دانشگاه صنعتي اميركبير
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش پروژه درس داده کاوی محاسباتی پروژه 3

مبانی محاسباتی رگرسیون (پایداری و مقیاس پذیری)

نگارش
هومن ذوالفقاری

استاد راهنما
دکتر مهدی قطعی

تدریس یار
مهندس بهنام یوسفی مهر

آذر 1404

صفحه فرم ارزیابی و تصویب پایان‌نامه- فرم تأیید اعضاي کمیته دفاع

در این صفحه (هر سه مقطع تحصیلي) باید تصویر فرم ارزیابي یا تأیید و تصویب پایان-نامه/رساله موسوم به فرم کمیته دفاع براي مقاطع کارشناسی‌ارشد و دکتری و تصویر فرم تصویب براي مقطع کارشناسي، موجود در پرونده آموزشی را قرار دهند.

نکات مهم:

- ✓ نگارش پایان‌نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان‌نامه‌های دانشگاه صنعتی امیرکبیر باشد. (دستورالعمل و راهنمای حاضر)؛
- ✓ تحویل پایان‌نامه به زبان انگلیسي، براي دانشجویان بین‌الملل با شرایط دستورالعمل حاضر بلامانع است و داشتن صفحه عنوان فارسي به همراه چکیده مبسوط فارسي، 30 صفحه براي پایان‌نامه کارشناسی‌ارشد و50 صفحه براي رساله دکتری در ابتداي آن الزامي است؛
- ✓ دریافت پایان‌نامه کارشناسي،کارشناسي‌ارشد و رساله دکتری، بصورت نسخه دیجیتال مطابق راهنمای وبسایت و دستورالعمل حاضر مي‌باشد؛
- ✓ درصورتي که یک عنوان پایان‌نامه داراي دو نویسنده است، فقط یکبار فایل و فرم اطلاعات آن با ذکر هر دو نویسنده بارگذاري و تکمیل گردد؛
- ✓ با توجه به اینکه در ورود2016یا بالاتر، احتمال تغییر ترتیب ذکر زیر فصل‌ها وجود دارد لطفا در انتها به شماره‌دهي زیر فصل‌ها توجه نمایید که بصورت صحیح باشد. از راست به چپ: شماره فصل-زیرفصل1-زیرفصل2-زیرفصل3 و ... .

به نام خدا

تاریخ:

تعهدنامه اصالت اثر

اینجانب هومن ذوالفقاری متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتي اميركبير بوده و به دستاوردهاي ديگران كه در اين پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذكر گرديده است. اين پايان نامه قبلاً براي احراز هيچ مدرك هم‌سطح يا بالاتر ارائه نگرديده است.

در صورت اثبات تخلف در هر زمان، مدرك تحصيلي صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پيگيري قانوني خواهد داشت.

كليه نتايج و حقوق حاصل از اين پايان نامه متعلق به دانشگاه صنعتي اميركبير مي‌باشد. هرگونه استفاده از نتايج علمي و عملي، واگذاري اطلاعات به ديگران يا چاپ و تكثير، نسخه‌برداري، ترجمه و اقتباس از اين پايان نامه بدون موافقت كتبي دانشگاه صنعتي اميركبير ممنوع است.
نقل مطالب با ذكر مآخذ بلامانع است.

در صفحه تعهدنامه اصالت اثر، در قسمت بالا سمت چپ، تاریخ دفاع خود را جایگزین تاریخ نوشته شده کنید.

همچنین در صفحه تعهدنامه اصالت اثر، در خط اول، نام و نام خانوادگی خود را به صورت کامل با نام و نام خانوادگی نمونه، جایگزین کنید. در انتهای متن تعهد، در قسمت امضا نیز باید نام و نام خانوادگی کامل خود را وارد نماید.

هومن ذوالفقاری

امضا

چکیده

در این پروژه، سه روش مختلف برای حل مسئلهٔ رگرسیون خطی شامل معادلات نرمال، گرادیان کاهشی دسته‌ای (BGD) و تجزیهٔ مقادیر منفرد (SVD) بررسی و مقایسه شدند. پایداری عددی این روش‌ها با استفاده از داده‌های هم‌خط (Collinear) تحلیل شد و نشان داده شد که روش SVD در شرایط ill-conditioned عملکرد به‌مراتب پایدارتر و دقیق‌تری نسبت به معادلات نرمال دارد. سپس با استفاده از داده‌های واقعیAuto MPG ، دو الگوریتم BGD و گرادیان کاهشی تصادفی (SGD)از نظر سرعت همگرایی و کارایی محاسباتی مورد مقایسه قرار گرفتند. نتایج نشان داد که SGD به دلیل پیچیدگی خطی و بروزرسانی‌های سریع، برای داده‌های بزرگ مناسب‌تر است. در نهایت یک مدل رگرسیون چندجمله‌ای درجه ۲ با استفاده از SGD آموزش داده شد و نشان داده شد که مدل غیرخطی می‌تواند رابطهٔ پیچیده بین Horsepower و MPG را بهتر از مدل خطی ساده ثبت کند. این پژوهش اهمیت پایداری عددی، مقیاس‌پذیری محاسباتی و انتخاب مدل مناسب را در مسائل یادگیری ماشین نشان می‌دهد.

واژه‌های کلیدی:

همخطی، داده کاوی، رگرسیون، بهینه سازی

فهرست مطالب

صفحه

فهرست اشکال

صفحه

فهرست جداول

صفحه

فهرست علائم

علائم لاتین

A	ماتریس داده
x	بردار داده
y	خروجی مدل رگرسیون

علائم یونانی

θ	بردار پارامتر ها

فصل اول مقدمه

مقدمه

هدف اصلی این پروژه، بررسی روش‌های مختلف حل مسئلهٔ رگرسیون خطی و تحلیل پایداری و مقیاس‌پذیری آن‌ها در شرایط مختلف داده‌ای است. ابتدا با یک مثال ساده، سه روش مهم یعنی **حل تحلیلی(Normal Equation)** ، **گرادیان کاهشی دسته‌ای (Batch Gradient Descent)** و **روش مستقیم مبتنی بر تجزیهٔ SVD** بررسی و مقایسه شدند. سپس برای نشان دادن اهمیت پایداری عددی، داده‌هایی با هم‌خطی (Collinearity) ایجاد شد تا تفاوت عملکرد روش‌های مختلف در شرایط نامناسب دیده شود.

در بخش دوم، داده‌های واقعی **Auto MPG** مورد استفاده قرار گرفتند تا بتوان کارایی روش‌های یادگیری تکراری را در یک مسئلهٔ واقعی ارزیابی کرد. در این مرحله، دو الگوریتم BGD و SGD پیاده‌سازی و از نظر سرعت همگرایی، رفتار تابع هزینه و مناسب‌بودن برای داده‌های بزرگ مقایسه شدند. در نهایت، در بخش سوم یک **رگرسیون غیرخطی (Polynomial Regression)** درجهٔ ۲ طراحی شد تا نقش مدل‌های غیرخطی در بهبود دقت پیش‌بینی بررسی گردد. این مدل نیز با استفاده از SGD آموزش داده شد و منحنی برازش‌شده روی داده‌ها رسم شد.

این پروژه مجموعه‌ای از مباحث مهم در یادگیری ماشین و داده کاوی شامل پایداری عددی، پیچیدگی محاسباتی، روش‌های بهینه‌سازی، نقش نرمال‌سازی داده‌ها و مدل‌سازی غیرخطی را به صورت عملی و کاربردی مورد بررسی قرار می‌دهد.

فصل دوم
مبانی محاسباتی

مبانی محاسباتی

2-1- داده اصلی

با A اصلی سه روش را اجرا کردیم؛ نتایج عددی (بردار ستون) به‌صورت زیر است:

روش معادلات نرمال (محاسبهٔ مستقیم): $\theta_{\text{normal}} \approx \begin{bmatrix} -1.04137931 \\ 2.03103448 \end{bmatrix}$

روش گرادیان کاهشی بچ (α=0.01, 1000 iter): $\theta_{\text{BGD}} \approx \begin{bmatrix} -0.55217831 \\ 1.91468290 \end{bmatrix}$

روش شبه‌معکوس با SVD : $\theta_{\text{SVD}} \approx \begin{bmatrix} -1.04137931 \\ 2.03103448 \end{bmatrix}$

نتایج معادلات نرمال و SVD عملاً یکی‌اند؛ اما BGD با گام 0.01 و 1000 تکرار هنوز کاملاً به جواب دقیق نرسیده (تقریب خوبی است ولی نه برابر)

2-2- داده با همخطی زیاد

حل با روش نرمال: مقادیر پارامترها برای دو ستون تقریبا همخطی بسیار بزرگ و با علامت مخالف شدند (جبران همدیگر). این مقادیر ظاهراً معقول نیستند و نشان‌دهندهٔ **نوسان عددی شدید** است. هنگام محاسبهٔ $(A^TA)^{-1}$ مقدار condition number ماتریس A^TA بسیار بزرگ شد (حدود 1.56×10^{11} — (که یعنی ماتریس نزدیک به تکین (singular) است و معکوس‌پذیری عددی پایدار نیست.

حل با روش SVD :

هم‌خطی (multicollinearity) یعنی وجود ستون‌های تقریباً خطی مستقل‌ناپذیر در ماتریس طراحی. وقتی دو ستون تقریبا یکسان باشند، ماتریس A^TAنزدیک به تکین (singular) می‌شود و معکوس آن بسیار حساس به نویز کوچک در داده است.

روش معادلات نرمال مستقیماً نیاز به محاسبهٔ $(A^TA)^{-1}$دارد. اگر A^TAخوب شرطی نشده باشد)یعنی condition number خیلی بزرگ باشد(، محاسبهٔ معکوس باعث ضرب خطاهای عددی و تولید ضریب‌های بسیار بزرگ و بی‌معنی می‌شود — همان چیزی که دیدیم (ضریب‌های ≈ ±3438). همچنین معادلات نرمال عملاً مربعِ condition number را وارد می‌کند(cond(ATA) ≈ cond(A)^2)، پس مشکل را تشدید می‌کند.

SVD شبه‌معکوس از طرف دیگر تجزیهٔ مقادیر منفرد $A = U\Sigma V^T$را می‌سازد و با بررسی مقادیر منفرد کوچک می‌تواند به صورت عددی پایدارتر عمل کند .pinv معمولاً مقادیر منفرد کمتر از یک آستانهٔ tol را معکوس نمی‌کند (یا آن‌ها را با مقدارهای کوچک‌تر جایگزین می‌کند)، که مؤثرِ نوعی «نرمال‌سازی» یا «تنظیم (regularization) «است و جلوی انفجار ضرایب را می‌گیرد. همین باعث می‌شود SVD در حالت‌های نزدیک به تکین، رفتاری کنترل‌شده‌تر و قابل تفسیرتر داشته باشد.

در **عمل :**حتی SVD وقتی ستون‌ها خیلی خیلی هم‌خطی باشند، ممکن است اعداد بزرگی بدهد (چون اطلاعات واقعی برای تشخیص ضریبِ جداگانهٔ دو ستون وجود ندارد). برای حلِ مسئلهٔ هم‌خطی جدی معمولاً روش‌هایی نظیر **Ridge regression (Tikhonov regularization)**یا حذف/ادغام ویژگی‌های هم‌خطی پیشنهاد می‌شود .Ridge عملاً معکوس $A^TA + \lambda I$را محاسبه می‌کند و شرطبندی را با افزودن λبهتر می‌کند.

فصل سوم
تحلیل مقیاس پذیری محاسباتی

تحلیل مقیاس پذیری محاسباتی

مراحل:

1. **BGD**:
   - در هر تکرار از تمام داده‌ها استفاده می‌کنیم.(gradient = 1/m * A^T(Aθ - y))
   - تعداد تکرارها: 1000.
2. **SGD**:
   - در هر تکرار فقط **یک نمونه تصادفی** برای محاسبه گرادیان استفاده می‌کنیم.
   - تعداد تکرارها: 5000.
   - تابع هزینه (Cost) هر بار روی **تمام داده‌ها** محاسبه می‌شود تا مسیر همگرایی قابل مقایسه باشد.
3. **جمع‌آوری هزینه**:
   - compute_cost تابع مربعات خطا تقسیم بر $2m$است (همان تابع هزینه استاندارد رگرسیون خطی).
4. **نمودار همگرایی**:
   - محور :X تعداد تکرارها
   - محور :Y مقدار هزینه کل$J(\theta)$
   - مسیر BGD معمولاً صاف و یکنواخت است.
   - مسیر SGD نوسان دارد ولی در نهایت همگرایی می‌کند.

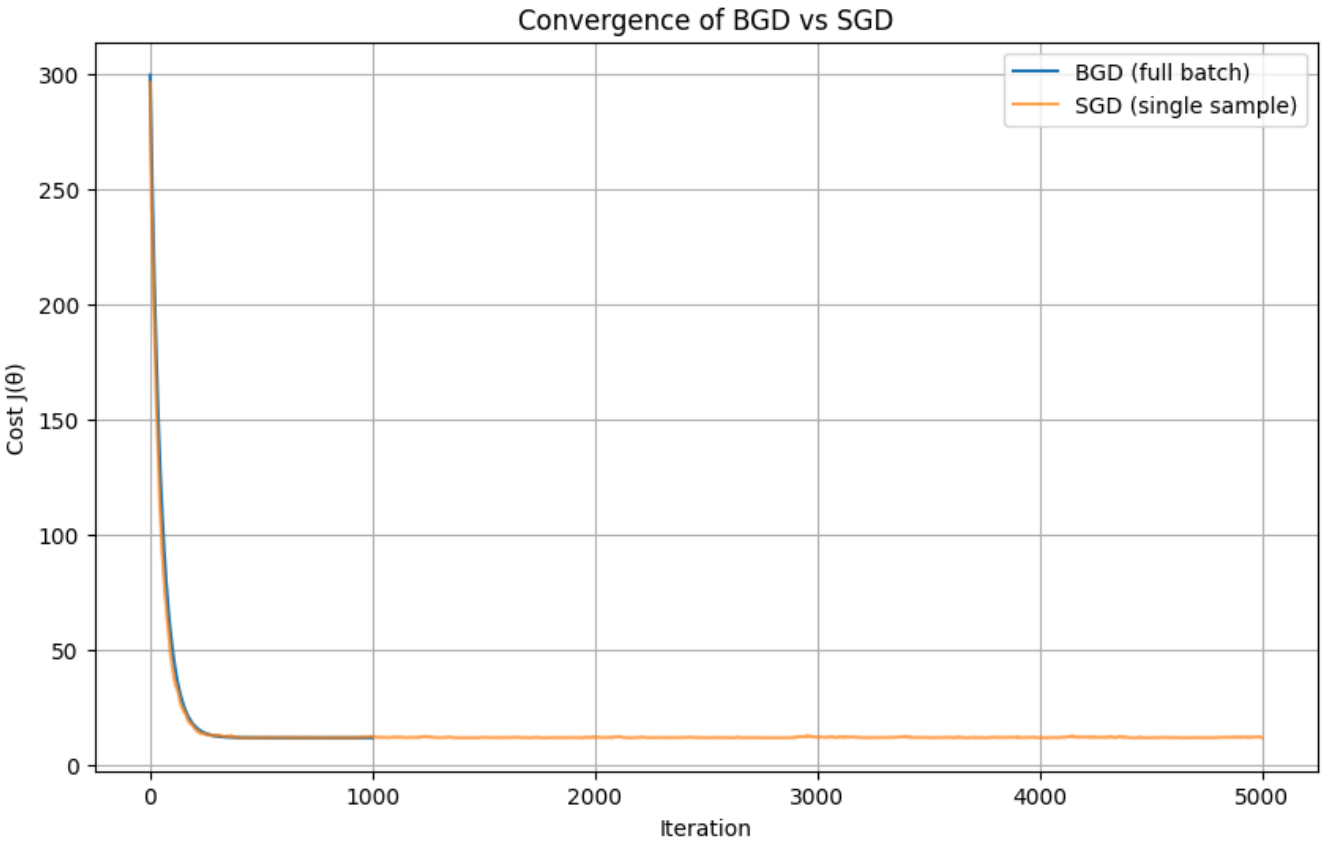
تحلیل مقیاس‌پذیری و استفاده ازSGD

1. **BGD**در داده‌های بزرگ:
   - هر تکرار نیاز به محاسبه گرادیان روی **تمام داده‌ها** دارد.
   - در داده‌های میلیون‌ها نمونه، خیلی کند و غیرعملی است.
2. **SGD (یا:Mini-Batch)**
   - فقط از یک یا چند نمونه در هر تکرار استفاده می‌کند.
   - محاسبات سبک‌تر و سریع‌تر است.

- نوسان دارد ولی تقریباً همان جواب BGD را در تعداد تکرار بیشتر می‌دهد.
- امکان **تدریجی یادگیری در داده‌های بسیار بزرگ** وجود دارد.

3. **جمع‌بندی**:

- برای داده‌های بزرگ، تقریباً همیشه از **SGDیا Mini-Batch SGD** استفاده می‌شود.
- BGDفقط برای داده‌های کوچک یا الگوریتم‌هایی که داده‌ها راحت در حافظه جای می‌گیرند مناسب است.



Convergence of BGD vs SGD1 Figure

فصل چهارم
رگرسیون غیر خطی

رگرسیون غیر خطی

1. **ماتریس طراحی برای درجه ۲**:

$$A_{\text{poly}} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & & \end{bmatrix}$$

- ستون اول = بایاس
- ستون دومx =
- ستون سومx^2 =

2. **استانداردسازی ستون‌های ویژگی**:

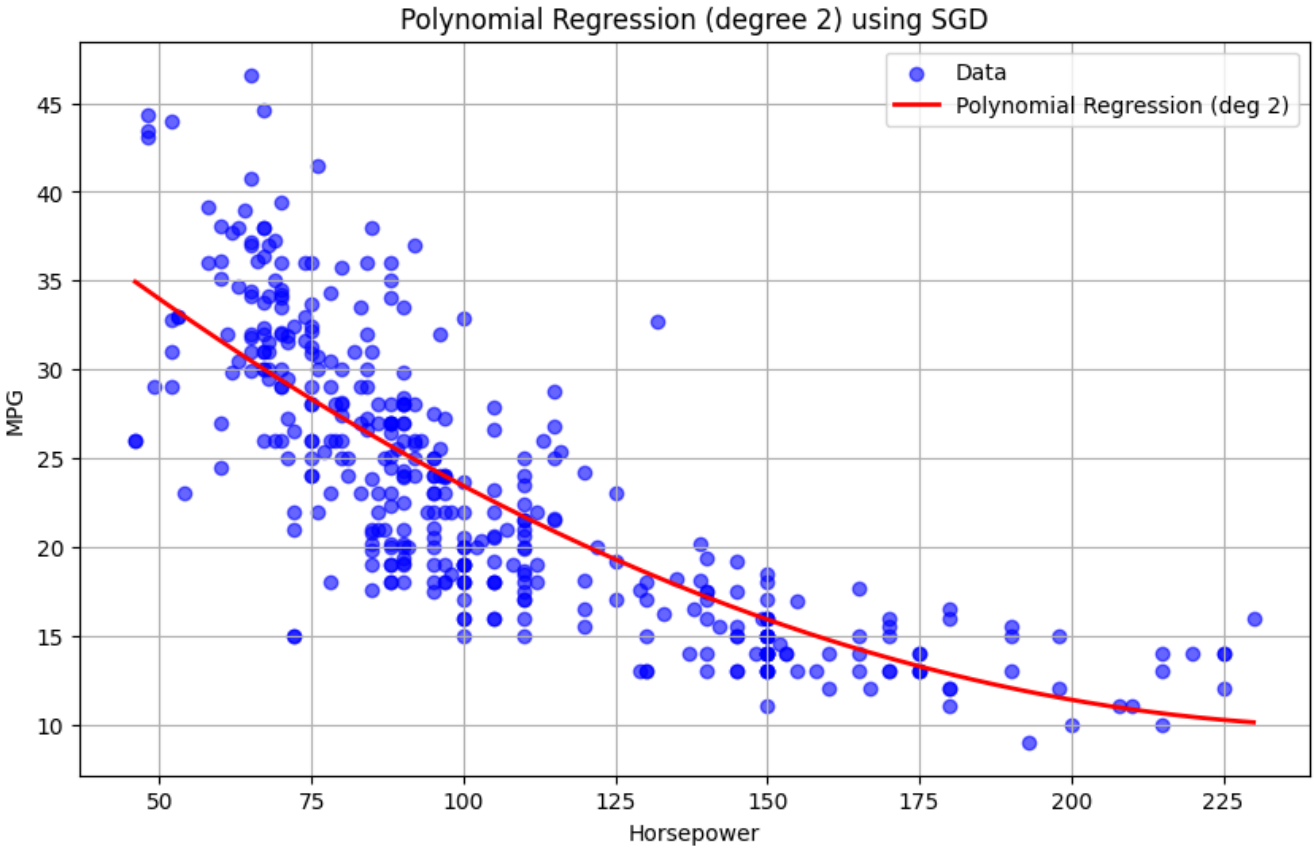
- ستون بایاس = 1 ⟵ استاندارد نمی‌شود
- ستون‌های x و StandardScaler ⟵ x^2 برای همگرایی بهترSGD

3. **SGDبرای رگرسیون غیرخطی**:

- الگوریتم همان SGD قبلی است، فقط اندازه θو Aبزرگ‌تر است (3 پارامتر)
- گرادیان برای یک نمونه$\nabla J_i(\theta) = A_i^T(A_i\theta - y_i)$:

4. **رسم منحنی**:

- نقاط برای رسم صاف تولید می‌شوند(np.linspace)
- قبل از ضرب درθ ، ستون‌های x و x^2 را با همان scaler استاندارد می‌کنیم
- منحنی قرمز = مدل چندجمله‌ای
- نقاط آبی = داده‌های واقعی



Polynomial Regression SGD2 Figure

نتیجه: یک **رگرسیون درجه ۲ با SGD** روی Horsepower → MPG داریم و منحنی قرمز روی داده‌ها قرار می‌گیرد و شکل غیرخطی رابطه را نشان می‌دهد.

فصل ششم
جمع‌بندي و نتيجه‌گيري

جمع‌بندي و نتيجه‌گيري

نتایج این پروژه نشان داد که انتخاب روش مناسب برای حل مسئلهٔ رگرسیون کاملاً به **ماهیت داده‌ها** و **اندازهٔ آن‌ها** وابسته است. در داده‌های کوچک، هر سه روشNormal Equation ، BGDو SVD قادر به یافتن راه‌حل هستند؛ اما در صورت وجود هم‌خطی بین ویژگی‌ها، روش معمول (Normal Equation) ناپایدار و نامناسب می‌شود، در حالی که **SVDپایدارترین و قابل‌اعتمادترین راه‌حل** را ارائه می‌دهد.

در کار با داده‌های واقعیAuto MPG ، مقایسهٔ الگوریتم‌های BGD و SGD نشان داد که **SGDبا وجود نوسان در مسیر همگرایی، بسیار سریع‌تر و کارآمدتر از BGD** است و به همین دلیل در یادگیری ماشین مدرن — به‌ویژه در داده‌های بزرگ — روش استاندارد محسوب می‌شود.

در بخش نهایی نیز مشاهده شد که مدل خطی ساده نمی‌تواند رابطهٔ غیرخطی بین Horsepower و MPGرا به خوبی ثبت کند؛ اما رگرسیون چندجمله‌ای درجه ۲ با استفاده از SGD توانست منحنی‌ای مناسب و سازگار با الگوی واقعی داده‌ها ارائه دهد.

به طور کلی، این پروژه اهمیت موارد زیر را نشان داد:

- نرمال‌سازی داده‌ها برای همگرایی درست الگوریتم‌های گرادیان
- برتری SVD در شرایط ناسازگار وill-conditioned
- اهمیت انتخاب روش بهینه‌سازی درست برای داده‌های بزرگ
- ضرورت استفاده از مدل‌های غیرخطی در مسائل واقعی با روابط پیچیده

در نهایت، ترکیب روش‌های عددی پایدار، الگوریتم‌های بهینه‌سازی مقیاس‌پذیر، و مدل‌های مناسب منجر به تحلیل دقیق‌تر و پیش‌بینی بهتر در مسائل یادگیری ماشین می‌شود.

منابع و مراجع

Matrix Methods in Data Mining and Pattern Recognition - Lars Elden	[1]
Introduction to Machine Learning with Python – Andreas Müller & Sarah Guido	[2]
Data Mining: Concepts and Techniques – Han, Kamber, Pei	[3]
Pattern Recognition and Machine Learning – Christopher M. Bishop	[4]

Abstract

In this project, three different approaches to solving linear regression—**Normal Equation**, **Batch Gradient Descent (BGD)**, and **Singular Value Decomposition (SVD)**—were implemented and compared. Numerical stability was evaluated using artificially generated collinear data, demonstrating that SVD performs significantly more reliably than the Normal Equation in ill-conditioned scenarios. Using the real-world Auto MPG dataset, the performance and convergence behavior of BGD and **Stochastic Gradient Descent (SGD)** were analyzed. Results showed that SGD, due to its fast updates and linear computational complexity, is more suitable for large-scale datasets. Finally, a **second-degree polynomial regression model** was trained using SGD, illustrating that nonlinear models can better capture the complex relationship between Horsepower and MPG compared to a simple linear model. Overall, this study highlights the importance of numerical stability, computational scalability, and proper model selection in machine learning tasks.

Key Words: Computational Data Mining – Collinearity – Regression - Optimization

Amirkabir University of Technology
(Tehran Polytechnic)

… Department …

MSc or PhD Thesis

Title of Thesis

By
Name

Supervisor
Dr.

Advisor
Dr.

Month & Year