها یا Mean Square Error (MSE) شناخته می‌شود.

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

۲.۲ آموزش رگرسیون خطی

۱.۲.۲ راه حل بسته

در حل بسته، باید رابطه زیر حل شود تا θ پیدا شود.

$$\theta = (X^TX)^{-1}X^T\vec{y}$$

البته ممکن است X^TX معکوس پذیر نباشد (که در تکلیف همین مورد برای داده‌های Train اتفاق می‌افتد.) که در این صورت می توان از Moore-Penrose pseudo-inverse در رابطه راه حل بسته استفاده کرد:

$$\theta = (X^TX)^{+}X^T\vec{y}$$

پاسخ به دست آمده:

$$y = \theta_0 + \theta_1 x$$

$$\theta_0 = -3.8957, \theta_1 = 1.1930$$

---

[۱] مقدمه با مدل GPT 3 نوشته شده است.
[۲] Closed Form Solution
[۳] Mean Squared Error (MSE)
[۴] Regularization
[۵] Stochastic Gradient Descent
[۶] Batch Gradient Descent

۲.۲.۲ گرادیان کاهشی تصادفی

در این روش از الگوریتم گرادیان کاهشی تصادفی برای بهینه سازی تابع هزینه استفاده می‌شود.
پاسخ به دست آمده:

$$y = \theta_0 + \theta_1 x$$

$$\theta_0 = -3.8481, \theta_1 = 1.0570$$

۳.۲.۲ گرادیان کاهشی دسته‌ای

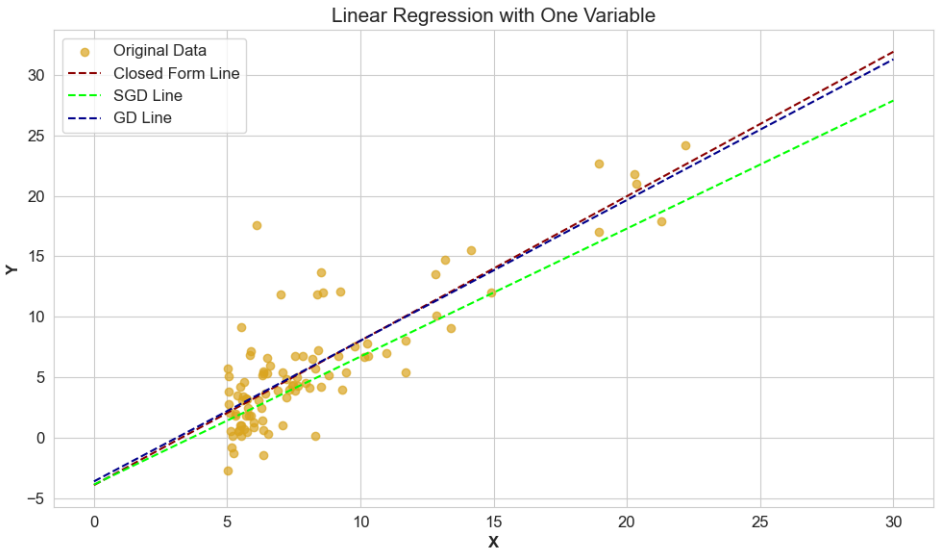
در این روش از الگوریتم گرادیان کاهشی دسته‌ای برای بهینه سازی تابع هزینه استفاده می‌شود.
پاسخ به دست آمده:

$$y = \theta_0 + \theta_1 x$$

$$\theta_0 = -3.5858, \theta_1 = 1.1619$$

۳.۲ مصور سازی مدل با داده‌ها

هر سه پاسخ به دست آمده در کنار داده‌ها در شکل ۱ به تصویر کشیده شده است. همانطور که دیده می شود گرادیان کاهشی دسته‌ای بهتر از تصادفی عمل کرده است.



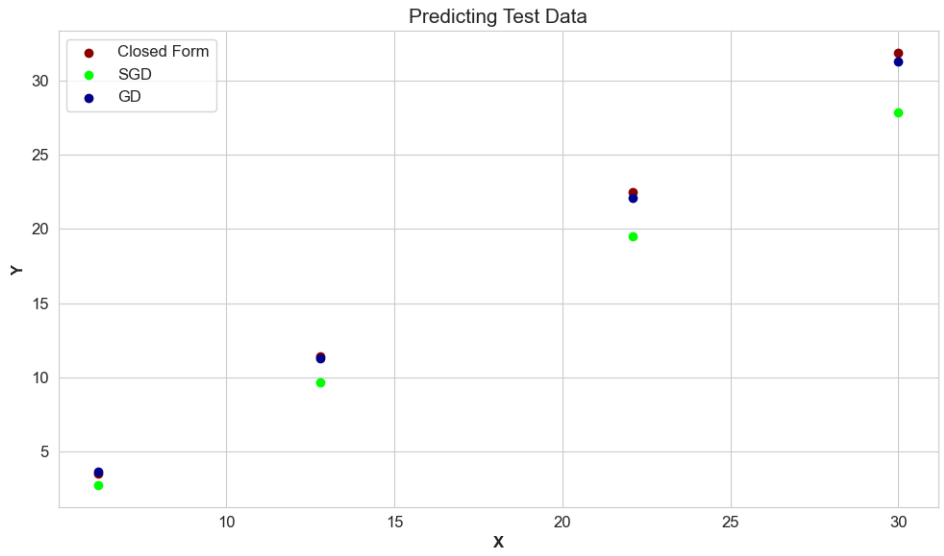
شکل ۱: خطوط به دست آمده از سه روش مختلف

۴.۲ مقایسه پیش‌بینی مدل‌ها

پس از آموزش هر مدل، برای تست کردن، به آن‌ها داده‌های

$$X = 6.2, 12.8, 22.1, 30$$

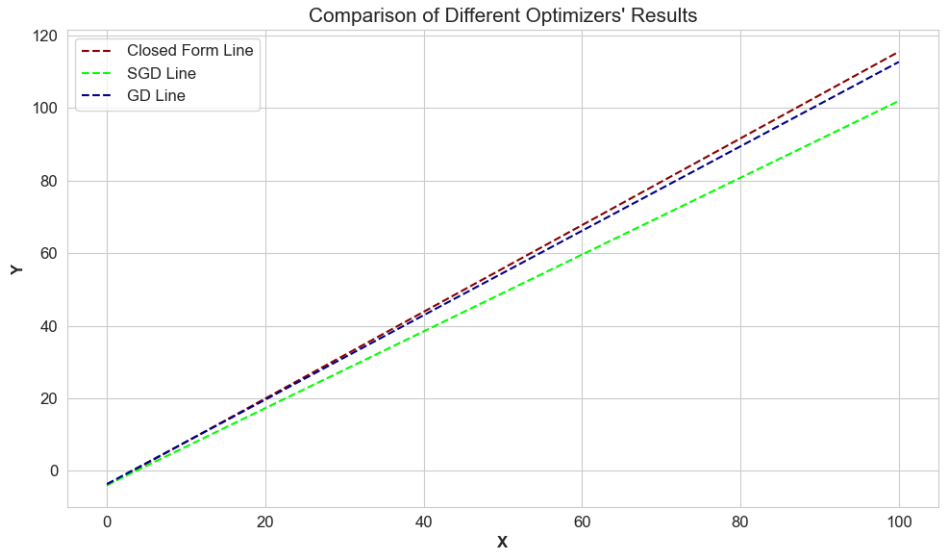
داده می‌شود. نزدیکی خروجی هر مدل، به مدلی که از راه حل بسته به دست آمده است می‌تواند مشخص کند آن شیوه بهینه سازی چقدر خوب توانسته پارامترهایی را پیدا کند که تابع هزینه را کمینه می‌کنند.



شکل ۲: خروجی مدل در روش

۵.۲ مقایسه مدل‌های آموزش دیده

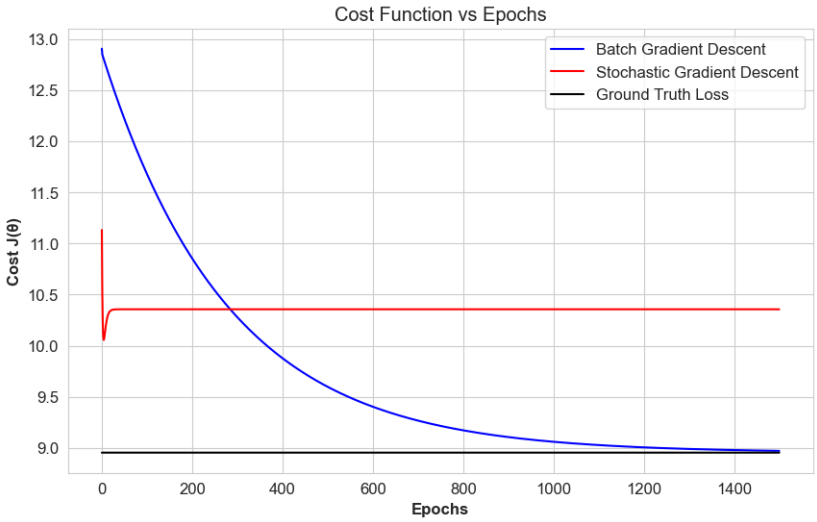
از آنحایی که پاسخ بسته بهترین پاسخی است که می‌توان به آن رسید، مقایسه هر خطی با پاسخ بسته می‌تواند نشان بدهد آن مدل چقدر خوب است.(شکل ۳)



شکل ۳: خروجی مدل در روش

۶.۲ بررسی رفتار تابع هزینه

از مهمترین روش‌های مانیتور کردن کیفیت یادگیری یک مدل، بررسی رفتار تابع هزینه در هنگام آموزش می‌باشد.
همانطور که در شکل ۴ دیده می شود، گرادیان تصادفی بعد از ۲۰ تکرار دیگر نمی‌تواند تابع هزینه را مینیمم کند و احتمالا در یک بهینه محلی گیرکرده است. از طرفی گرادیان کاهشی دسته‌ای توانسته به نزدیکی مقدار مینیمم کلی تابع هزینه برسد.



شکل ۴: نمودار مقدار تابع هزینه با توجه به هر تکرار در زمان یادگیری

۷.۲ کدام روش بهینه سازی ترجیح داده می‌شود؟

با توجه به بخش ۶.۲ و شکل ۴ می‌توان گفت از آنجایی که گرادیان کاهشی دسته‌ای نسبت به تصادفی بیشتر به کل داده‌ها دید دارد، بهتر می‌تواند مسیر خود را به سمت مینیمم تابع هزینه پیدا کند، مخصوصا زمانی که داده‌ها کم هستند و میتوانند در رم باشند.

۳ رگرسیون چند متغیره

۱.۳ پیش پردازش داده‌ها

۱.۱.۳ انکود داده‌ها

بعضی از متغیرهای دیتاست کتگوریکال هستند و نیاز بود که اینها متغیرها انکود شوند. از One Hot Encoding (OHE) و Integer Encoding (IE) برای انکودینگ استفاده شد. متغیر *region* با استفاده از OHE انکود شده و برای ویژگی‌های *smoker, gender* از روش دوم انکودینگ استفاده شده است.

سوال: چرا برای این سه ویژگی از روش های متفاوت استفاده شد؟

زمانی از IE برای انکودینگ استفاده می‌کنیم که بدانیم ترتیب در مقادیر ویژگی مهم است. ممکن است این سوال پیش بیاید که پس چرا برای ویژگی *gender* از این روش استفاده می‌شود. می‌توان گفت عملا تفاوتی نمی‌کند چون تنها دو مقدار برای این ویژگی وجود دارد. و اما زمانی که از روش OHE استفاده می‌کنیم دیگر ترتیب اثری ندارد و با هر ویژگی ایجاد شده که مقدار ۰ یا ۱ را دارد به طور عادلانه برخورد می‌شود، همانطور که از ماهیت ویژگی *region* مشخص است.

۲.۱.۳ استانداردسازی

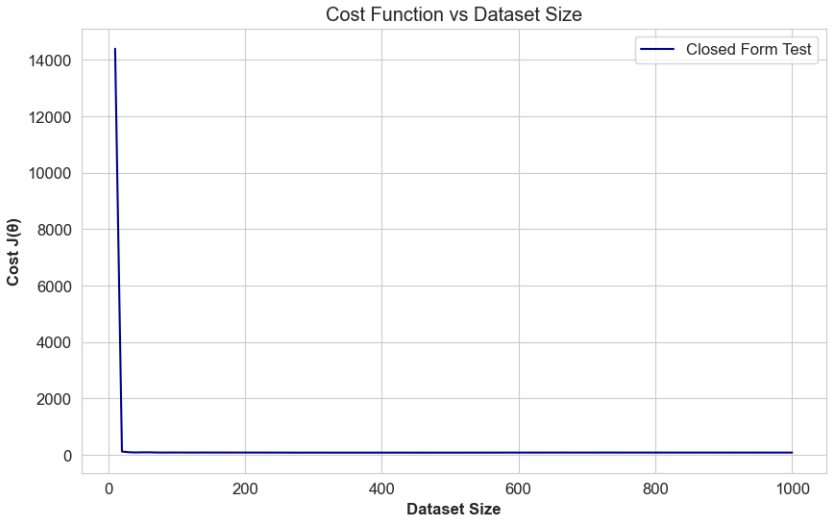
در زمان انجام تکلیف، دیده شد که عدم استانداردسازی داده‌ها باعث انفجار گرادیان و پارامترها می‌شود. به همین علت استانداردسازی روی داده‌ها های آموزشی انجام شد و با همان پارامترها، داده های تست نیز استانداردسازی شدند.

۲.۳ آموزش مدل

۱.۲.۳ راه حل بسته

در این بخش، چون نمی‌توان از X^TX (که X ماتریس داده آموزشی باشد) معکوس گرفت، از Moore-Penrose pseudo inverse استفاده می‌کنیم.

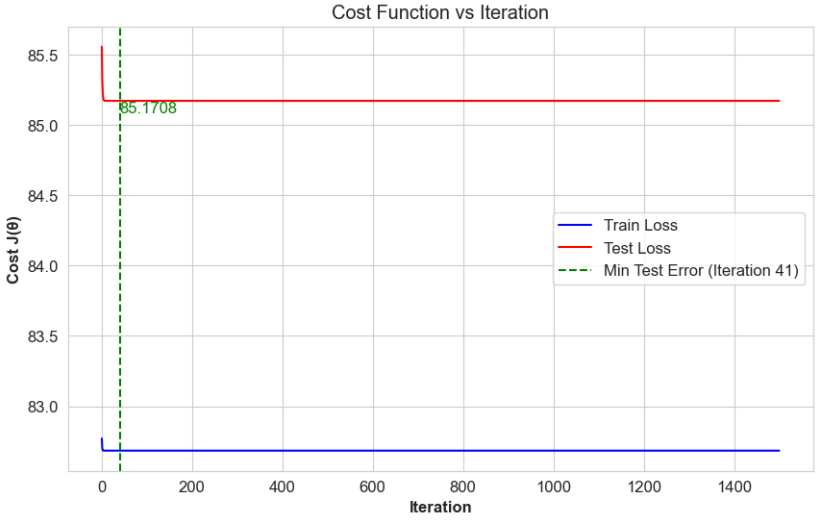
همچنین با توجه به نمودار ۵ می‌توان مشاهده کرد که عملکرد راه حل بسته با مقدار مختلف از داده‌های یادگیری، در ابتدا بسیار کم بوده اما بعد از حدود ۲۰ داده توانسته به خطای ثابتی برسد.



شکل ۵: نمودار نشان‌دهنده مقدار میانگین خطا نسبت به اندازه دیتاست آموزش

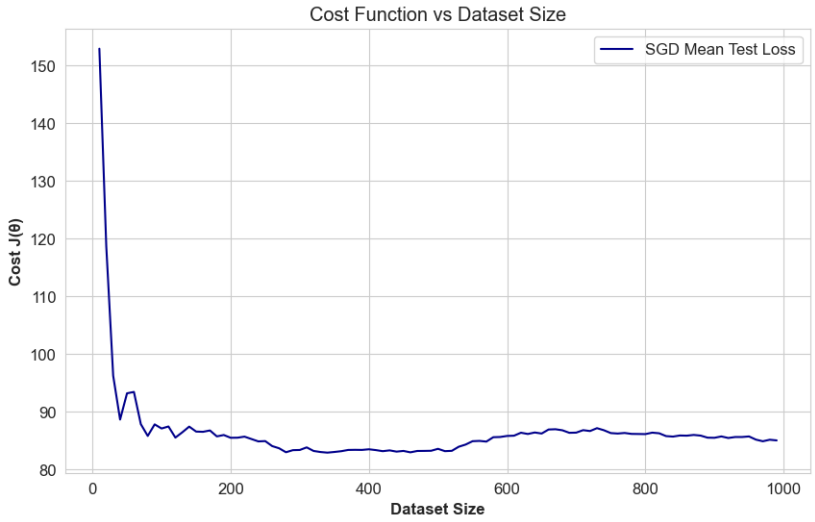
۲.۲.۳ گرادیان کاهشی تصادفی

مدل رگرسیون خطی را با گرادیان کاهشی تصادفی اموزش میدهیم. نمودار ۶ نشان دهنده روند Test Loss و Train Loss را برای یادگیری روی تمام داده‌های بخش Train نشان می‌دهد.



شکل ۶: نمودار رفتار خطا تست و خطا آموزش برای گرادیان کاهشی تصادفی (کمترین خطا در تکرار ۴۱ با مقدار ۸۵.۱۷ می‌باشد.)

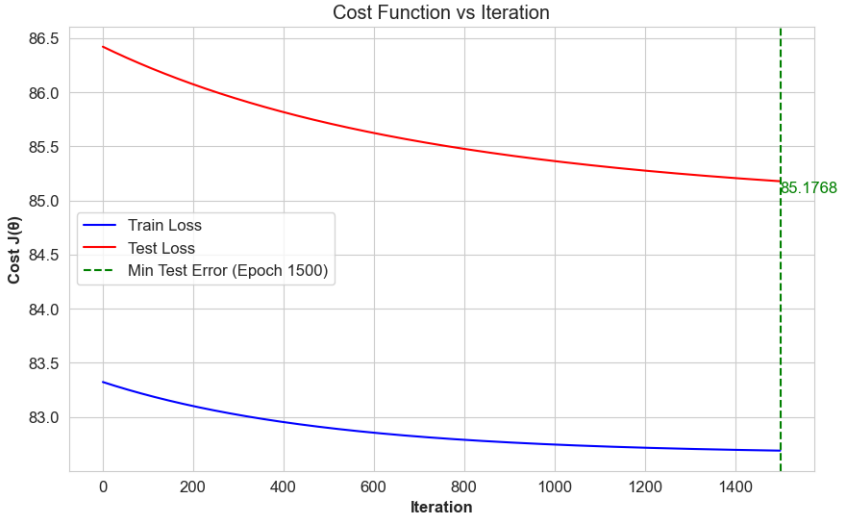
همچنین میانگین خطا نسبت به اندازه دیتاست در شکل ۷ بررسی شده است.



شکل ۷: میانگین خطا مدل آموزش داده شده روی اندازه‌های مختلف دیتاست

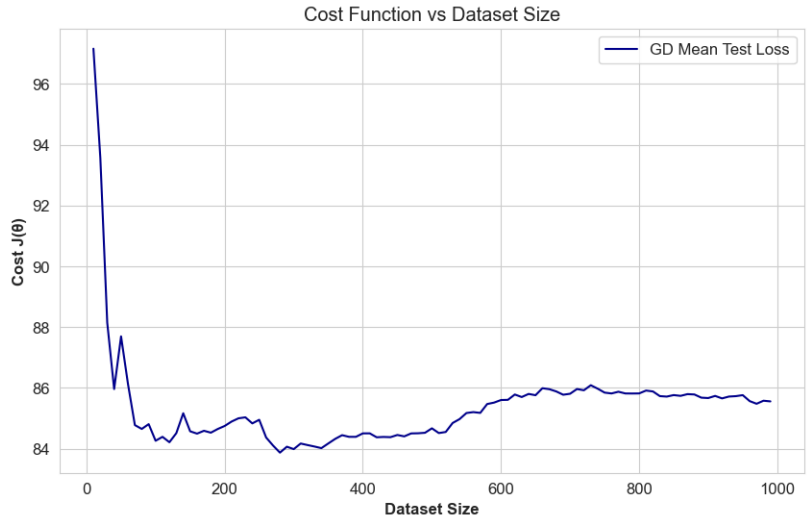
۳.۲.۳ گرادیان کاهشی دسته‌ای

مدل رگرسیون خطی را با گرادیان کاهشی دسته‌ای اموزش میدهیم. نمودار ۸ نشان دهنده روند Test Loss و Train Loss را برای یادگیری روی تمام داده‌های بخش Train نشان می‌دهد.



شکل ۸: نمودار رفتار هزینه تست و هزینه آموزش برای گرادیان کاهشی دسته‌ای (کمترین خطا در تکرار ۱۵ با مقدار ۸۵.۱۷ می‌باشد.)

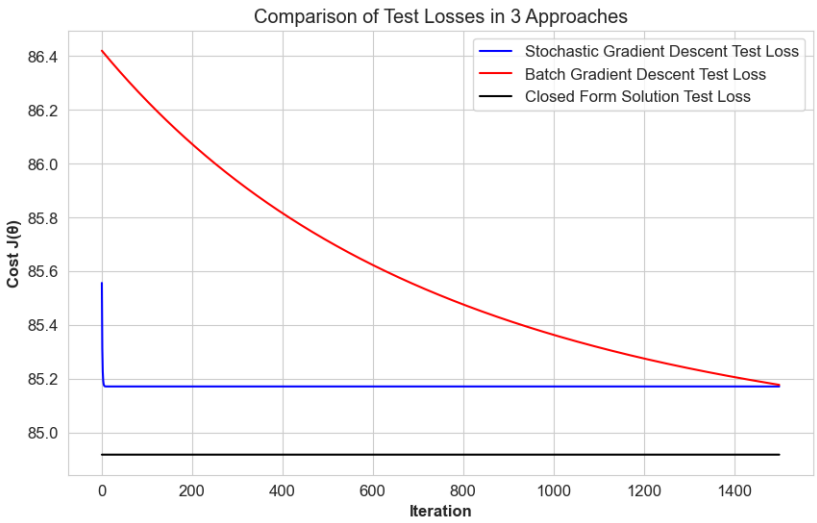
همچنین میانگین خطا نسبت به اندازه دیتاست در شکل ۹ بررسی شده است.



شکل ۹: میانگین خطا مدل آموزش داده شده روی اندازه‌های مختلف دیتاست

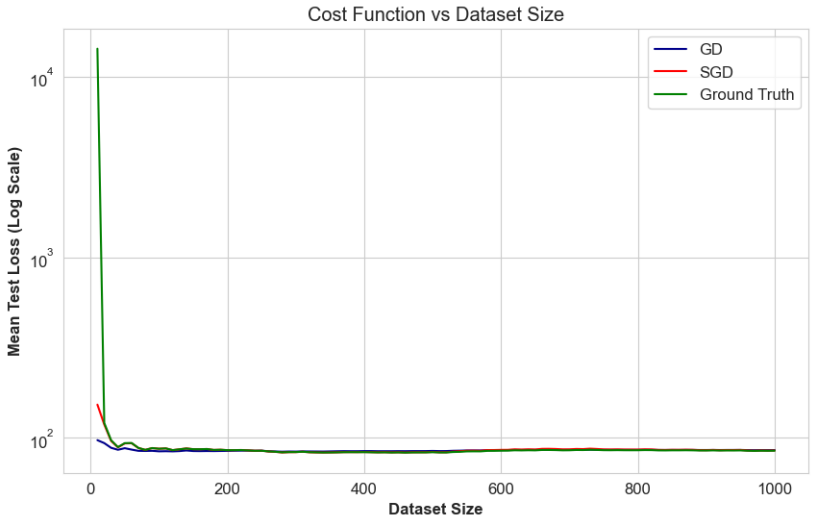
۴.۲.۳ مقایسه سه روش

به طور کلی در شکل روند کاهش مقدار خطا روی داده‌های تست نشان داده شده است. مدلی که از طریق بهینه سازی SGD آموزش دیده بود، سریعا به همگرایی رسید اما روش BGD با روندی کندتر، اما ثابت و پیشبینی پذیر به همگرایی میرسد. دیده می‌شود (شکل ۱۰) که هر دو روش با کمترین خطای ممکن کمتر از ۰.۲ فاصله دارند.



شکل ۱۰: روند کاهش خطای تست در سه روش SGD, SGD و راه حل بسته

یکی دیگر از مقایسه‌ها، تفاوت مقدار میانگین خطا تست نسبت به اندازه دیتاست می‌باشد. همانطور که در شکل ۱۱ دیده می‌شود، در ابتدا روش‌های جستجو مانند گرادیان کاهشی بهتر از راه حل بسته ظاهر شده اند.



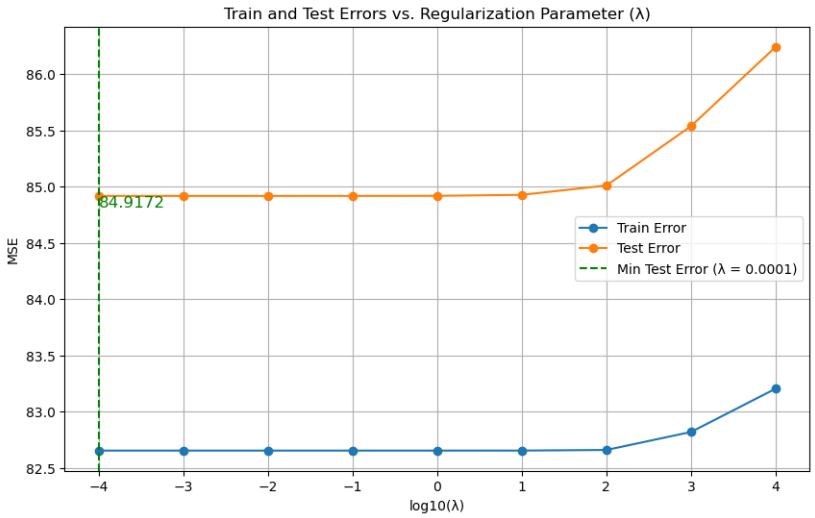
شکل ۱۱: میانگین خطا تست به اندازه دیتاست

۳.۳ اضافه کردن L_2 به راه حل بسته

با اضافه کردن ترم L_2 به تابع هزینه، به رابطه زیر برای به دست آوردن پارامتر θ می‌رسیم.

$$\theta = (X^T X + \lambda I)^{-1} X^T \vec{y}$$

برای به دست آوردن مقدار مناسب λ، مقدار خطای تست پرای پارامترهای به دست آمده با مقادیر مختلف λ در شکل ۱۲ دیده شده است.



شکل ۱۲: میانگین خطا تست برای مقادیر متفاوت λ (کمترین مقدار خطای تست، با مقدار $\lambda = 10^{-4}$ به دست می‌اید)