گزارش تحلیل رگرسیون خطی

هیوا ابوالهادی زاده ۴۰۰۴۰۵۰۰۴
درس یادگیری ماشین

۶ آذر ۱۴۰۳

۱ مقدمه

در این گزارش به بررسی و پیاده‌سازی روش Linear Regression پرداخته شده است. مسائل بررسی‌شده شامل دو بخش زیر می‌باشند:

- رگرسیون خطی با یک متغیر
- رگرسیون خطی با چند متغیر

هدف از این تحلیل، پیش‌بینی مقادیر خروجی و ارزیابی دقت مدل‌ها با استفاده از روش‌های مختلف مانند حل بسته (Closed-form Solution)، روش نزول گرادیان (Gradient Descent) و تنظیم L2 است. نتایج شامل پیش‌بینی‌ها، نمودارها، و مقایسه مدل‌ها ارائه می‌شود.

۲ بخش اول: رگرسیون خطی با یک متغیر

۱.۲ فرمول تابع هزینه

برای رگرسیون خطی، تابع هزینه به صورت زیر تعریف می‌شود:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 \quad (۱)$$

که در آن:

- $h_\theta(x_i) = \theta_0 + \theta_1 x_i$: مقدار پیش‌بینی‌شده.
- m: تعداد نمونه‌ها.
- y_i: مقدار واقعی خروجی.

۲.۲ روش‌های تخمین پارامترها

برای تخمین مقادیر θ_0 و θ_1، سه روش زیر استفاده شده است:

۱. حل بسته با روش **MSE**: با حل مستقیم معادلات، پارامترها محاسبه شدند.

۲. نزول گرادیان تصادفی (**Stochastic GD**): به ازای هر نمونه ورودی، پارامترها با استفاده از شیب محاسبه شده به‌روز شدند.

۳. نزول گرادیان دسته‌ای (**Batch GD**): به‌روزرسانی پارامترها با استفاده از میانگین گرادیان کل داده‌ها.

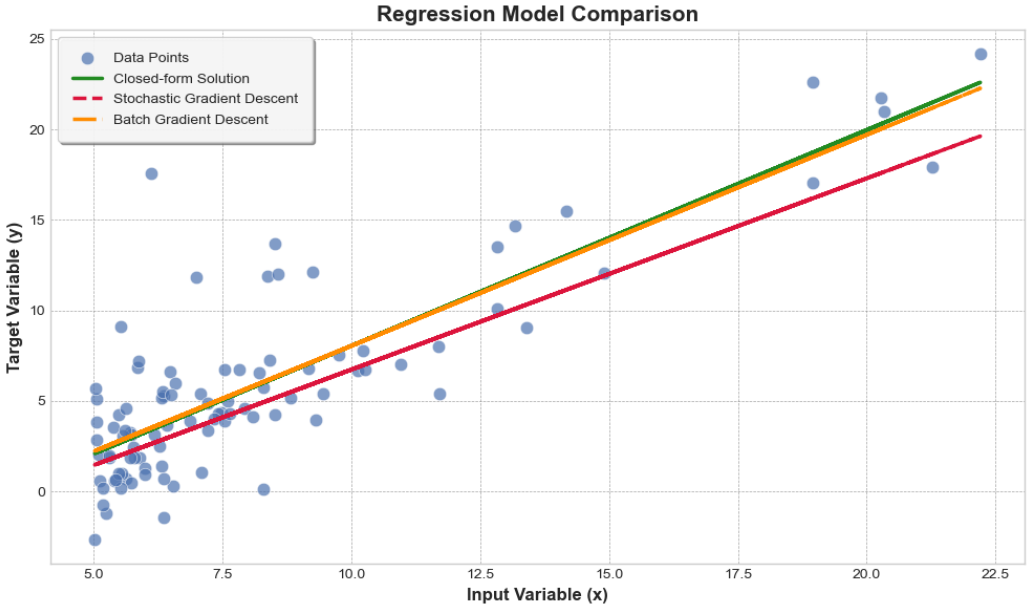
۳.۲ نمودارها

- نمودار داده‌های ورودی و خروجی (Scatter Plot) رسم شده است.
- خطوط مدل‌های تخمین‌زده‌شده توسط هر سه روش روی نمودار رسم شده‌اند.

این نمودار مقایسه‌ای بین سه روش مختلف برای رگرسیون خطی با یک متغیر را نشان می‌دهد. محور افقی نشان‌دهنده متغیر ورودی (x) و محور عمودی متغیر خروجی (y) است. نقاط آبی‌رنگ نشان‌دهنده داده‌های واقعی (نمونه‌ها) هستند و سه خط رنگی مربوط به مدل‌های تخمین زده‌شده است:

۱. خط سبز (حل بسته **Closed-form Solution**):

- این خط نشان‌دهنده مدل بهینه‌ای است که با حل مستقیم معادلات به دست آمده است.



شکل ۱: نمودار ۱

- همان‌طور که مشاهده می‌شود، این خط دقیقاً از مرکز داده‌ها عبور کرده و به نظر می‌رسد به بهترین شکل ممکن خطای میانگین مربعات (MSE) را حداقل کرده است.

۲. خط قرمز (نزول گرادیان تصادفی **Stochastic Gradient Descent**):

- این خط از الگوریتم نزول گرادیان تصادفی به‌دست آمده است.
- گرچه این روش عملکرد خوبی دارد، اما به‌وضوح کمی کمتر دقیق‌تر از روش حل بسته عمل کرده و خطای بیشتری را نسبت به آن نشان می‌دهد.
- علت این اختلاف می‌تواند به دلیل نوسانات ناشی از انتخاب تصادفی داده‌ها در هر تکرار باشد.

۳. خط نارنجی (نزول گرادیان دسته‌ای **Batch Gradient Descent**):

- این خط با استفاده از الگوریتم نزول گرادیان دسته‌ای (استفاده از کل داده‌ها در هر تکرار) ایجاد شده است.
- تطابق این خط با روش حل بسته بسیار زیاد است و تفاوت محسوسی مشاهده نمی‌شود، زیرا این روش گرادیان میانگین تمام داده‌ها را محاسبه می‌کند که دقت بیشتری دارد.

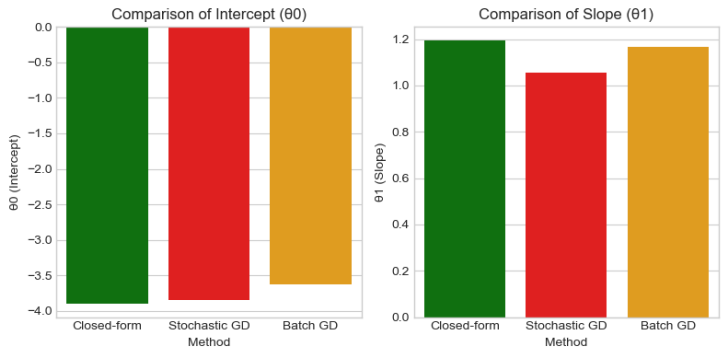
۳ تحلیل کلی

- روش حل بسته (**Closed-form Solution**): دقت بسیار بالایی دارد و بهترین خط را نسبت به داده‌ها ارائه داده است.
- نزول گرادیان دسته‌ای (**Batch Gradient Descent**): عملکرد مشابه روش حل بسته دارد اما ممکن است در داده‌های بسیار بزرگ به زمان بیشتری نیاز داشته باشد.
- نزول گرادیان تصادفی (**Stochastic Gradient Descent**): سرعت بالایی دارد، اما دقت آن نسبت به روش‌های دیگر کمتر است و معمولاً با نوسانات همراه است.

۱.۳ مقایسه پارامترها

مقدار θهای تخمین‌زده‌شده توسط هر روش در یک نمودار به‌صورت مقایسه‌ای نشان داده شده است. این نمودار مقایسه مقادیر θ_0 (عرض از مبدأ) و θ_1 (شیب) را در سه روش Closed-form Solution، Stochastic Gradient Descent و Batch Gradient Descent نشان می‌دهد.

- مقادیر θ_0 و θ_1 در روش Closed-form Solution و Batch Gradient Descent بسیار نزدیک به یکدیگر هستند.
- روش Stochastic Gradient Descent نسبت به دو روش دیگر اختلاف اندکی دارد که به دلیل ماهیت تصادفی آن است.
- این نتایج نشان می‌دهد که روش‌های Closed-form Solution و Batch Gradient Descent دقت بیشتری دارند، درحالی‌که روش Stochastic Gradient Descent سرعت را به دقت ترجیح می‌دهد.



شکل ۲: مقایسه ضرایب θ_0 و θ_1 در روش‌های مختلف

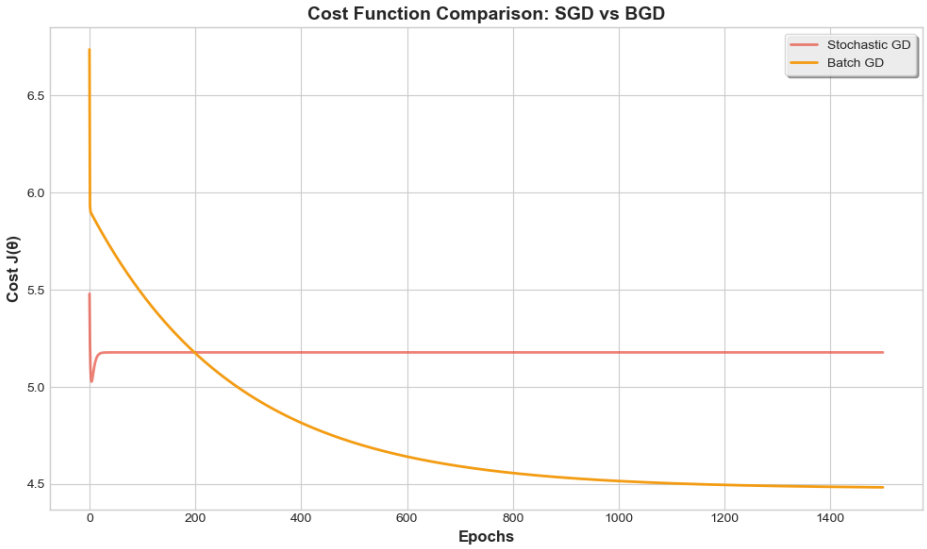
تحلیل نمودار هزینه: SGD در مقابل BGD

این نمودار هزینه‌ی $J(\theta)$ را بر اساس تعداد دوره‌ها (Epochs) برای دو روش بهینه‌سازی نشان می‌دهد:

- **Batch Gradient Descent (BGD)**: کاهش یکنواخت و پایدار هزینه.
- **Stochastic Gradient Descent (SGD)**: نوسانات بیشتر در کاهش هزینه و رسیدن به یک مقدار ثابت بزرگ‌تر.

دلایل این رفتار:

- در SGD، وزن‌ها برای هر نمونه داده به‌روزرسانی می‌شوند که منجر به نوسانات بیشتر می‌گردد.
- در BGD، به‌روزرسانی وزن‌ها پس از مشاهده کل داده‌ها انجام می‌شود که نتایج پایدارتر و دقیق‌تری دارد.



شکل ۳:

نتیجه‌گیری

با توجه به نمودار، روش Batch Gradient Descent به دلیل کاهش یکنواخت‌تر و پایدارتر هزینه و دستیابی به مقدار بهینه‌ی پایین‌تر، انتخاب بهتری نسبت به Stochastic Gradient Descent در این مسئله است.

۴ بخش دوم: رگرسیون خطی با چند متغیر

۱.۴ داده‌ها و پیش‌پردازش

داده‌ها شامل ویژگی‌هایی مانند جنسیت، منطقه، سیگار کشیدن و غیره می‌باشند. برای پیش‌پردازش:

- متغیرهای جنسیت و سیگار کشیدن با استفاده از Integer Encoding رمزگذاری شدند.
- متغیر منطقه با استفاده از One-Hot Encoding (OHE) رمزگذاری شد.

۲.۴ چرا روش **One-Hot Encoding** بهتر است؟

روش One-Hot Encoding برای کدگذاری ویژگی‌های منطقه‌ای مناسب‌تر است، زیرا برخلاف Integer Encoding، از ایجاد روابط عددی اشتباه بین مناطق جلوگیری کرده و استقلال هر منطقه را حفظ می‌کند.

۳.۴ روش‌های تخمین

۱. حل بسته با **MSE**: پارامترهای مدل با فرض BMI^2 به‌عنوان ویژگی پایه محاسبه شدند.

۲. افزایش تدریجی داده‌ها: خطای تست (MSE) به‌ازای افزایش تدریجی تعداد نمونه‌های آموزشی از ۱۰ تا ۱۰۰۰ نمونه محاسبه و گزارش شده است.

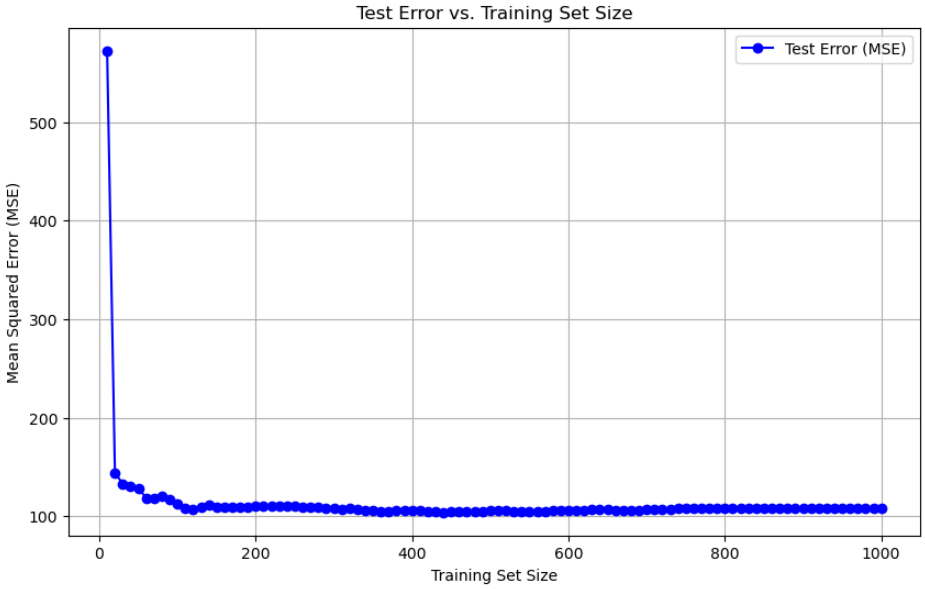
۳. نزول گرادیان دسته‌ای و تصادفی: دو روش Batch GD و Stochastic GD برای حل مسئله استفاده شده‌اند.

نمودار حل با روش بسته نشان می‌دهد که:

- با افزایش اندازه مجموعه داده‌های آموزشی، مقدار خطای آزمون (MSE) به طور چشمگیری کاهش می‌یابد.
- پس از یک حد مشخص از اندازه داده، خطا به مقدار ثابتی همگرا می‌شود.

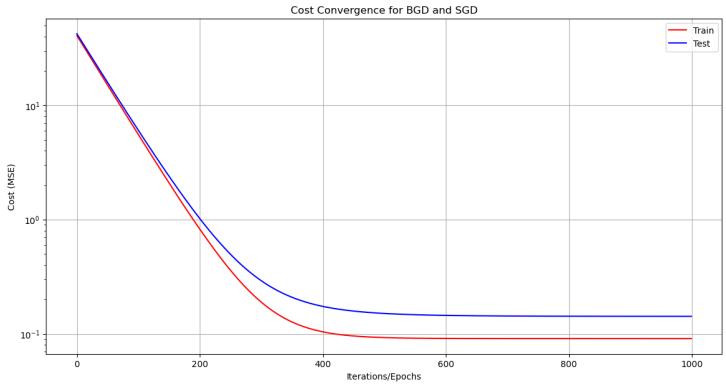
این رفتار نشان‌دهنده‌ی این است که استفاده از داده‌های آموزشی بیشتر می‌تواند دقت مدل را تا یک حد مشخص بهبود دهد، اما پس از آن، افزایش داده تأثیر قابل توجهی نخواهد داشت.

نمودار ۵ نشان می‌دهد که:



شکل ۴:

- خطای آموزش (Train Error) و خطای آزمون (Test Error) هر دو با افزایش تعداد تکرارها کاهش می‌یابند.
- خطاها به مقدار ثابتی همگرا می‌شوند که نشان‌دهنده‌ی بهینه‌سازی مناسب مدل است.
- فاصله‌ی بین خطای آموزش و آزمون کم است که نشان‌دهنده‌ی عدم وجود overfitting قابل توجه است.



شکل ۵: نمودار ۵

۴.۴ تنظیم L2

یک منظم‌کننده L2 به تابع هزینه اضافه شده است و مقدار بهینه λ در بازه 10^{-4} تا 10^{4} محاسبه شده است. نمودار خطای آموزشی و تست نسبت به مقادیر مختلف λ رسم شده است. نمودار ۶ نشان می‌دهد:

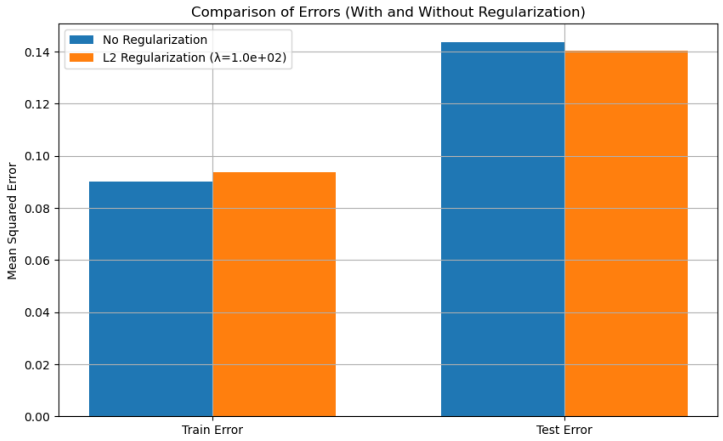
- با افزایش قدرت منظم‌سازی (λ)، خطای آموزش به‌طور پیوسته کاهش می‌یابد.
- خطای آزمون ابتدا کاهش و سپس افزایش می‌یابد، که نشان‌دهنده وقوع بیش‌برازش (Overfitting) است.
- مقدار بهینه‌ی λ حدود $\log_{10}(\lambda) = 0$ است، جایی که خطای آزمون کمینه می‌شود و مدل تعادل مناسبی بین پیچیدگی و تعمیم‌پذیری دارد.
- بعد از $\log_{10}(\lambda) = 0$، خطای آزمون افزایش یافته و مدل به سمت کم‌برازش (Underfitting) می‌رود.

شکل ۶: اثر L2

هدف انتخاب مقدار λ بهینه است که خطای آزمون را به حداقل برساند.

۵.۴ مقایسه مدل‌ها با و بدون تنظیم

مدل تنظیم‌شده خطای کمتری در داده‌های تست داشته و تقریبا از Overfitting جلوگیری کرده است.



شکل ۷: اثر L2

۵ نتیجه‌گیری

- روش Closed-form Solution سریع و دقیق است، اما برای داده‌های بزرگ محدودیت‌هایی دارد.
- روش‌های Gradient Descent برای داده‌های بزرگ مقیاس‌پذیرتر هستند. روش Stochastic GD سرعت بالاتری دارد، درحالی‌که روش Batch GD دقت بیشتری ارائه می‌دهد.
- افزودن تنظیم L2 باعث کاهش Overfitting و تعادل بهتر بین خطای آموزشی و تست شده است.