دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش تمرین سوم درس داده‌کاوی محاسباتی

مبانی محاسباتی رگرسیون (پایداری و مقیاس پذیری)

نگارش
زهرا براتی

استاد درس
دکتر مهدی قطعی

تدریسیار
آقای بهنام یوسفی‌مهر

آذر ۱۴۰۴

چکیده

در این پژوهش، سه رویکرد مختلف برای حل مسائل رگرسیون خطی و غیرخطی بررسی شده است. در ابتدا، با استفاده از یک مجموعه‌داده کوچک، عملکرد روش‌های مستقیم شامل معادلات نرمال و SVD ارزیابی شد. نتایج نشان داد که معادلات نرمال در حضور هم‌خطی دچار ناپایداری عددی می‌شود، در حالی‌که SVD با کنترل مقادیر منفرد کوچک راه‌حلی پایدار و قابل اعتماد ارائه می‌کند. در ادامه و با استفاده از داده بزرگ‌تر Auto MPG، رفتار الگوریتم‌های تکراری گرادیان کاهشی دسته‌ای و تصادفی مقایسه شد. مشاهده شد که هرچند BGD مسیر همگرایی یکنواخت‌تری دارد، هزینه محاسباتی آن برای داده‌های بزرگ بالا است و SGD با وجود نوسان، کاراتر و سریع‌تر به ناحیه بهینه می‌رسد. در بخش پایانی، یک مدل رگرسیون چندجمله‌ای درجه دو برای تقریب رابطه Horsepower و MPG ساخته شد. منحنی به‌دست‌آمده نشان داد که مدل غیرخطی به‌دلیل توانایی در ثبت تغییرات شیب، توصیف دقیق‌تری نسبت به مدل خطی ارائه می‌دهد. در مجموع، تحلیل‌ها نشان می‌دهد که انتخاب روش مناسب در رگرسیون به اندازه داده، ساختار ویژگی‌ها و الزامات پایداری عددی بستگی دارد.

واژه‌های کلیدی:

رگرسیون خطی، رگرسیون چندجمله‌ای، معادلات نرمال، تجزیه مقادیر منفرد، گرادیان کاهشی دسته‌ای، گرادیان کاهشی تصادفی

فهرست مطالب

صفحه

فصل اول

مقدمه

مقدمه

رگرسیون یکی از پایه‌ای‌ترین ابزارها برای تحلیل رابطه بین متغیرهاست و در بیشتر کاربردهای یادگیری ماشین نقش مهمی دارد. با این حال، روش‌هایی که برای حل رگرسیون استفاده می‌کنیم بسته به اندازه داده، نوع ویژگی‌ها و محدودیت‌های محاسباتی می‌توانند رفتار کاملاً متفاوتی داشته باشند. در این پروژه تلاش شده است که این تفاوت‌ها به‌صورت مرحله‌به‌مرحله بررسی شود.

در بخش اول، با یک مجموعه‌داده کوچک کار شد تا عملکرد روش‌های مستقیم مثل معادلات نرمال و تجزیه SVD بررسی شود و مشخص شود که این روش‌ها در حضور هم‌خطی چطور دچار ناپایداری عددی می‌شوند. در بخش دوم، داده بزرگ‌تری انتخاب شد تا رفتار الگوریتم‌های گرادیان کاهشی دسته‌ای و تصادفی از نظر سرعت و مقیاس‌پذیری مقایسه شود. در نهایت، در بخش سوم یک مدل چندجمله‌ای درجه دو ساخته شد تا نشان داده شود که استفاده از مدل‌های غیرخطی چگونه می‌تواند الگوهای واقعی داده را بهتر از مدل خطی ساده ثبت کند. هدف اصلی این گزارش این است که نشان دهد انتخاب روش مناسب برای رگرسیون کاملاً به شرایط مسئله و ویژگی‌های داده بستگی دارد.

فصل دوم

مبانی محاسباتی (داده‌های کوچک)

مبانی محاسباتی (داده‌های کوچک)

در این بخش،

۲-۱- روش مستقیم (Equations Normal)

معادلات نرمال[1] راه حل تحلیلی و مستقیم برای مسئله کمترین مربعات خطی است. برای یافتن بردار ضرایب θ که خطای $\| A\theta - y \|^2$ را کمینه می‌کند، به جای بهینه‌سازی تکراری، می‌توان مستقیماً سیستم معادلات خطی $A^T A\theta = A^T y$ را حل کرد.

راه‌حل به صورت $\theta = (A^T A)^{-1} A^T y$ است. این روش برای d(تعداد ویژگی‌ها) کم، بسیار سریع است، اما دو مشکل محاسباتی اساسی دارد:

۱. هزینه معکوس کردن ماتریس $A^T A$ از مرتبه $O(d^3)$ است که برای d بزرگ بسیار سنگین است.

۲. اگر ویژگی‌ها هم‌خطی داشته باشند، ماتریس $A^T A$ تکین[2] یا بدحالت[3] شده و معکوس آن از نظر عددی ناپایدار یا غیرممکن می‌شود.

مقدار θ به دست آمده از این روش برابر $\begin{bmatrix}[-1.04137931] \\ [\ 2.03103448]\end{bmatrix}$ است. این بردار نشان می‌دهد که مدل خطی به‌شکل تقریباً $\hat{y} = -1.0414 + 2.0310\ x$ داده‌ها را توصیف می‌کند.

۲-۲- روش تکراری (Batch Gradient Descent)

گرادیان کاهشی دسته‌ای[4] یک روش بهینه‌سازی تکراری است که با شروع از یک θ اولیه، به‌صورت گام‌به‌گام به سمت کمینه تابع هزینه $J(\theta)$ حرکت می‌کند.

---

[1] Equations Normal

[2] Singular

[3] ill-conditioned

[4] Batch Gradient Descent (BGD)

در هر گام، گرادیان (مشتق) تابع هزینه محاسبه شده و θ در خلاف جهت آن حرکت می‌کند.

در این روش «دسته‌ای[5]» به این معناست که برای محاسبه گرادیان در هر گام، از تمام n نمونه موجود در مجموعه داده استفاده می‌شود: $\nabla J(\theta) = \frac{1}{m} A^T(A\theta - y)$.

همچنین رابطه به‌روزرسانی θ در الگوریتم گرادیان کاهشی دسته‌ای برابر است با: $\theta^{(k+1)} = \theta^{(k)} - \eta\, \nabla J(\theta^{(k)})$. در این تمرین، با انتخاب نرخ یادگیری $\eta = 0.01$ و تعداد تکرار $n_{\text{iterations}} = 1000$، θ برابر $\begin{bmatrix}[-0.55217831] \\ [\ 1.9146829\ ]\end{bmatrix}$ و تابع هزینه تقریباً برابر $J(\theta_{BGD}) = 0.01873$ به دست آمد.

این نتیجه نشان می‌دهد که الگوریتم گرادیان کاهشی با پارامترهای انتخاب‌شده هنوز به مقدار بهین θ (که در بخش قبل محاسبه شد) نرسیده است. با افزایش تعداد تکرارها یا تنظیم نرخ یادگیری می‌توان مقدار θ را به جواب دقیق‌تری نزدیک کرد.

۲-۳- روش مستقیم پایدار (SVD)

روشی بسیار پایدارتر برای حل مستقیم رگرسیون، استفاده از شبه‌معکوس مور-پنروز (A^+) است. بهترین و استانداردترین راه محاسباتی برای یافتن A^+، استفاده از تجزیه مقادیر منفرد[6] است.

اگر $A = U\Sigma V^T$ باشد، آنگاه: $A^+ = V\Sigma^+ U^T$. این روش حتی اگر ماتریس $A^T A$ تکین باشد (یعنی هم‌خطی کامل وجود داشته باشد)، یک راه‌حل منحصر به فرد و بهینه (با کمترین نرم) ارائه می‌دهد و از نظر عددی بسیار پایدار است.

در نتیجه محاسبات برای این روش θ برابر $\begin{bmatrix}[-1.04137931] \\ [\ 2.03103448]\end{bmatrix}$ که دقیقاً با θ حاصل از روش معادلات نرمال برابر است.

[5] Batch

[6] Singular Value Decomposition (SVD)

۲-۴- تحلیل پایداری محاسباتی

در این بخش به تحلیل پایداری محاسباتی می‌پردازیم.

۲-۴-۱- مقایسه مقادیر θ در سه روش پیشین

در جدول زیر مقادیر محاسبه شده قابل مشاهده است.

روش	θ_0	θ_1
Equations Normal	-1.04137931	2.03103448
BGD	-0.55217831	1.9146829
SVD	-1.04137931	2.03103448

همانطور که می‌بینیم، θ به‌دست‌آمده از SVD دقیقاً با نتیجه معادلات نرمال همسان است، درحالی‌که گرادیان کاهشی به‌دلیل تعداد تکرار محدود مقادیری متفاوت ارائه می‌دهد. با افزایش تعداد تکرارها یا تنظیم نرخ یادگیری می‌توان نتیجه BGD را به θ دقیق نزدیک‌تر کرد.

۲-۴-۲- ساخت ماتریس با هم‌خطی شدید

پس از ساخت ماتریس جدید با روابط داده شده

```
01 noise = np.random.rand(m, 1) * 0.0001
02  A_collinear = np.hstack((A, A[:, [1]] + noise))
```

و حل مجدد مسئله با روش معادلات نرمال و تجزیه مقادیر منفرد، نتایج زیر حاصل شد:

روش	θ_0	θ_1	θ_2
Equations Normal	-8.53782625e-01	3.43827452e+03	-3.43623267e+03
SVD	-8.53775599e-01	3.43826298e+03	-3.43622113e+03

لازم به ذکر است با قرار دادن noise = 0 محاسبات حاصل از روش معادلات نرمال با خطای «Singular matrix» مواجه شد. درحالی‌که روش تجزیه مقادیر منفرد بدون خطا اجرا شد.

۲-۴-۳- تحلیل

وقتی در حضور هم‌خطی شدید، ماتریس A^TA تقریباً تکین و از نظر عددی بدحالت می‌شود. این وضعیت باعث می‌شود روش معادلات نرمال هنگام معکوس‌گیری این ماتریس، به مقادیر ناپایدار و ضرایب بسیار بزرگ یا حتی شکست محاسباتی منجر شود. مقالات موجود نیز همین پدیده را گزارش می‌کنند و نشان می‌دهند که وابستگی خطی میان ویژگی‌ها خطاهای گردشی (round-off) را تقویت کرده و واریانس ضرایب مدل را به‌شدت افزایش می‌دهد.

در مقابل، روش SVD از معکوس‌گیری مستقیم A^TA پرهیز می‌کند و به‌جای آن ماتریس A را به صورت $A = U\Sigma V^T$ تجزیه می‌کند. وقتی هم‌خطی وجود داشته باشد، یکی از مقادیر منفرد σ_i بسیار کوچک می‌شود؛ SVD این مقادیر کوچک را کنترل کرده و در شبه‌معکوس آن‌ها را به شکلی پایدار مدیریت می‌کند. به همین دلیل مقالات تأکید می‌کنند که SVD خطاهای عددی را تقویت نمی‌کند و حتی در شرایط هم‌خطی یا رتبه ناکامل، پایدارترین تخمین ممکن (کمترین‌نُرم) را ارائه می‌دهد.

به‌صورت خلاصه، معادلات نرمال به‌دلیل وابستگی به $(A^TA)^{-1}$ در برابر هم‌خطی حساس و ناپایدار است، در حالی که SVD با کنترل مقادیر منفرد کوچک، یک راه‌حل پایدار و مقاوم در برابر نویز و خطای محاسباتی فراهم می‌کند.

فصل سوم

تحلیل مقیاس پذیری محاسباتی (داده‌های بزرگ)

تحلیل مقیاس‌پذیری محاسباتی (داده‌های بزرگ)

در این بخش، کارایی روش‌های مبتنی بر گرادیان در مقیاس بزرگ بررسی می‌شود. برخلاف بخش اول که مسئله تنها یک ویژگی و چهار نمونه داشت، داده Auto MPG شامل چند صد نمونه است و از این جهت می‌توان تفاوت رفتاری روش‌های تکراری را بهتر مشاهده کرد.

در ادامه، داده Auto MPG بارگذاری و پاک‌سازی می‌شود، ماتریس طراحی برای مدل خطی ساخته و نرمال‌سازی و سپس هر یک از دو الگوریتم BGD و SGD به‌طور جداگانه اجرا و مسیر همگرایی آن‌ها تحلیل می‌شود.

۳-۱- بارگذاری داده‌ها و آماده‌سازی ماتریس‌ها

در این مرحله داده Auto MPG از مخزن UCI بارگذاری و مقادیر گمشده ستون Horsepower حذف و برچسب MPG و ویژگی ورودی Horsepower انتخاب شد. برای بهبود همگرایی روش‌های مبتنی‌بر گرادیان، ستون Horsepower به‌صورت استاندارد نرمال‌سازی شده و سپس ماتریس طراحی A مطابق مدل $y = \theta_0 + \theta_1 \times Horsepower$ ساخته شد که شامل یک ستون ثابت و یک ستون ویژگی نرمال‌شده است. داده آماده‌شده در مراحل بعد برای اجرای BGD و SGD مورد استفاده قرار می‌گیرد.

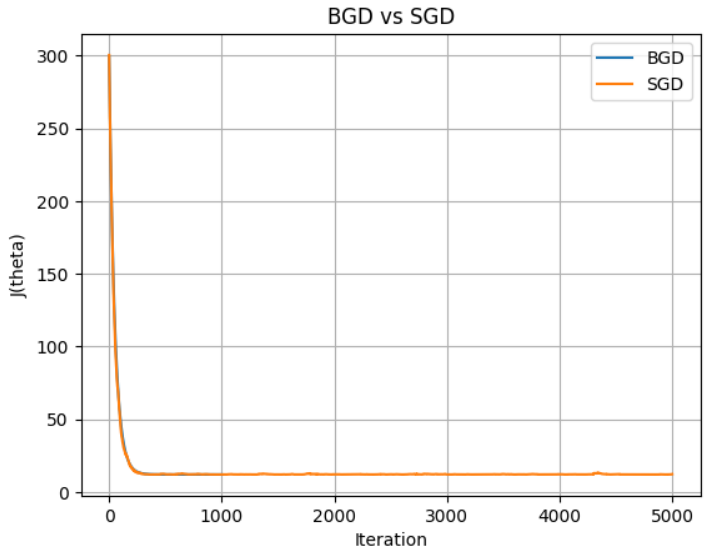
۳-۲- مقایسه محاسباتی BGD در مقابل SGD

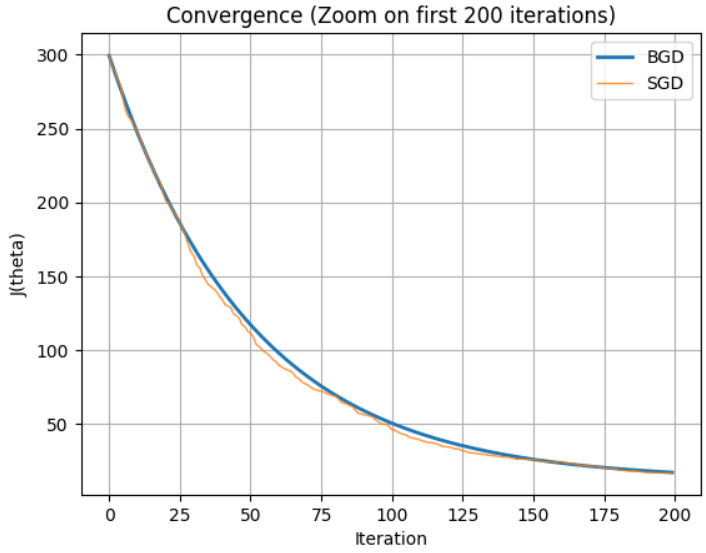
در این مرحله پس از پیاده‌سازی BGD و SGD، نتایج زیر حاصل شد:

روش	θ_0	θ_1	$Last\ J$
BGD	23.44490618	-6.06761045	11.971832015876284
SGD	23.82669705	-6.68718906	12.236104231855803

همچنین نمودار همگرایی بر حسب تکرار نشان می‌دهد که هر دو روش در تکرارهای ابتدایی کاهش تند دارند و سریعاً به ناحیه کمینه می‌رسند. در شکل بزرگ‌مقیاس، منحنی‌های دو روش تقریباً روی یکدیگر قرار می‌گیرند، اما در بزرگ‌نمایی تکرارهای نخست، می‌توان مشاهده کرد که BGD مسیر یکنواخت‌تری

دارد، در حالی که SGD به دلیل استفاده از نمونه‌های تصادفی دارای نوسان‌های کوچک است. با این وجود، هر دو روش به مقدار تقریباً مشابهی از $J(\theta)$ همگرا می‌شوند که نشان‌دهنده سادگی مدل و اثر نرمال‌سازی ویژگی است.





۳-۳- تحلیل محاسباتی

در مسائل یادگیری ماشین مقیاس‌بزرگ، استفاده از روش گرادیان کاهشی دسته‌ای به دلیل هزینه

محاسباتی زیاد هر تکرار کارآمد نیست؛ زیرا در هر گام لازم است گرادیان بر روی کل داده‌ها محاسبه شود که برای مجموعه‌داده‌های بزرگ بسیار سنگین است. در مقابل، روش گرادیان کاهشی تصادفی در هر تکرار تنها از یک نمونه داده‌ها استفاده می‌کند. اگرچه مسیر همگرایی در SGD نویزی‌تر از BGD است، اما به دلیل تعداد بسیار بیشتر به‌روزرسانی‌ها در زمان ثابت، در عمل بسیار سریع‌تر به ناحیه‌ی کمینه تابع هزینه نزدیک می‌شود. بنابراین، در محیط‌های محاسباتی بزرگ و کاربردهای عملی یادگیری ماشین، روش‌های مبتنی‌بر SGD بر BGD به دلیل هزینه‌ی سنگین محاسباتی ارجحیت دارد.

فصل چهارم

کاربرد (رگرسیون غیرخطی)

کاربرد (رگرسیون غیرخطی)

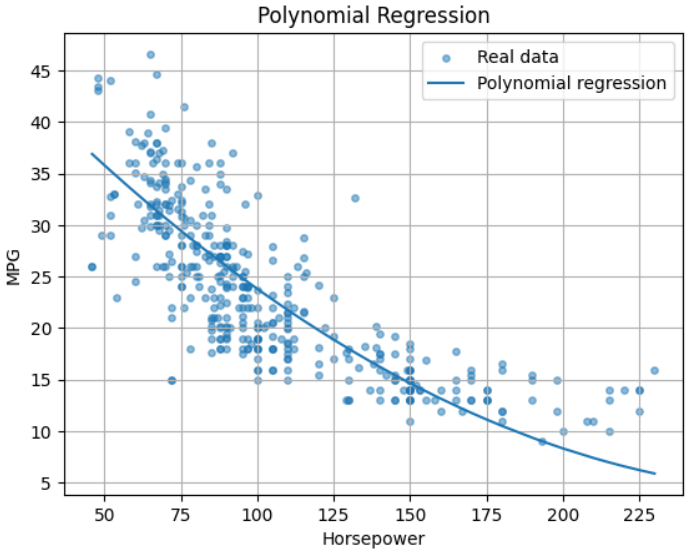
در این بخش، با استفاده از ویژگی Horsepower از مجموعه‌داده Auto MPG، یک مدل رگرسیون چندجمله‌ای درجه دو به فرم $y = \theta_0 + \theta_1 x + \theta_2 x^2$ ساخته شد. برای بهبود همگرایی، ستون‌های x و x^2 به‌صورت جداگانه نرمال‌سازی شدند و ماتریس A_{poly} شامل سه ستون $[1, x_{\text{norm}}, x_{\text{norm}}^{(2)}]$ تشکیل شد.

پس از اجرای روش SGD بر روی ماتریس جدید، نتایج زیر حاصل شد:

روش	θ_0	θ_1	θ_2	$Last\ J$
SGD	23.7286363	-12.48909616	5.55257232	10.719091833697766

این مقادیر نشان می‌دهد که مدل یاد گرفته است که چگونه رابطه غیرخطی بین توان موتور (Horsepower) و مصرف سوخت (MPG) را تقریب بزند. ضریب منفی θ_1 نشان‌دهنده‌ی کاهش اولیه MPG با افزایش Horsepower است، در حالی که ضریب مثبت θ_2 نقش انحنای مدل را تعیین می‌کند و باعث می‌شود مدل بتواند روند ملایم‌تر و خمیده قسمت‌های انتهایی داده‌ها را دنبال کند.

در نهایت با رسم منحنی رگرسیون به دست آمده (مدل چندجمله ای) روی نمودار پراکندگی داده‌ها، مشاهده شد نتایج به دست آمده انطباق مناسبی با داده‌های واقعی دارند.



فصل پنجم

جمع‌بندی و نتیجه‌گیری

جمع‌بندی و نتیجه‌گیری

در این پروژه، سه رویکرد مختلف برای برازش مدل‌های خطی و غیرخطی بر روی داده‌ها بررسی شد. در بخش نخست، تحلیل روی داده‌های کوچک نشان داد که روش‌های مستقیم مانند معادلات نرمال در مسائل هم‌خطی حساسیت بالایی دارند و ممکن است به پاسخ‌های ناپایدار ختم شوند، در حالی‌که روش SVD با کنترل مقادیر منفرد کوچک، راه‌حلی پایدارتر ارائه می‌کند. این تفاوت در شرایطی آشکارتر شد که ستون‌های ماتریس طراحی وابستگی شدید داشتند و معادلات نرمال دچار مشکل شدند.

در بخش دوم، با استفاده از داده بزرگ‌تر Auto MPG، رفتار الگوریتم‌های تکراری بررسی شد. نتایج نشان داد که گرادیان کاهشی دسته‌ای گرچه مسیری یکنواخت و بدون نوسان دارد، اما از نظر محاسباتی برای داده‌های بزرگ ناکارآمد است؛ زیرا هر تکرار مستلزم پردازش کل مجموعه داده است. در مقابل، روش گرادیان کاهشی تصادفی با وجود نوسان در مسیر همگرایی، به‌طور قابل توجهی از نظر زمانی مقرون‌به‌صرفه‌تر است و در عمل با تعداد زیادی به‌روزرسانی سبک، سریع‌تر به ناحیه کمینه نزدیک می‌شود. این رفتار، دلیل اصلی استفاده گسترده از روش‌های مبتنی بر SGD در کاربردهای یادگیری ماشین مقیاس‌بزرگ است.

در بخش سوم نیز یک مدل چندجمله‌ای درجه دو بر اساس ویژگی Horsepower ساخته شد. منحنی برازش‌شده نشان داد که افزودن جمله درجه دو موجب افزایش انعطاف مدل و بهبود تطابق آن با روند واقعی داده‌ها می‌شود.

در مجموع، نتایج نشان می‌دهد که انتخاب روش مناسب برای حل مسئله رگرسیون، به اندازه داده‌ها، ساختار ویژگی‌ها و الزامات پایداری عددی بستگی دارد. روش‌های مستقیم برای مسائل کوچک مناسب‌اند، روش‌های تجزیه‌ای مانند SVD در حضور هم‌خطی برتری دارند، و برای داده‌های بزرگ، الگوریتم‌های گرادیان تصادفی کارآمدترین انتخاب هستند. همچنین افزودن پیچیدگی کنترل‌شده به مدل (مانند درجه دوم) می‌تواند برازش بهتری نسبت به مدل خطی فراهم کند، بدون آن‌که هزینه محاسباتی چشمگیری تحمیل کند.

منابع و مراجع

[1] R. F. G.V. Haines, "Modeling by singular value decomposition and the elimination of statistically insignificant coefficients," *Computers & Geosciences,* pp. 19-28, 2013.

https://numpy.org

پیوست‌ها

لینک گیت‌هاب پروژه:

https://github.com/ZahraBarati99/Computational-Fundamentals-of-Regression-Stability-and-Scalability