دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

داده کاوی محاسباتی

گزارش ۳ (مبانی محاسباتی رگرسیون (پایداری و مقیاس پذیری))

نگارش
محمدامین نعمتی

شماره دانشجویی
۴۰۴۱۲۴۹۰۴

استاد درس
آقای دکتر قطعی

نیمسال اول ۱۴۰۵-۱۴۰۴

چکیده

در این پروژه روش‌های مختلف حل رگرسیون خطی بررسی و مقایسه شده‌اند. نتایج نشان می‌دهد روش مبتنی بر معادلات نرمال در برابر هم‌خطی ناپایدار است، در حالی‌که تجزیه مقادیر منفرد پاسخی پایدار ارائه می‌کند. همچنین روش‌های کاهشی، به‌ویژه نوع تصادفی، برای داده‌های حجیم کارآمدتر و سریع‌تر هستند. در نهایت، ساخت مدل چندجمله‌ای درجه دو نشان داد که با نرمال‌سازی مناسب می‌توان رفتارهای غیرخطی را نیز به‌خوبی مدل‌سازی کرد.

واژه‌های کلیدی:

رگرسیون خطی، هم‌خطی، تجزیه مقادیر منفرد، روش‌های کاهشی

فهرست مطالب

صفحه

چکیده..... أ

فصل اول مقدمه..... ۱

فصل دوم مبانی محاسباتی..... ۵

فصل سوم مقایسه محاسباتی BGD در مقابل SGD..... ۷

فصل چهارم تحلیل مقیاس پذیری..... ۱۰

فصل پنجم کاربرد (پیاده‌سازی رگرسیونِ غیرخطی)..... ۱۳

فصل ششم نتیجه‌گیری..... ۱۷

Abstract..... ۱۹

فصل اول

مقدمه

مقدمه

رگرسیون خطی یکی از بنیادی‌ترین ابزارهای تحلیل داده و یادگیری ماشینی به شمار می‌آید و شیوه حل آن نقش مهمی در دقت و پایداری نتایج دارد. با گسترش داده‌ها و پیچیده‌تر شدن ساختار آن‌ها، انتخاب روش مناسب برای برآورد پارامترها بیش از گذشته اهمیت یافته است. در این پروژه، ابتدا پایه‌های محاسباتی چند روش اصلی حل رگرسیون خطی بررسی شده و رفتار آن‌ها در شرایط مختلف، از جمله وجود هم‌خطی، تحلیل می‌شود، سپس با استفاده از داده‌های واقعی، کارایی روش‌های تکراری کاهشی از نظر سرعت، هزینه محاسباتی و کیفیت همگرایی سنجیده شده و در ادامه به کاربرد مدل‌های چندجمله‌ای برای بازنمایی الگوهای غیرخطی پرداخته می‌شود. هدف کلی این پژوهش، ارائه درکی روشن از نقاط قوت و ضعف هر رویکرد و تبیین معیارهای انتخاب روش مناسب بر اساس وضعیت داده و هدف تحلیل است.

فصل دوم

مبانی محاسباتی

مبانی محاسباتی

۱ .مقایسهٔ روش‌ها در حالت بدون هم‌خطی(Non-Collinear)

پس از اجرای سه روش، مقادیر به‌دست‌آمده برای پارامترهای مدل در جدول (۱–۱) گزارش شده‌اند. همان‌طور که مشاهده می‌شود، دو روش معادلات نرمال و SVD دقیقاً یک پاسخ تولید کرده‌اند؛ زیرا هر دو روش در این حالت کوچک از نظر محاسباتی دچار مشکل عددی نمی‌شوند. روش گرادیان‌کاهشی اگرچه با روندی همگرا به سمت مقدار صحیح حرکت می‌کند، اما به‌دلیل نرخ یادگیری و تعداد تکرار محدود، هنوز به پاسخ دقیق نرسیده و انحراف کوچکی نسبت به دو روش دیگر دارد.

جدول ۱–۱ مقایسهٔ بردار پارامترها در سه روش (بدون هم‌خطی)

Parameter	Normal Equations	Batch GD	SVD
θ_0	-1.041379	-0.552178	-1.041379
θ_1	2.031034	1.914683	2.031034

۲ .ایجاد هم‌خطی قوی(Strong Collinearity)

برای بررسی پایداری عددی، یک ویژگی جدید ساخته شد که تقریباً کپی دقیق ستون دوم ویژگی‌هاست؛ اما با نویزی بسیار کوچک؛ این کار باعث شد ماتریس طراحی سه‌ستونه جدید تقریباً دچار هم‌خطی کامل شود و شرایطی ایجاد گردد که در آن روش معادلات نرمال انتظار می‌رود دچار ناپایداری شود.

نتایج حاصل از اعمال معادلات نرمال و SVD بر روی این ماتریس هم‌خط در جدول (۲–۱) ارائه شده است. همان‌گونه که در جدول دیده می‌شود، روش معادلات نرمال در مواجهه با هم‌خطی، مقادیر غیرواقعی بسیار بزرگ تولید می‌کند. دلیل این رفتار، تکین‌شدن ماتریس و در نتیجه ناپایداری شدید هنگام وارون‌گیری آن است. در مقابل، روش SVD حتی در حضور هم‌خطی شدید، یک جواب پایدار و معنادار ارائه می‌دهد؛ زیرا به‌جای وارون‌گیری مستقیم، از شبه‌وارون پایدار (Pseudo-inverse) استفاده می‌کند.

جدول ۲–۱ مقایسهٔ پارامترهای برآوردشده در حضور هم‌خطی شدید

Method	θ_0	θ_1	θ_2
Normal Equations	-0.849826	1.353459×10^5	-1.353439×10^5
SVD	-1.080676	1.328667×10^5	-1.328646×10^5

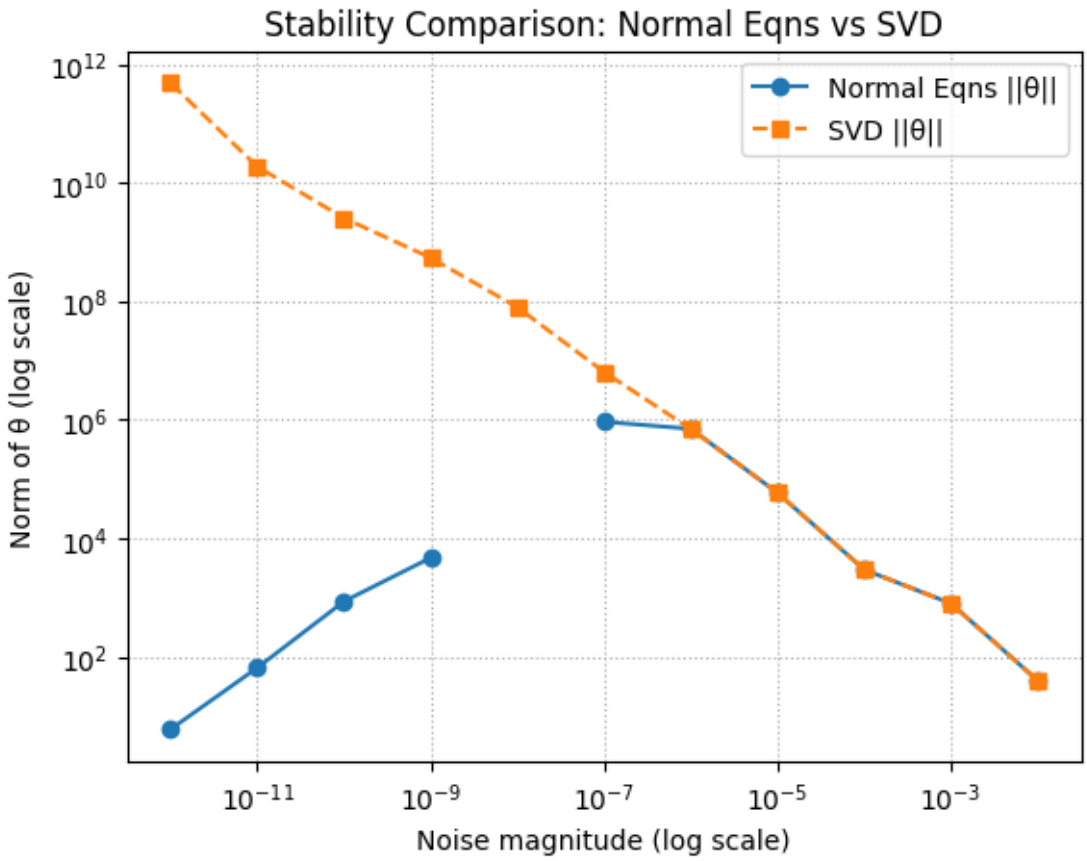
همان‌طور که مشاهده می‌شود، هر دو روش مقدارهای بزرگ برای θ_1 و θ_2 تولید کرده‌اند.

۳ .تحلیل پایداری عددی و تفسیر نمودار

برای بررسی دقیق‌تر پایداری، مقدار نویز کنترل‌شده‌ای به ستون هم‌خط اضافه شد و بزرگی نورم پارامترهای تخمینی در هر دو روش برای ۱۰ سطح مختلف نویز محاسبه گردید. نمودار حاصل در **شکل ۱–۱** ترسیم شد.

این شکل رفتار دو روش را در مقیاس لگاریتمی نشان می‌دهد. منحنی مربوط به معادلات نرمال با کاهش نویز به‌سرعت ناپایدار شده و در برخی سطوح نویز حتی به‌دلیل تکین‌شدن ماتریس، مقدار محاسبه‌نشده (NaN) تولید می‌کند که باعث قطع‌شدن منحنی در نمودار می‌شود. این پدیده یکی از نشانه‌های مشخص ضعف روش معادلات نرمال در برابر هم‌خطی است.

از سوی دیگر، منحنی SVD در تمام سطوح نویز پایدار باقی می‌ماند و مقادیر کنترل‌شده‌ای تولید می‌کند. دلیل این ویژگی آن است که در روش SVD ، مقادیر منفرد کوچک (small singular values) که عامل اصلی ناپایداری هستند، در فرآیند شبه‌وارون به‌گونه‌ای مدیریت می‌شوند که باعث انفجار عددی نشوند. به‌همین دلیل SVD همواره توصیه‌شده‌ترین روش برای داده‌های دارای هم‌خطی یا شرایط ill-conditioned است.



شکل ۱–۱ مقایسهٔ پایداری روش معادلات نرمال و SVD تحت نویزهای مختلف

محور افقی بزرگی نویز و محور عمودی نرم پارامترها در مقیاس لگاریتمی

فصل سوم

مقایسه محاسباتی BGD در مقابل SGD

مقایسه محاسباتی BGD در مقابل SGD

در این بخش، داده‌های مجموعه **Auto MPG** بارگذاری و پس از حذف مقادیر نامعتبر ستون *Horsepower*، ماتریس طراحی برای مدل خطی ساخته شد. جدول زیر نمونه‌ای از داده‌های پاک‌سازی‌شده را نشان می‌دهد:

جدول ۳–۱ نمونه‌ای از داده‌های ورودی

MPG	Cyl	Disp	HP	Weight	Accel	Year	Origin	CarName
18.0	8	307.0	130.0	3504.0	12.0	70	1	chevrolet ...
15.0	8	350.0	165.0	3693.0	11.5	70	1	buick skylark ...
18.0	8	318.0	150.0	3436.0	11.0	70	1	plymouth ...
16.0	8	304.0	150.0	3433.0	12.0	70	1	amc rebel ...
17.0	8	302.0	140.0	3449.0	10.5	70	1	ford torino

در ادامه **ماتریس طراحی A** و **بردار خروجی y** در یک جدول ترکیب شده‌اند:

جدول ۳–۲ ماتریس ترکیبی A و y ۵ سطر اول

HP	MPG
130	18
165	15
150	18
150	16
140	17

پس از اجرای دو روش گرادیان‌کاهشی، هزینه نهایی هر روش مطابق جدول زیر به‌دست آمد:

جدول ۳–۳ مقایسه هزینه نهایی روش‌های BGD و SGD

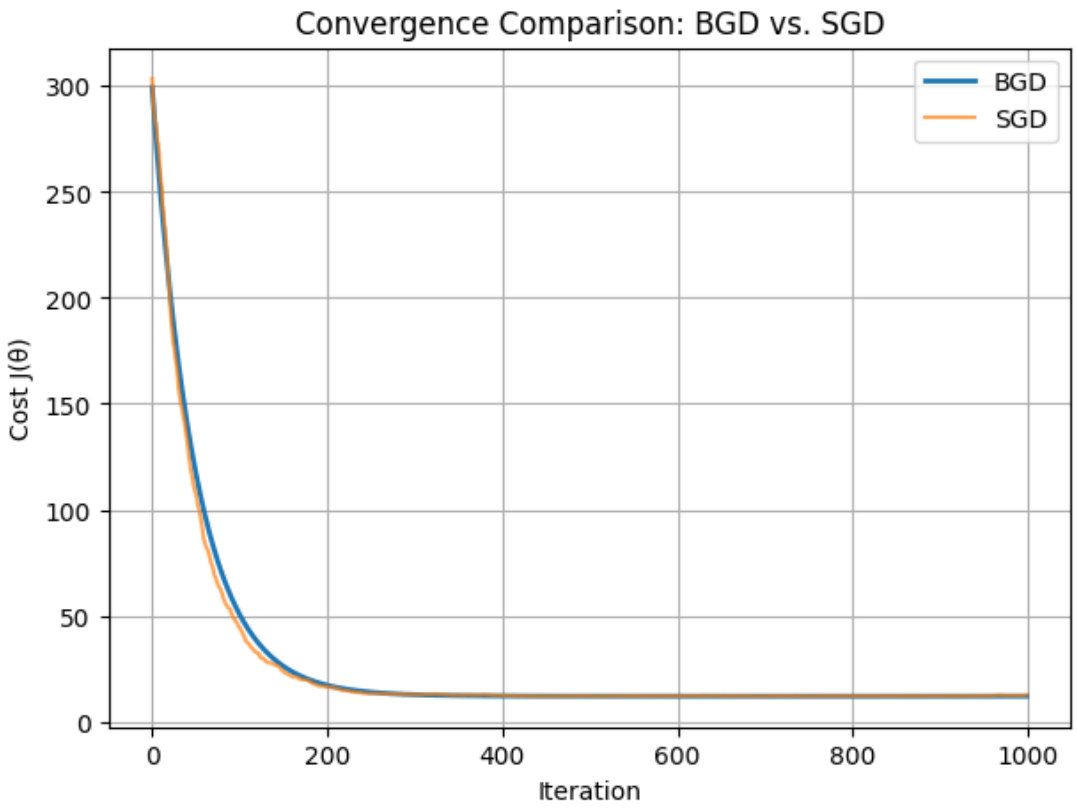
روش	هزینه نهایی
BGD	11.9718
SGD	12.3248

در این مقایسه، BGD به دلیل استفاده از کل داده‌ها در هر به‌روزرسانی، کاهش هزینه‌ای یکنواخت‌تر و مقدار نهایی کمی دقیق‌تر ارائه داده است. در مقابل SGD با وجود نوسان بیشتر، همگرایی سریع‌تری در مراحل ابتدایی داشته و با هزینه محاسباتی کمتر به مقدار قابل قبولی رسیده است.

فصل چهارم

تحلیل مقیاس‌پذیری

تحلیل مقیاس‌پذیری



شکل ۱-۴ مقایسه همگرایی BGD و SGD

نمودار نشان می‌دهد که هر دو روش در نهایت به مقدار مشابهی از تابع هزینه می‌رسند، اما **SGD** به‌دلیل به‌روزرسانی‌های سریع‌تر، کاهش اولیه سریع‌تری نسبت به **BGD** دارد. مسیر **BGD** یکنواخت و بدون نوسان است، زیرا هر تکرار گرادیان کل داده را محاسبه می‌کند، درحالی‌که **SGD** نوسان بیشتری دارد اما زودتر به ناحیه مینیمم نزدیک می‌شود.

در داده‌کاوی محاسباتی و یادگیری ماشینِ مقیاس بزرگ، تقریباً همیشه از **SGD** و یا نسخه **Mini-Batch** آن استفاده می‌شود، زیرا هر به‌روزرسانی فقط نیازمند پردازش یک زیرمجموعه کوچک داده است و هزینه هر تکرار وابسته به اندازه کل داده نیست. این موضوع باعث **مقیاس‌پذیری بالا، سرعت همگرایی بیشتر و امکان استفاده در مجموعه‌داده‌های بسیار بزرگ** می‌شود، در حالی که **BGD** به‌دلیل نیاز به پردازش کل داده در هر تکرار، در عمل برای داده‌های بزرگ قابل استفاده نیست.

فصل پنجم

کاربرد (پیاده‌سازی رگرسیونِ غیرخطی)

کاربرد (پیاده‌سازی رگرسیونِ غیرخطی)

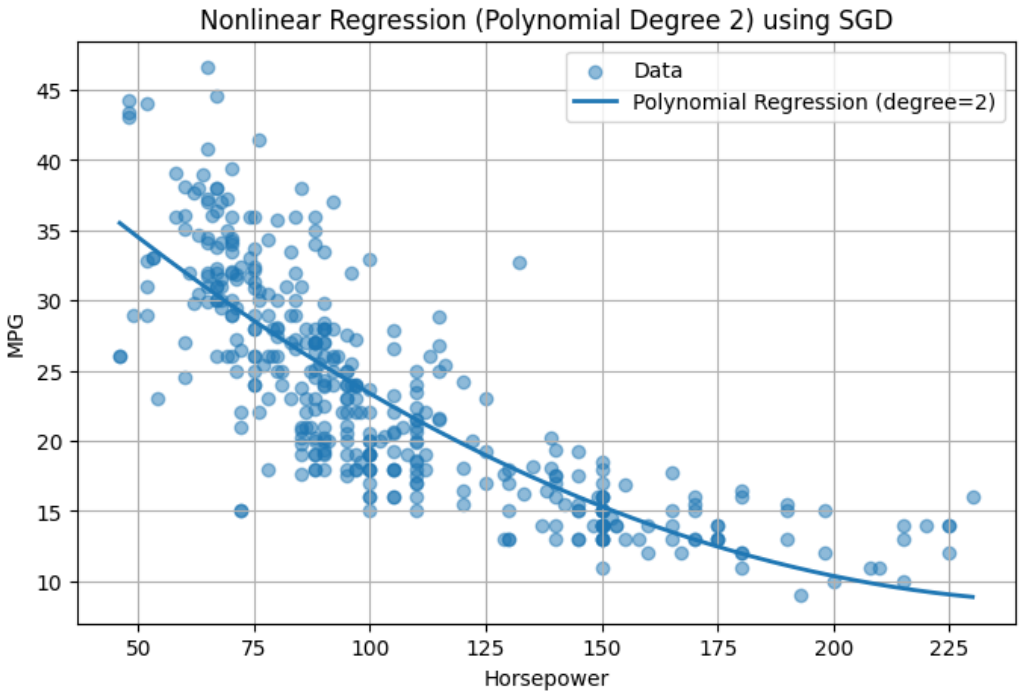
مجموعه داده Auto-MPG با حذف نمونه‌های ناقص بارگذاری شد و ماتریس طراحی درجه دو با نرمال‌سازی ستون‌ها ایجاد گردید که ابعاد آن ۳۹۲ در ۳ می‌باشد. الگوریتم SGD با نرخ یادگیری ۰.۰۱ و ۵۰۰۰ تکرار اجرا شد. در هر تکرار یک نمونه تصادفی انتخاب و ضرایب به‌روزرسانی شدند. مقدار هزینه هر ۱۰ تکرار ثبت شد.

ضرایب نهایی برای عرض از مبدأ، ضریب خطی و ضریب درجه‌دو به شکل جدول زیر هستند:

θ	مقدار
θ_0	23.4913
θ_1	-11.0904
θ_2	5.2666

پس از تولید نقاط Horsepower، ماتریس طراحی استاندارد شده و با ضرب در θ منحنی پیش‌بینی ترسیم شد. منحنی درجه‌دو به‌خوبی الگوی غیرخطی داده را بازتولید می‌کند.

رگرسیون چندجمله‌ای درجه‌دو با SGD ضرایب پایدار و منحنی سازگار با داده تولید کرد. نرمال‌سازی ورودی‌ها و نرخ یادگیری مناسب در همگرایی مدل مؤثر بودند.



شکل ۳–۱ رگرسیون چندجمله‌ای (درجه ۲) با استفاده از SGD

فصل ششم

جمع‌بندی و نتیجه‌گیری

جمع‌بندی و نتیجه‌گیری

بررسی‌های انجام‌شده نشان می‌دهد که روش‌های مختلف حل رگرسیون خطی رفتار یکسانی ندارند و هرکدام در شرایطی خاص بر دیگری برتری می‌یابند. روش مبتنی بر معادلات نرمال تنها در داده‌های خوش‌شرط پایدار است و با ایجاد هم‌خطی دچار ناپایداری می‌شود، در حالی که تجزیه مقادیر منفرد حتی در شرایط بدحالت نیز پاسخی قابل اتکا ارائه می‌کند. از سوی دیگر، روش‌های کاهشی به‌ویژه نوع تصادفی، برای داده‌های حجیم و کاربردهای مقیاس‌پذیر بسیار مناسب هستند و امکان رسیدن سریع به ناحیه بهینه را فراهم می‌کنند. همچنین ساخت مدل چندجمله‌ای نشان داد که با پیش‌پردازش مناسب می‌توان روابط غیرخطی را نیز با دقت مطلوب مدل‌سازی کرد. در مجموع، این پژوهش اهمیت توجه هم‌زمان به ویژگی‌های داده، پایداری عددی و هزینه محاسباتی را در انتخاب روش مناسب برای رگرسیون خطی برجسته می‌سازد.

Abstract

This project examines and compares various methods for solving linear regression. The results show that the normal-equation approach becomes unstable in the presence of multicollinearity, whereas singular-value decomposition provides a stable solution. Moreover, iterative reduction methods—especially the stochastic variant—are more efficient and faster for large datasets. Finally, constructing a second-degree polynomial model demonstrated that, with proper normalization, nonlinear patterns can also be modeled effectively.

Key Words: Linear regression, multicollinearity, singular-value decomposition, iterative reduction methods

GITHUB LINK

https://github.com/nematiamin57-pixel/data_mining1404

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Science

Computational Data Mining

Report 3: Computational Foundations of Regression (Stability and Scalability)

By
Mohammadamin Nemati

Student ID
404124904

Lecturer
Dr. Ghatee

First Semester of the 2025–2026 Academic Year