گزارش پروژه فاز دوم: رگرسیون چندجمله‌ای با گرادیان کاهشی

مقدمه

در این پروژه، قصد داریم مسئله رگرسیون غیرخطی را با استفاده از مدل‌های چندجمله‌ای و الگوریتم گرادیان کاهشی مورد بررسی قرار دهیم. برخلاف فاز اول که به دنبال بهترین خط راست بودیم، در این فاز به دنبال بهترین منحنی چندجمله‌ای هستیم که بر داده‌ها منطبق شود. مدل چندجمله‌ای علی‌رغم غیرخطی بودن نسبت به ورودی `x`، نسبت به پارامترهای خود `w` خطی است. این ویژگی به ما اجازه می‌دهد تا از همان الگوریتم گرادیان کاهشی ساده‌شده در فاز اول برای یافتن پارامترها استفاده کنیم.

هدف اصلی این گزارش، بررسی تأثیر دو پارامتر مهم بر عملکرد مدل است:

** .1درجه چندجمله‌ای :**(m) که مستقیماً بر پیچیدگی مدل تأثیر می‌گذارد و می‌تواند منجر به پدیده‌ی بیش‌برازش (Overfitting) شود.

** .2تعداد داده‌های آموزشی :**(n) که بر توانایی مدل در تعمیم‌دهی (Generalization) تأثیرگذار است.

تمامی تحلیل‌های زیر بر اساس پیاده‌سازی الگوریتم گرادیان کاهشی از پایه و با استفاده از زبان برنامه‌نویسی پایتون انجام شده است.

بخش ۱: بارگذاری و تحلیل اولیه داده‌ها

در این بخش، داده‌ها بارگذاری و مورد بررسی اولیه قرار گرفتند.

الف) بارگذاری داده‌ها و رسم Scatter Plot

داده‌های پروژه شامل ۵۰۰ نمونه است که به دو بخش آموزشی (۳۲۰ نمونه) و آزمون (۱۸۰ نمونه) تقسیم شده‌اند. هر نمونه فقط یک ویژگی (تک‌متغیره) دارد. پس از بارگذاری فایل‌های `trainX.npy`, `trainY.npy`, `testX.npy`, `testY.npy`، نمودار پراکندگی (Scatter Plot) داده‌ها رسم شد.

همانطور که در شکل مشخص است، داده‌ها از یک الگوی غیرخطی و سینوسی‌شکل پیروی می‌کنند. نقاط آبی رنگ داده‌های آموزشی و نقاط قرمز رنگ داده‌های آزمون را نشان می‌دهند. هر دو مجموعه داده، دامنه‌ی مشابهی از ورودی (بین ۳- تا ۱-) را پوشش می‌دهند.

ب) تحلیل و پاسخ به سوال: چرا مدل چندجمله‌ای برای این داده‌ها مناسب است؟

با توجه به نمودار پراکندگی داده‌ها، به وضوح مشاهده می‌شود که رابطه بین ورودی (x) و خروجی (y) خطی نیست و از یک الگوی منحنی و نوسانی پیروی می‌کند. یک مدل خطی ساده (خط راست) نمی‌تواند این پیچیدگی را به خوبی مدل‌سازی کند و با خطای زیادی مواجه خواهد شد. مدل‌های چندجمله‌ای به دلیل ساختار غیرخطی خود که شامل توان‌های مختلفی از x است، قادر به برازش این نوع روابط هستند. بنابراین، انتخاب یک مدل چندجمله‌ای برای این مجموعه داده، انتخابی منطقی و مناسب به نظر می‌رسد.

بخش ۲: بررسی تأثیر درجه چندجمله‌ای

در این بخش، با ثابت نگه‌داشتن تعداد داده‌های آموزشی (۲۰ داده اول) تأثیر درجه چندجمله‌ای (m) بر عملکرد مدل بررسی شده است.

الف) نتایج و تحلیل مدل‌ها برای درجات مختلف

برای هر یک از درجات مختلف(m = 1, 2, 4, 8, 10, 13, 15, 17, 19) ، مدل مربوطه با الگوریتم گرادیان کاهشی آموزش داده شد. برای جلوگیری از ناپایداری عددی) مانند (exploding gradient در درجات بالا، داده‌ها نرمال‌سازی (Standardization) شدند و نرخ یادگیری متناسب با درجه کاهش یافت. نتایج به دست آمده در جداول زیر قابل مشاهده است.

الف.۱) رسم منحنی‌های چندجمله‌ای

با رسم منحنی‌های به‌دست‌آمده روی ۲۰ نقطه آموزشی، روند تغییرات مدل با افزایش درجه به وضوح قابل مشاهده است:

m=1 -(خطی):** مدل یک خط راست است و به هیچ وجه نمی‌تواند نوسانات داده‌ها را دنبال کند. این پدیده که مدل برای یادگیری الگوی داده بسیار ساده است، **کم‌برازش **(Underfitting) نامیده می‌شود.

m=4, m=8, m=10: -با افزایش درجه، مدل انعطاف‌پذیرتر شده و بهتر می‌تواند خود را با الگوی داده‌ها وفق دهد. در این درجات، مدل تا حد زیادی الگوی کلی حاکم بر داده‌ها را یاد گرفته است.

m=15, m=17, m=19: -در این درجات بالا، مدل شروع به ایجاد نوسانات شدید بین نقاط می‌کند تا بتواند از همه نقاط آموزشی عبور کند. این پدیده که مدل به جای یادگیری الگوی کلی، نویز و جزییات داده‌ها را نیز به خاطر می‌سپارد، **بیش‌برازش **(Overfitting) نام دارد.

الف.۲) رسم نرم L2 پارامترها

نمودار نرم L2 پارامترها (بدون در نظر گرفتن بایاس) بر حسب درجه رسم شد. مشاهده می‌شود که:

-برای درجه ۲، نرم پارامترها به طور قابل توجهی افزایش یافته است.

-با افزایش درجه به ۴ و ۸، نرم پارامترها مجدداً کاهش می‌یابد.

-برای درجات بسیار بالا (مثلاً ۱۹)، نرم پارامترها بسیار کوچک می‌شود.

این رفتار نشان‌دهنده‌ی این است که مدل‌های بسیار ساده (m=1) و بسیار پیچیده (m=19) به دنبال یافتن مقادیر پارامترهایی با اندازه‌های متفاوت هستند. در مدل‌های با بیش‌برازش، اگرچه نرم پارامترها می‌تواند کوچک باشد، اما نوسانات شدید بین نقاط نشان از عدم تعمیم‌پذیری مناسب دارد.

ب) رسم MSE برای داده‌های Train و Test و پاسخ به سوالات

نمودار خطای MSE برای داده‌های آموزشی و آزمون بر حسب درجه رسم شد. تحلیل این نمودار پاسخ سوالات مطرح‌شده را به وضوح نشان می‌دهد:

** -چرا خطا به ازای m کوچک روی داده‌های آموزش و آزمون زیاد است؟**

درm های کوچک (مثلاً ۱ و ۲)، مدل از پیچیدگی کافی برای یادگیری الگوی غیرخطی داده‌ها برخوردار نیست (کم‌برازش). بنابراین، هم روی داده‌های آموزشی و هم روی داده‌های جدید (آزمون) عملکرد ضعیفی دارد و خطای آن در هر دو بالا است.

** -چرا با افزایش مقدار m خطا روی داده‌های آموزش کم می‌شود؟**

با افزایشm ، مدل پیچیده‌تر و منعطف‌تر می‌شود و توانایی بیشتری برای انطباق با داده‌های آموزشی پیدا می‌کند. این انطباق بیشتر باعث کاهش خطا روی داده‌هایی می‌شود که مدل با آنها آموزش دیده است. در درجات بسیار بالا، خطای آموزش به مقادیر بسیار نزدیک به صفر می‌رسد.

** -چرا خطا به ازای m بزرگ روی داده‌های آموزش کم ولی روی داده‌های آزمون زیاد است؟**

این دقیقاً تعریف پدیده‌ی **بیش‌برازش** است. مدل‌های با درجه بالا) مثلاً (m>=15 آن‌قدر پیچیده شده‌اند که به جای یادگیری الگوی کلی، نویز و نقاط داده‌های آموزشی را "حفظ" کرده‌اند. در نتیجه، وقتی با داده‌های جدید و دیده‌نشده (آزمون) مواجه می‌شوند، عملکرد بسیار ضعیفی داشته و خطای آنها به شدت افزایش می‌یابد.

** -بهترین درجه چندجمله‌ای چند است و چرا؟**

با توجه به نمودار خطا، بهترین درجه، **m=4**است. در این نقطه، خطای آموزش به میزان قابل توجهی کاهش یافته و خطای آزمون نیز در پایین‌ترین مقدار خود قرار دارد. این نشان‌دهنده‌ی یک مصالحه (Trade-off) مناسب بین کم‌برازش و بیش‌برازش است. مدل با درجه ۴ به خوبی الگوی اصلی داده را یاد گرفته و توانایی تعمیم آن به داده‌های جدید نیز بالاست.

ج) گزارش بهترین پارامترها

بهترین مدل با درجه m=4 به دست آمد.

د) رسم منحنی نهایی و تحلیل آن

منحنی مدل نهایی با درجه ۴ بر روی داده‌ها رسم شد. تحلیل این منحنی:

** -آیا منحنی به‌دست‌آمده خوب برازش شده است؟**

بله، منحنی به‌دست‌آمده به خوبی الگوی اصلی و سینوسی حاکم بر داده‌ها را دنبال می‌کند. بدون ایجاد نوسانات اضافی، از میان نقاط عبور کرده و شکل کلی داده‌ها را نشان می‌دهد.

** -خطا بیشتر کجا وجود دارد؟**

خطا در نقاطی که داده‌ها پراکندگی بیشتری دارند (به خصوص در ابتدا و انتهای بازه و در نقاط اوج و فرود منحنی) بیشتر است. این طبیعی است، زیرا مدل سعی دارد یک منحنی صاف را بر داده‌های نویزی برازش دهد.

** -آیا نشانه‌ای از بیش‌برازش مشاهده می‌شود؟**

خیر، در مدل با درجه ۴ هیچ نشانه‌ای از بیش‌برازش مشاهده نمی‌شود. منحنی صاف و بدون پیچ‌وخم‌های اضافی است و در بازه‌های بدون داده، رفتار معقولی از خود نشان می‌دهد. همچنین، اختلاف بین خطای آموزش و آزمون در این درجه حداقل است که تأییدی بر عدم وجود بیش‌برازش است.

###بخش ۳: بررسی تأثیر تعداد داده‌های آموزشی

در این بخش، درجه چندجمله‌ای را روی مقدار بالای ۱۹ ثابت نگه داشته و تأثیر افزایش تعداد داده‌های آموزشی (n)را بر عملکرد مدل و پدیده‌ی بیش‌برازش بررسی می‌کنیم.

الف) آموزش مدل با تعداد داده‌های مختلف

مدل با درجه ۱۹ برای تعداد داده‌های مختلف (n = 10, 20, 40, 80, 160, 320) آموزش داده شد. با توجه به گستره‌ی وسیع مقادیر ویژگی `x` (که پس از نرمال‌سازی به توان ۱۹ می‌رسد)، برای جلوگیری از ناپایداری عددی، از نرخ‌های یادگیری بسیار کوچک استفاده شد.

تحلیل جدول پارامترها:

با افزایش تعداد داده‌های آموزشی، یک الگوی مشخص در مقادیر پارامترها مشاهده می‌شود:

-پارامترهای مرتبط با توان‌های بالا) مانند w13 تا (w19 که در مدل با n=10 نوسانات زیادی داشتند، با افزایش nبه سمت مقادیر کوچک‌تر و نزدیک به صفر میل می‌کنند. این رفتار نشان می‌دهد که مدل برای تعمیم‌دهی بهتر، به این نتیجه رسیده است که جملات با درجه بالا اهمیت کمی در پیش‌بینی خروجی دارند.

-در مقابل، پارامترهای با درجات پایین‌تر) مانند (w1, w2, w3 به مقادیر نسبتاً پایداری همگرا می‌شوند.

این مشاهدات نشان می‌دهد که مدل چندجمله‌ای نهایی با درجه ۱۹، پس از آموزش با داده‌های کافی، عملاً به مدلی با درجه مؤثر بسیار پایین‌تر (احتمالاً ۳ یا ۴) تبدیل می‌شود و جملات با درجه بالا عملاً حذف می‌شوند. این یکی از

مکانیسم‌های **منظم‌سازی (Regularization) ضمنی** است که در آن داده‌های بیشتر، مدل را به سمت سادگی سوق می‌دهند.

ب) رسم MSE برای تعداد داده‌های مختلف

نمودار خطای آموزش و آزمون بر حسب تعداد داده‌های آموزشی رسم شد. (محور افقی به صورت لگاریتمی انتخاب شد).

ج) تحلیل تأثیر تعداد داده‌ها بر بیش‌برازش و پاسخ به سوال

-با تعداد داده‌های بسیار کم (n=10 و ۲۰)، مدل با درجه ۱۹ به شدت دچار بیش‌برازش می‌شود. خطای آموزش بسیار پایین (نزدیک به صفر) است، اما خطای آزمون بسیار بالاست. مدل نویز موجود در داده‌های محدود را به خاطر سپرده است.

-با افزایش تدریجی تعداد داده‌ها) تا(n=80 ، خطای آموزش اندکی افزایش یافته و خطای آزمون به شدت کاهش می‌یابد. این بدان معناست که مدل دیگر نمی‌تواند نویز همه داده‌ها را به خاطر بسپارد و مجبور است الگوی اصلی حاکم بر داده‌ها را یاد بگیرد که باعث بهبود عملکرد آن روی داده‌های جدید می‌شود.

-در ادامه با افزایش تعداد داده‌ها (n=160 و ۳۲۰)، هر دو خطای آموزش و آزمون به طور همزمان افزایش می‌یابند. این پدیده به دلیل بزرگی بیش از حد درجه مدل (۱۹) نسبت به داده‌هاست. حتی با داده‌های زیاد، یک مدل بسیار پیچیده ممکن است همچنان در تطبیق با الگوی اصلی داده‌ها با مشکل مواجه شود و خطای ذاتی (Bias) آن بالا بماند.

نتیجه‌گیری نهایی: افزایش تعداد داده‌های آموزشی تأثیر چشمگیری در کاهش پدیده‌ی بیش‌برازش دارد. با داده‌های بیشتر، مدل‌های پیچیده) مانند (m=19 مجبور به یادگیری الگوی کلی داده‌ها شده و از حفظ کردن نویز و نقاط منفرد منصرف می‌شوند. در نتیجه، فاصله بین خطای آموزش و آزمون کاهش یافته و عملکرد مدل روی داده‌های جدید بهبود می‌یابد. با این حال، انتخاب یک درجه بسیار بالا (مانند ۱۹) حتی با داده‌های زیاد هم می‌تواند منجر به خطای پایه (Bias) بالا شود. بهترین عملکرد مدل نیز در همین نقطه‌ی مصالحه بین پیچیدگی و تعداد داده‌ها حاصل می‌شود.