دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گزارش پروژه درس داده کاوی محاسباتی

پروژه ۳

نگارش

محمدصادق قلی زاده

استاد راهنما

دکتر مهدی قطعی

بسم الله الرحمن الرحیم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ:

اینجانب **محمدصادق قلی زاده** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

محمدصادق قلی زاده

امضا

نویسنده پایان نامه، درصورت تمایل میتواند برای سپاسگزاری پایان نامه خود را به شخص یا اشخاص و یا ارگان خاصی تقدیم نماید.

سپاس گزاری

نویسنده پایان‌نامه می‌تواند مراتب امتنان خود را نسبت به استاد راهنما و استاد مشاور و یا دیگر افرادی که طی انجام پایان‌نامه به نحوی او را یاری و یا با او همکاری نموده‌اند ابراز دارد.

محمدصادق قلی‌زاده

چکیده

این مطالعه به بررسی مبانی محاسباتی و مقیاس‌پذیری الگوریتم‌های رگرسیون خطی می‌پردازد و به‌طور خاص معادله نرمال (Normal Equation)، تجزیه مقدار منفرد (Singular Value Decomposition یا SVD)، گرادیان نزولی دسته‌ای (Batch Gradient Descent یا BGD) و گرادیان نزولی تصادفی (Stochas-tic Gradient Descent یا SGD) را با یکدیگر مقایسه می‌کند. با استفاده از یک مجموعه‌داده مصنوعی کوچک، ابتدا نشان می‌دهیم که اگرچه معادله نرمال یک راه‌حل دقیق ارائه می‌دهد، اما در شرایط هم‌خطی چندگانه (multicollinearity) از نظر عددی ناپایدار می‌شود. در مقابل، روش SVD با محاسبه شبه‌معکوس (pseudoinverse) قادر است ماتریس‌های تکین را به‌طور مؤثر مدیریت کرده و از پایداری بالایی برخوردار باشد. در ادامه، مقیاس‌پذیری را با استفاده از مجموعه‌داده واقعی Auto MPG تحلیل می‌کنیم. نتایج ما نشان می‌دهد که اگرچه BGD همگرایی همواری دارد، اما برای مجموعه‌داده‌های بزرگ از نظر محاسباتی پرهزینه است. در مقابل، SGD با وجود مسیر همگرایی پرنوسان، به دلیل به‌روزرسانی تدریجی پارامترها، گزینه‌ای بسیار کارآمدتر برای کاربردهای مقیاس بزرگ محسوب می‌شود. در نهایت، ما با موفقیت SGD را برای یک مدل رگرسیون چندجمله‌ای به کار می‌بریم و نشان می‌دهیم که الگوریتم‌های خطی می‌توانند از طریق مهندسی ویژگی، روابط غیرخطی را نیز مدل‌سازی کنند؛ به شرط آنکه نرمال‌سازی مناسب داده‌ها اعمال شود. GitHub

واژه‌های کلیدی:

الگوریتم‌های رگرسیون خطی، بهینه‌سازی با گرادیان نزولی، پایداری عددی، هم‌خطی چندگانه

فهرست مطالب

عنوان | صفحه

۱ **مقدمه** ۱

۱-۱ مقدمه ۱

۲ **مروری بر ادبیات** ۳

۲-۱ تعاریف مفاهیم پایه ۳

۲-۱-۱ معادلات نرمال ۳

۲-۱-۲ تجزیه مقدار منفرد و شبه‌معکوس ۴

۲-۱-۳ گرادیان نزولی دسته‌ای ۵

۲-۱-۴ گرادیان نزولی تصادفی ۵

۲-۱-۵ پایداری عددی ۵

۳ **روش شناسی و دیتاست** ۷

۳-۱ روش‌شناسی ۷

۳-۱-۱ روش‌های حل مستقیم ۷

۳-۱-۲ روش‌های بهینه‌سازی تکرارشونده ۸

۳-۱-۳ آماده‌سازی داده و مهندسی ویژگی ۹

۴ **نگاهی عمیق تر** ۱۰

۴-۱ نتایج و تحلیل ۱۰

۴-۱-۱ پایداری محاسباتی (مجموعه‌داده‌های کوچک) ۱۰

۴-۱-۲ مقیاس‌پذیری محاسباتی (مجموعه‌داده‌های بزرگ) ۱۲

۴-۱-۳ کاربرد در مدل‌های غیرخطی ۱۲

۵ **جمع‌بندی و نتیجه‌گیری و پیشنهادات** ۱۵

۵-۱ نتیجه‌گیری ۱۵

منابع و مراجع ۱۷

فهرست اشکال

شکل صفحه

۴-۱ مقایسه نتایج BGD و SGD ۱۳
۴-۲ نتایج Polynomial Regression با SGD ۱۴

فهرست جداول

جدول صفحه

فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
$\mathbb{S}^n$	کره یکه n بعدی
M^m	خمینه m-بعدی M
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی M
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار یکه روی (M,g)
$\Omega^p(M)$	مجموعه p-فرمی‌های روی خمینه M
Q	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
ric	تانسور ریچی
L	مشتق لی
Φ	۲-فرم اساسی خمینه تماسی
∇	التصاق لوی-چویتای
Δ	لاپلاسین ناهموار
∇^*	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
g_s	متر ساساکی
∇	التصاق لوی-چویتای وابسته به متر ساساکی
Δ	عملگر لاپلاس-بلترامی روی p-فرم‌ها

فصل ۱

مقدمه

۱-۱ مقدمه

رگرسیون خطی به‌عنوان یکی از پایه‌ای‌ترین اجزای یادگیری ماشین شناخته می‌شود و روشی ساده اما قدرتمند برای مدل‌سازی روابط میان متغیرها ارائه می‌دهد. با این حال، انتخاب الگوریتم مورد استفاده برای حل یک مسئله رگرسیون چه به‌صورت تحلیلی و چه تکرارشونده تأثیر چشمگیری بر کارایی، پایداری و مقیاس‌پذیری دارد. این گزارش با پیاده‌سازی و تحلیل چهار روش متمایز حل، یعنی معادله نرمال [۱]، تجزیه مقدار منفرد [۲]، گرادیان نزولی دسته‌ای [۳] و گرادیان نزولی تصادفی [۴]، به بررسی این موازنه‌ها می‌پردازد.

این مطالعه در سه مرحله اصلی سازمان‌دهی شده است. در مرحله نخست، مبانی محاسباتی رگرسیون را بر روی یک مجموعه‌داده مصنوعی کوچک بررسی می‌کنیم. تمرکز این بخش بر پایداری عددی است و به‌طور خاص رفتار روش مستقیم معادله نرمال و روش SVD را در مواجهه با هم‌خطی چندگانه (ویژگی‌های به‌شدت هم‌بسته) مقایسه می‌کند. این تحلیل محدودیت‌های وارون‌سازی سنتی ماتریس‌ها را در شرایطی که داده‌ها به‌خوبی شرطی نشده‌اند، برجسته می‌سازد.

در مرحله دوم، به مقیاس‌پذیری محاسباتی با استفاده از مجموعه‌داده واقعی Auto MPG می‌پردازیم. با افزایش اندازه داده‌ها، راه‌حل‌های تحلیلی از نظر محاسباتی بسیار پرهزینه می‌شوند. در این بخش،

[۱] (Normal Equation)
[۲] (Singular Value Decomposition یا SVD)
[۳] (Batch Gradient Descent یا BGD)
[۴] (Stochastic Gradient Descent یا SGD)

رفتار همگرایی گرادیان نزولی دسته‌ای را با گرادیان نزولی تصادفی مقایسه می‌کنیم تا نشان دهیم چرا روش‌های تصادفی به استاندارد اصلی در وظایف داده‌کاوی مقیاس بزرگ تبدیل شده‌اند.

در نهایت، انعطاف‌پذیری مدل را با اعمال SGD بر یک مسئله غیرخطی بررسی می‌کنیم. با گسترش مدل خطی از طریق ویژگی‌های چندجمله‌ای، نشان می‌دهیم که چگونه الگوریتم‌های خطی می‌توانند برای برازش توزیع‌های داده خمیده سازگار شوند. این بخش بر نقش حیاتی مهندسی ویژگی و نرمال‌سازی در زنجیره‌های مدرن یادگیری ماشین تأکید دارد.

فصل ۲

مروری بر ادبیات

۲-۱ تعاریف مفاهیم پایه

۲-۱-۱ معادلات نرمال

معادلات نرمال[۱] یک راه‌حل تحلیلی و مستقیم برای مسئله کمترین مربعات خطی[۲] ارائه می‌دهند [۱، ۲]. در این روش، به‌جای بهینه‌سازی تکرارشونده، می‌توان مستقیماً سیستم معادلات خطی

$$\|A\theta - y\|^{۲}$$

را که خطای مربعی را کمینه می‌کند، حل کرد. در این‌جا θ بردار ضرایب مدل است. با صفر قرار دادن گرادیان تابع هزینه، دستگاه معادلات نرمال به‌صورت زیر به‌دست می‌آید:

$$A^T A\theta = A^T y \qquad \Rightarrow \qquad \theta = (A^T A)^{-۱} A^T y$$

این روش برای حالتی که تعداد ویژگی‌ها d کم باشد بسیار سریع است، اما دارای دو مشکل محاسباتی اساسی است [۲]:

۱. هزینه محاسبه معکوس ماتریس $A^T A$ از مرتبه $\mathcal{O}(d^{۳})$ است که برای d بزرگ بسیار سنگین

[۱] (Normal Equations)

[۲] (Ordinary Least Squares یا OLS)

می‌شود.

۲. اگر ویژگی‌ها هم‌خطی باشند، ماتریس A^TA ممکن است تکین (Singular) یا بدحالت (ill-conditioned) شود که در این صورت معکوس آن از نظر عددی ناپایدار یا حتی غیرممکن خواهد بود.

۲-۱-۲ تجزیه مقدار منفرد و شبه‌معکوس

روشی بسیار پایدارتر برای حل مستقیم رگرسیون خطی، استفاده از تجزیه مقدار منفرد [۳] و شبه‌معکوس[۴] است [۲]. شبه‌معکوس مور□پنروز[۵] به‌صورت زیر تعریف می‌شود [۳]:

$$\theta = A^{+}y$$

اگر تجزیه SVD ماتریس A به‌شکل

$$A = U\Sigma V^T$$

باشد، آنگاه شبه‌معکوس آن برابر است با:

$$A^{+} = V\Sigma^{+}U^T$$

این روش حتی در صورتی که ماتریس A^TA تکین باشد (یعنی هم‌خطی کامل بین ویژگی‌ها وجود داشته باشد)، یک راه‌حل یکتا و بهینه با کمترین نرم ارائه می‌دهد و از نظر عددی بسیار پایدار است. به همین دلیل، استفاده از SVD به‌عنوان استانداردترین روش محاسباتی برای حل رگرسیون خطی در نظر گرفته می‌شود [۲].

[۳] (Singular Value Decomposition یا SVD)

[۴] (Pseudoinverse)

[۵] (Moore,Penrose Pseudoinverse)

۲-۱-۳ گرادیان نزولی دسته‌ای

گرادیان نزولی دسته‌ای [۶] یک روش بهینه‌سازی تکرارشونده است که با شروع از یک مقدار اولیه برای θ، در خلاف جهت گرادیان تابع هزینه حرکت می‌کند [۴]. در هر گام، گرادیان تابع هزینه $J(\theta)$ محاسبه شده و پارامترها به‌صورت گام‌به‌گام به سمت کمینه حرکت می‌کنند.

در این روش، برای محاسبه گرادیان در **هر گام**، از **تمام** داده‌های آموزشی استفاده می‌شود:

$$\nabla J(\theta) = \frac{۱}{m} A^T (A\theta - y)$$

زمانی که تعداد نمونه‌ها n بسیار بزرگ باشد، این روش از نظر محاسباتی بسیار کند خواهد بود، زیرا هر به‌روزرسانی مستلزم پردازش کل مجموعه‌داده است [۴].

۲-۱-۴ گرادیان نزولی تصادفی

گرادیان نزولی تصادفی [۷] یک راه‌حل مقیاس‌پذیر برای غلبه بر محدودیت‌های گرادیان نزولی دسته‌ای است [۵]. در این روش، به‌جای استفاده از کل داده‌ها، در هر گام تنها **یک نمونه** به‌صورت تصادفی انتخاب شده و گرادیان بر اساس همان نمونه محاسبه می‌شود:

$$\nabla J_i(\theta) = (A_i\theta - y_i) A_i^T$$

اگرچه مسیر همگرایی SGD به دلیل نویز، نوسانات زیادی دارد، اما به دلیل هزینه محاسباتی بسیار کم هر به‌روزرسانی، در مجموع سریع‌تر از BGD به جواب نزدیک می‌شود. به همین دلیل، SGD به استاندارد طلایی برای آموزش مدل‌های یادگیری ماشین مقیاس بزرگ (Large-scale Machine Learning) تبدیل شده است [۵].

۲-۱-۵ پایداری عددی

پایداری عددی [۸] به توانایی یک الگوریتم در تولید نتایج دقیق و قابل اعتماد در حضور خطاهای کوچک محاسباتی، مانند خطاهای گرد کردن در محاسبات ممیز شناور، اطلاق می‌شود [؟]. الگوریتم‌های ناپایدار،

[۶] (Batch Gradient Descent یا BGD)

[۷] (Stochastic Gradient Descent یا SGD)

[۸] (Numerical Stability)

مانند معکوس کردن ماتریس‌های بدحالت، می‌توانند این خطاهای کوچک را تقویت کرده و به نتایج کاملاً نادرست منجر شوند. ازاین‌رو، روش‌هایی مانند SVD که پایداری عددی بالایی دارند، در مسائل رگرسیون و یادگیری ماشین ترجیح داده می‌شوند [۲].

فصل ۳

روش شناسی و دیتاست

۳-۱ روش‌شناسی

این مطالعه چهار رویکرد متمایز برای حل مسئله رگرسیون خطی را پیاده‌سازی و مقایسه می‌کند که با مدل

$$y = A\theta + \epsilon$$

تعریف می‌شود. روش‌شناسی به دو دسته اصلی تقسیم می‌شود: راه‌حل‌های تحلیلی مستقیم و تکنیک‌های بهینه‌سازی تکرارشونده.

۳-۱-۱ روش‌های حل مستقیم

دو روش تحلیلی برای محاسبه مستقیم بردار پارامترها (θ) پیاده‌سازی شد:

معادلات نرمال معادلات نرمال[۱] با صفر قرار دادن مشتق تابع هزینه $J(\theta)$ آن را کمینه می‌کنند. راه‌حل به‌صورت

$$\theta = (A^T A)^{-1} A^T y$$

به‌دست می‌آید. اگرچه این روش از نظر ریاضی دقیق است، اما نیازمند محاسبه وارون ماتریس $A^T A$

[۱] (Normal Equations)

بوده که از نظر محاسباتی پرهزینه (از مرتبه $\mathcal{O}(n^{۳})$) است و در صورتی که ماتریس تکین یا بدشرط (ill-conditioned) باشد، از نظر عددی ناپایدار می‌شود.

تجزیه مقدار منفرد برای رفع مشکل ناپایداری، از تجزیه مقدار منفرد [۲] جهت محاسبه شبه‌معکوس (A^{+}) استفاده شد. در این روش، ماتریس به‌صورت

$$A = U\Sigma V^T$$

تجزیه شده و سپس

$$\theta = A^{+}y = V\Sigma^{+}U^T y$$

محاسبه می‌شود. با مدیریت صریح مقادیر منفرد نزدیک به صفر، SVD حتی در شرایطی که $A^T A$ به دلیل هم‌خطی چندگانه غیرقابل وارون‌سازی باشد، یک راه‌حل پایدار با «کمترین نرم» ارائه می‌دهد.

۳-۱-۲ روش‌های بهینه‌سازی تکرارشونده

برای یادگیری مقیاس‌پذیر، الگوریتم‌های گرادیان نزولی پیاده‌سازی شدند که به‌صورت تکراری θ را برای کمینه‌سازی خطای میانگین مربعات [۳] به‌روزرسانی می‌کنند:

گرادیان نزولی دسته‌ای گرادیان نزولی دسته‌ای [۴] در هر گام، گرادیان را با استفاده از کل مجموعه‌داده (شامل m نمونه) محاسبه می‌کند:

$$\nabla J(\theta) = \frac{۱}{m} A^T (A\theta - y)$$

این روش مسیر همگرایی همواری را تضمین می‌کند، اما در هر تکرار به حافظه و محاسبات قابل توجهی نیاز دارد.

[۲] (Singular Value Decomposition یا SVD)

[۳] (Mean Squared Error یا MSE)

[۴] (Batch Gradient Descent یا BGD)

گرادیان نزولی تصادفی در گرادیان نزولی تصادفی [۵]، گرادیان با استفاده از یک نمونه تصادفی $(x^{(i)}, y^{(i)})$ در هر گام تقریب زده می‌شود:

$$\nabla J_i(\theta) = (x^{(i)}\theta - y^{(i)})x^{(i)T}$$

اگرچه این کار باعث ایجاد نوسان در مسیر همگرایی می‌شود، اما هزینه محاسباتی هر به‌روزرسانی را به‌طور چشمگیری کاهش داده و پیشرفت سریع‌تری را در مجموعه‌داده‌های بزرگ ممکن می‌سازد.

۳-۱-۳ آماده‌سازی داده و مهندسی ویژگی

مجموعه‌داده مصنوعی یک مجموعه‌داده کوچک با ابعاد ۱×۴ برای بررسی پایداری عددی ایجاد شد. هم‌خطی چندگانه به‌صورت مصنوعی و با افزودن یک نسخه نویزی از ستون ویژگی $(x_۲ \approx x_۱)$ القا شد که منجر به ایجاد یک ماتریس طراحی تقریباً تکین گردید.

مجموعه‌داده Auto MPG داده‌های دنیای واقعی با حذف مقادیر گمشده پردازش شدند و ویژگی Horsepower برای پیش‌بینی MPG استخراج گردید.

نرمال‌سازی برای الگوریتم‌های گرادیان نزولی (بخش ۲) و رگرسیون چندجمله‌ای (بخش ۳)، ویژگی‌ها با استفاده از استانداردسازی (Z-score Normalization) به میانگین صفر و واریانس واحد تبدیل شدند. این کار از غالب شدن ویژگی‌هایی با مقیاس بزرگ (مانند $x^۲$) بر گرادیان و ناپایدار شدن الگوریتم جلوگیری می‌کند.

ویژگی‌های چندجمله‌ای برای مدل‌سازی روابط غیرخطی، فضای ویژگی با افزودن جمله‌های درجه دوم $(x^۲)$ گسترش داده شد و مسئله به شکل زیر درآمد:

$$y = \theta_۰ + \theta_۱ x + \theta_۲ x^۲$$

[۵] Stochastic Gradient Descent یا SGD)

فصل ۴

نگاهی عمیق تر

۴-۱ نتایج و تحلیل

این بخش یافته‌های عددی حاصل از سه مرحله آزمایشی را ارائه می‌کند و پایداری، مقیاس‌پذیری و انعطاف‌پذیری الگوریتم‌های رگرسیون پیاده‌سازی‌شده را مورد تحلیل قرار می‌دهد.

۴-۱-۱ پایداری محاسباتی (مجموعه‌داده‌های کوچک)

در آزمایش اولیه بر روی یک مجموعه‌داده تمیز با ابعاد ۱ × ۴، سه الگوریتم بردارهای پارامتر θ زیر را تولید کردند:

- **معادلات نرمال:**

$$\begin{bmatrix} -۱٫۰۴ \\ ۲٫۰۳ \end{bmatrix}$$

- **روش SVD:**

$$\begin{bmatrix} -۱٫۰۴ \\ ۲٫۰۳ \end{bmatrix}$$

- **گرادیان نزولی دسته‌ای (BGD):**

$$\begin{bmatrix} -۰٫۳۸ \\ ۱٫۸۷ \end{bmatrix}$$

تحلیل روش‌های معادلات نرمال و SVD راه‌حل‌های یکسان و دقیقی ارائه دادند که اعتبار آن‌ها را به‌عنوان حل‌کننده‌های تحلیلی مستقیم تأیید می‌کند. در مقابل، گرادیان نزولی دسته‌ای یک جواب تقریبی تولید کرد (به‌طور مثال $\theta_۱ \approx ۱٫۸۷$ در مقایسه با مقدار دقیق ۲٫۰۳). این اختلاف یکی از ویژگی‌های کلیدی روش‌های تکرارشونده را نشان می‌دهد: این روش‌ها به صفر تحلیلی دقیق نمی‌رسند، بلکه به‌تدریج به آن همگرا می‌شوند. در نتیجه، برای دستیابی به دقت روش‌های مستقیم، BGD به تعداد تکرار بیشتر یا تنظیم مناسب‌تر ابرپارامترها نیاز دارد.

تأثیر هم‌خطی چندگانه با وارد کردن هم‌خطی شدید (افزودن یک ویژگی تکراری همراه با نویز)، پایداری حل‌کننده‌ها مورد آزمون قرار گرفت. ضرایب حاصل به‌صورت زیر بودند:

- **معادلات نرمال:**

$$\theta \approx \begin{bmatrix} ۰٫۱۰۷ \\ ۱۰۴۱۰ \\ -۱۰۴۰۸ \end{bmatrix}$$

- **روش SVD:**

$$\theta \approx \begin{bmatrix} ۰٫۱۰۷ \\ ۱۰۴۱۰ \\ -۱۰۴۰۸ \end{bmatrix}$$

تحلیل هر دو روش پدیده «انفجار ضرایب» را نشان دادند. وزن‌های مربوط به ویژگی‌های هم‌بسته ($x_۱$ و $x_۲$) به مقادیری با قدر مطلق بسیار بزرگ و با علامت‌های مخالف (در حدود ۱۰,۰۰۰ و ۱۰,۰۰۰−) تبدیل شدند. از دیدگاه ریاضی، مدل تلاش می‌کند با استفاده از تفاضل دو ستون تقریباً یکسان، نویز داده را برازش کند که این امر به بیش‌برازش [۱] منجر می‌شود. اگرچه SVD از نظر تئوری قادر است با حذف مقادیر منفرد کوچک این مشکل را برطرف کند، اما در پیاده‌سازی استاندارد (بدون آستانه‌گذاری دستی)، تفاوت‌های بسیار کوچک ناشی از نویز به‌عنوان اطلاعات معتبر در نظر گرفته شده و در نتیجه همان ناپایداری معادلات نرمال ایجاد شده است. این موضوع ضرورت استفاده از منظم‌سازی (مانند رگرسیون ریج) یا حذف دستی مقادیر منفرد کوچک در داده‌های هم‌خط واقعی را نشان می‌دهد.

[۱] (overfitting)

۴-۱-۲ مقیاس‌پذیری محاسباتی (مجموعه‌داده‌های بزرگ)

در مجموعه‌داده Auto MPG، رفتار همگرایی گرادیان نزولی دسته‌ای (BGD) و گرادیان نزولی تصادفی (SGD) مقایسه شد:

- **هزینه نهایی BGD**: ۱۱٫۹۷۱۸
- **هزینه نهایی SGD**: ۱۲٫۳۲۴۸

تحلیل BGD به هزینه نهایی اندکی پایین‌تر دست یافت که قابل انتظار است، زیرا در هر گام گرادیان دقیق را با استفاده از کل داده‌ها محاسبه می‌کند و می‌تواند دقیقاً روی کمینه قرار گیرد. در مقابل، SGD هزینه نهایی کمی بالاتر (حدود +۰٫۳۵) تولید کرد که بازتاب ماهیت تصادفی آن است؛ این الگوریتم به‌جای قرار گرفتن دقیق روی کمینه سراسری، در اطراف آن نوسان می‌کند. با این حال، از منظر مقیاس‌پذیری، این کاهش جزئی در دقت کاملاً قابل قبول است. برای مجموعه‌داده‌های بسیار بزرگ، SGD یک راه‌حل «به‌اندازه کافی خوب» را چندین برابر سریع‌تر از زمانی که BGD برای تنها یک گام به آن نیاز دارد، ارائه می‌دهد.

۴-۱-۳ کاربرد در مدل‌های غیرخطی

در نهایت، الگوریتم SGD بر روی یک مدل چندجمله‌ای درجه دوم

$$y = \theta_۰ + \theta_۱ x + \theta_۲ x^۲$$

با ویژگی‌های نرمال‌شده اعمال شد. پارامترهای آموخته‌شده به‌صورت زیر بودند:

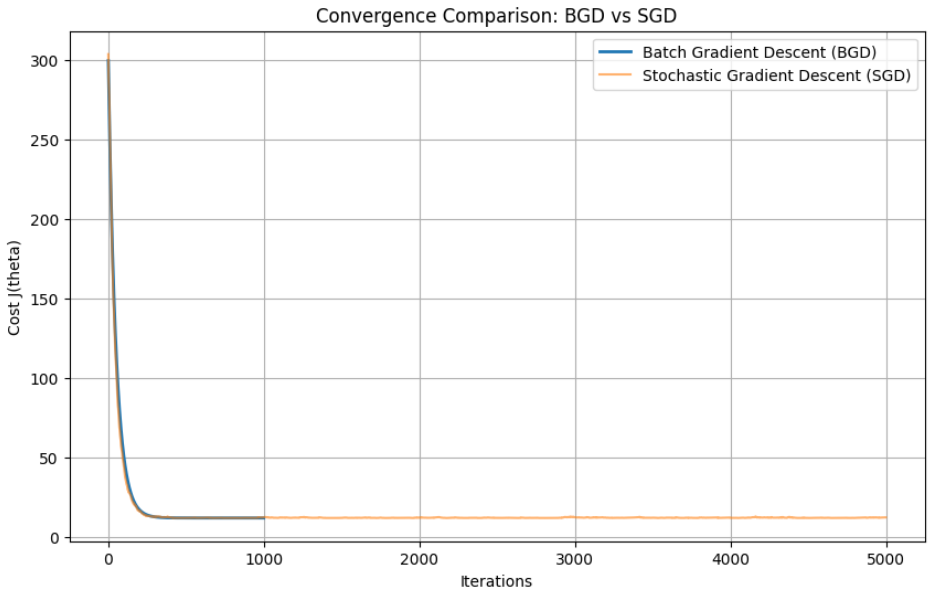
$$\theta = \begin{bmatrix} ۲۳٫۴۳ \\ -۱۵٫۲۵ \\ ۹٫۵۹ \end{bmatrix}$$

تفسیر

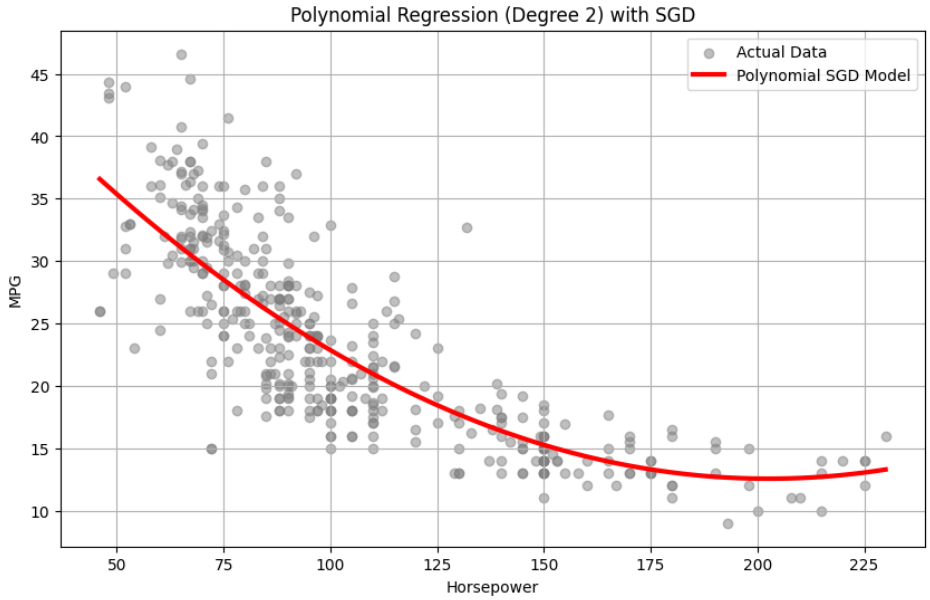
- **بایاس** ($\theta_۰ = ۲۳٫۴۳$): نمایانگر مقدار پایه MPG برای یک خودروی متوسط در مجموعه‌داده نرمال‌شده است.

- **جمله خطی** ($\theta_۱ = -۱۵٫۲۵$): ضریب منفی بزرگ نشان‌دهنده رابطه معکوس قوی است؛ با افزایش Horsepower، مقدار MPG به‌طور قابل توجهی کاهش می‌یابد.

- **جمله درجه دوم** ($\theta_۲ = ۹٫۵۹$): ضریب مثبت بیانگر انحنای محدب است. این بدان معناست که اگرچه MPG با افزایش Horsepower کاهش می‌یابد، اما نرخ این کاهش برای خودروهای بسیار پرقدرت کمتر می‌شود (بازده نزولی)، و در نتیجه منحنی حاصل شکلی خمیده و تخت‌شونده دارد.



شکل ۴-۱: مقایسه نتایج BGD و SGD



شکل ۴-۲: نتایج Polynomial Regression با SGD

فصل ۵

جمع‌بندی و نتیجه‌گیری و پیشنهادات

۵-۱ نتیجه‌گیری

این مطالعه به‌طور جامع موازنه‌ها میان روش‌های تحلیلی و تکرارشونده در رگرسیون خطی را ارزیابی کرد و مزایای متمایز هر یک را در سناریوهای محاسباتی مختلف برجسته ساخت.

یافته‌های کلیدی

پایداری عددی تحلیل ما بر روی مجموعه‌داده‌های کوچک نشان داد که روش‌های مستقیم مانند معادله نرمال برای داده‌های با بُعد کم از نظر محاسباتی کارآمد هستند، اما در برابر هم‌خطی چندگانه بسیار آسیب‌پذیرند. زمانی که ویژگی‌ها به‌شدت هم‌بسته باشند، وارون‌سازی ماتریس ناپایدار شده و به پدیده انفجار ضرایب منجر می‌شود. اگرچه تجزیه مقدار منفرد (Singular Value Decomposition یا SVD) از نظر تئوری از طریق شبه‌معکوس راه‌حلی پایدار ارائه می‌دهد، اما در عمل نیازمند حذف یا برش دقیق مقادیر منفرد کوچک است تا از بروز مشکلات بیش‌برازش مشابه جلوگیری شود.

مقیاس‌پذیری در زمینه داده‌کاوی مقیاس بزرگ، گرادیان نزولی تصادفی (Stochastic Gradient Descent یا SGD) به‌عنوان رویکرد برتر ظاهر شد. با وجود آنکه گرادیان نزولی دسته‌ای (Batch Gradient Descent یا BGD) به هزینه نهایی اندکی کمتر دست یافت (۱۱٫۹۷ در مقابل ۱۲٫۳۲)، الزام آن به پردازش کل مجموعه‌داده در هر به‌روزرسانی، این روش را برای داده‌های عظیم از نظر محاسباتی غیرعملی می‌سازد. توانایی SGD در به‌روزرسانی آنی پارامترها با سربار حافظه‌ای ناچیز، آن را ـ علیرغم مسیر همگرایی

پرنوسان ـ به استاندارد اصلی یادگیری ماشین مدرن تبدیل کرده است.

انعطاف‌پذیری مدل کاربرد SGD در رگرسیون چندجمله‌ای نشان داد که الگوریتم‌های خطی محدود به داده‌های خطی نیستند. با مهندسی مناسب ویژگی‌ها (مانند افزودن جمله $x^۲$) و اعمال نرمال‌سازی صحیح، توانستیم رابطه غیرخطی میان Horsepower و MPG را با موفقیت مدل‌سازی کنیم و پدیده بازده نزولی مصرف سوخت در خودروهای پرقدرت را به‌خوبی ثبت نماییم.

توصیه نهایی

برای مجموعه‌داده‌های کوچک و تمیز، روش‌های تحلیلی مبتنی بر SVD به دلیل سرعت و دقت بالا گزینه مناسب‌تری هستند. اما برای مسائل دنیای واقعی، مقیاس بزرگ یا غیرخطی، گرادیان نزولی تصادفی (به‌همراه نرمال‌سازی ویژگی‌ها) مقاوم‌ترین و مقیاس‌پذیرترین انتخاب است که توازن مناسبی میان کارایی محاسباتی و توان پیش‌بینی برقرار می‌کند.

منابع و مراجع

[1] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2 ed. , 2009.

[2] Golub, Gene H. and Van Loan, Charles F. Matrix Computations. Johns Hopkins University Press, Baltimore, 4 ed. , 2013.

[3] Penrose, Roger. A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society, 51(3):406–413, 1955.

[4] Boyd, Stephen and Vandenberghe, Lieven. Convex Optimization. Cambridge University Press, Cambridge, 2004.

[5] Bottou, Léon, Curtis, Frank E., and Nocedal, Jorge. Optimization methods for large-scale machine learning. SIAM Review, 60(2):223–311, 2018.

Abstract

This study explores the computational foundations and scalability of linear regression algorithms, specifically comparing the Normal Equation, Singular Value Decomposition (SVD), Batch Gradient Descent (BGD), and Stochastic Gradient Descent (SGD). Using a synthetic small dataset, we first demonstrate that while the Normal Equation provides an exact solution, it becomes numerically unstable under conditions of multicollinearity. In contrast, SVD proves robust by effectively handling singular matrices through pseudoinverse calculation. Subsequently, we analyze scalability using the real-world Auto MPG dataset. Our results show that while BGD offers smooth convergence, it is computationally expensive for large datasets. SGD, despite its noisy convergence path, provides a far more efficient alternative for large-scale applications by updating parameters iteratively. Finally, we successfully apply SGD to a polynomial regression model, demonstrating that linear algorithms can capture nonlinear relationships through feature engineering, provided that proper data normalization is applied.

Key Words:

Linear Regression Algorithms, Gradient Descent Optimization, Numerical Stability, Multicollinearity, Polynomial Regression

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Comptuer Science

M. Sc. Thesis

Third CDM Project

By

Mohammad Sadegh Gholizadeh

Supervisor

Dr. Mahdi Ghatei