هوش مصنوعی

مدرس: دکتر عادله بیطرفان[1]
دستیاران آموزشی: هلیا رنجبر[2]، مهرناز حسینی[3] و روژین پناو[4]

۱۲۰ نمره (۲۰ نمره اختیاری) | **پروژه پایانی: فاز دوم** | **موعد تحویل: ۳۰ بهمن ماه**

هدف از این پروژه حل مسئله رگراسیون خطی با استفاده از مسئله بهینه‌سازی است. این پروژه در ۲ فاز طراحی شده است. در فاز اول می‌آموزید هدف از بهینه‌سازی چیست، چه طور انجام می‌شود و رفتار آن را روی یک مجموعه داده کوچک تحلیل خواهید کرد. در فاز دوم مسئله رگراسیون غیر خطی را روی یک مجموعه داده واقعی با استفاده از الگوریتم گرادیان کاهشی که در فاز اول آموختید حل خواهید کرد.

فاز دوم: رگراسیون غیرخطی

بخش۱. مقدمه

در فاز اول با مسئله رگرسیون خطی و الگوریتم گرادیان کاهشی آشنا شدید. دیدیم که چگونه می‌توان یک مدل خطی را با استفاده از بهینه‌سازی و گرادیان کاهشی روی داده‌ها برازش داد. اما در بسیاری از مسائل واقعی، رابطه بین ورودی و خروجی خطی نیست. بنابراین مسئله تبدیل به یک مسئله غیرخطی خواهد شد و باید از مدل‌های غیرخطی کمک گرفت. یکی از ساده‌ترین و در عین حال پرکاربردترین مدل‌ها غیرخطی، مدل‌های چندجمله‌ای[5] هستند.

در این فاز، هدف این است که

- یک مدل غیرخطی تک‌متغیره از نوع چندجمله‌ای را در نظر بگیریم،
- نشان دهیم که این مدل اگرچه نسبت به ورودی غیرخطی است، اما نسبت به پارامترها خطی است،
- و با استفاده از همان الگوریتم گرادیان کاهشی فاز اول، پارامترهای آن را یاد بگیریم.

[1] adeleh.bitarafan@sharif.edu
[2] helia.ranjbar04@ut.ac.ir
[3] mehrnazhosseini4@ut.ac.ir
[4] rozhin.panaw@ut.ac.ir
[5] Polynomial Model

بخش۲. مدل چندجمله‌ای درجه m

فرض کنید نمونه داده ما به صورت زوج مرتب (x, y) باشد. مدل چندجمله‌ای درجه m به صورت زیر تعریف می‌شود:

$$\hat{y} = w_0 + xw_1 + x^2w_2 + x^3w_3 + \cdots + x^m w_m \quad (۱)$$

که در آن

- x ورودی تک ویژگی (تک‌متغیره) است،
- y خروجی واقعی است،
- $\hat{y}$ خروجی پیش‌بینی‌شده توسط مدل است،
- $\{w_0, w_1, w_2, \dots, w_m\}$ پارامترهای مدل هستند.

توجه کنید که اگرچه این مدل نسبت به x غیرخطی است، اما اگر بردار ویژگی زیر را تعریف کنیم:

$$\emptyset(x) = [1, x, x^2, \dots, x^m]^{\boldsymbol{T}} \quad (۲)$$

می‌توان مدل را به صورت برداری زیر نوشت:

$$\hat{y} = \boldsymbol{w}^T \emptyset(x) \quad (۳)$$

که در آن:

$$\boldsymbol{w} = [w_0, w_1, w_2, \dots, w_m]^{\boldsymbol{T}} \quad (۴)$$

در نتیجه، این مدل از دید پارامترها یک مدل خطی محسوب می‌شود و می‌توان آن را دقیقاً مشابه رگراسیون خطی فاز اول آموزش داد.

یافتن بهترین مقادیر پارامترها با استفاده از گرادیان کاهشی

برای به دست آوردن پارامترهای مدل غیرخطی با استفاده از الگوریتم گرادیان کاهشی، کافی است با شروع از یک مقدار اولیه تصادفی برای پارامترها، آن‌ها را با استفاده از فرمول به‌روزرسانی زیر به روزرسانی کنیم تا به همگرایی برسیم:

$$w_i = w_i - \alpha \times \frac{\partial \hat{y}}{\partial w_i}, \qquad i = 1, \dots, m \quad (۵)$$

که در رابطه (۵) گرادیان نسبت به هریک از پارامترها ($\frac{\partial \hat{y}}{\partial w_i}$) با توجه به رابطه (۱) برابر خواهد بود با:

$$\frac{\partial \hat{y}}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(\hat{y} - y_i)$$

$$\frac{\partial \hat{y}}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n} x_i(\hat{y} - y_i) \qquad (۶)$$

$$\frac{\partial \hat{y}}{\partial w_2} = \frac{2}{n}\sum_{i=1}^{n} {x_i}^2(\hat{y} - y_i)$$

$$\dots$$

$$\frac{\partial \hat{y}}{\partial w_m} = \frac{2}{n}\sum_{i=1}^{n} {x_i}^m(\hat{y} - y_i)$$

بنابراین، در این فاز دیگر به دنبال «بهترین خط» نیستیم، بلکه به دنبال بهترین منحنی چندجمله‌ای هستیم که از میان داده‌ها عبور کند. به‌هرحال، تنظیم درجه m مهم است چرا که با افزایش درجه m، خطر بیش برازش[1] ممکن است اتفاق بیفتد.

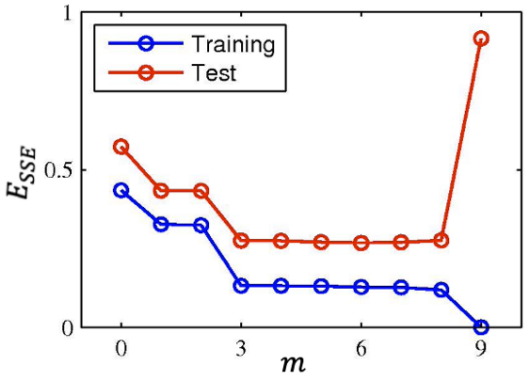
تعریف بیش‌برازش:

بیش‌برازش زمانی رخ می‌دهد که مدل بیش از حد به داده‌های آموزشی[2] «بچسبد» و به جای یاد گرفتن الگوی کلی حاکم بر داده‌ها، نویز و جزئیات تصادفی آن‌ها را نیز یاد بگیرد. در این حالت، اگرچه مدل روی داده‌های آموزش خطای بسیار کمی دارد، اما روی داده‌های جدید (داده‌های دیده‌نشده در طول پیدا کردن پارامترها) عملکرد ضعیفی خواهد داشت. این پدیده معمولاً زمانی اتفاق می‌افتد که مدل بیش از حد پیچیده باشد (مثلاً درجه‌ی چندجمله‌ای خیلی بزرگ انتخاب شود) یا تعداد پارامترها نسبت به تعداد داده‌ها زیاد باشد، و باعث می‌شود مدل به جای تعمیم‌پذیری، صرفاً داده‌های موجود را حفظ کند.

در حالت اتفاق افتادن بیش‌برازش مشاهده خواهید کرد خطا روی داده‌های آموزشی خیلی کم و نزدیک به صفر است ولی خطا روی داده‌های آزمون[3] (داده‌های جدید) خیلی زیاد است. به عنوان مثال، شکل (۱) خطا روی داده‌های آموزش (منحنی آبی رنگ) و خطا روی داده‌های آزمون (منحنی قرمز رنگ) به ازای mهای مختلف را نشان می‌دهد. همان طور که مشاهده می‌شود به ازای $m = ۹$ مشکل بیش‌برازش اتفاق افتاده است.

---

[1] Overfitting
[2] Training Data
[3] Test Data



شکل (۱). نمودار خطا روی داده‌های آموزشی (منحنی آبی رنگ) و خطا روی داده‌های آزمون (منحنی قرمز رنگ) به ازای مقادیر مختلف m.

بخش۳. پیاده‌سازی و تحلیل رفتار گرادیان کاهشی در مدل‌های غیرخطی چندجمله‌ای

در این فاز، شما با یک مجموعه داده واقعی سروکار دارید که شامل ۵۰۰ نمونه است. از این تعداد، ۳۲۰ نمونه به عنوان داده‌های آموزش و ۱۸۰ نمونه به عنوان داده‌های آزمون در نظر گرفته شده‌اند که در قالب چهار ماتریس TrainX، TrainY، TestX و TestY در اختیار شما قرار گرفته است.

تعریف داده‌های آموزشی و آزمایشی: داده‌های آموزش، داده‌هایی هستند که مدل با استفاده از آن‌ها یاد می‌گیرد. یعنی پارامترهای مدل ($\{w_0, w_1, w_2, \dots, w_m\}$) با دیدن این داده‌ها تنظیم می‌شوند. الگوریتم گرادیان کاهشی فقط با استفاده از داده‌های آموزش گرادیان را حساب می‌کند و پارامترها را به‌روزرسانی می‌کند. در مقابل، داده‌های آزمون، داده‌هایی هستند که مدل در طول فرآیند یادگیری آن‌ها را نمی‌بیند. این داده‌ها فقط برای این استفاده می‌شوند که بعد از تمام شدن آموزش (پس از یادگیری و تنظیم مقادیر پارامترها) بررسی کنیم مدل ما روی داده‌های جدید و دیده‌نشده چقدر خوب عمل می‌کند.

چرا داده‌ها را به آموزش و آزمون تقسیم می‌کنیم؟ اگر عملکرد مدل را فقط روی داده‌های آموزش بسنجیم ممکن است مدل صرفاً داده‌ها را حفظ کرده باشد (یعنی بیش‌برازش) و روی داده‌های جدید عملکرد ضعیفی داشته باشد. بنابراین ارزیابی واقعی مدل باید روی داده‌هایی انجام شود که در زمان آموزش دیده نشده‌اند.

درنهایت در این پروژه، هر نمونه داده شامل یک زوج مرتب به صورت (x_i, y_i) است که در آن x_i ورودی مدل است و y_i خروجی واقعی است. در مجموعه داده درنظر گرفته شده، هر داده فقط یک ویژگی[1] یا به عبارتی فقط یک متغیر ورودی دارد. یعنی به ازای هر نمونه، فقط یک عدد x به مدل داده می‌شود و مدل باید تنها بر اساس همین یک عدد، مقدار y_i را پیش‌بینی کند. به چنین مسائلی، مسائل تک‌متغیره یا تک‌ویژگی[2] گفته می‌شود. پس اگرچه مدل ما ممکن است چندین پارامتر داشته باشد (مثلاً در مدل چندجمله‌ای درجه m پارامترهای $\{w_0, w_1, w_2, \dots, w_m\}$)، اما ورودی مدل همچنان فقط یک متغیر x است.

---

[1] Feature

[2] Univariate Regression

بخش۱،۳. بارگذاری و تحلیل اولیه داده‌ها (۱۰ نمره)

الف) مجموعه داده‌ای که در پوشه "data" در اختیار شما قرار داده شده است را بارگذاری کرده و داده‌های آموزشی را در فضای دوبعدی رسم کنید (محور افقی ورودی x_i و محور عمودی خروجی داده‌ها (y_i) را درنظر بگیرید).

ب) توضیح دهید چرا یک مدل چندجمله‌ای می‌تواند برای این داده مناسب باشد.

بخش۲،۳. بررسی تاثیر درجه چندجمله‌ای (۷۰ نمره)

در این بخش فرض کنید تنها ۲۰ داده اول از ۳۲۰ مجموعه داده آموزشی را به عنوان نمونه‌های آموزشی استفاده کنیم. درصورتی که برای رگراسیون از تابع زیان $MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$ استفاده شود، به طوری‌که:

$$\widehat{y_i} = w_0 + x_i w_1 + x_i^2 w_2 + x_i^3 w_3 + \cdots + x_i^m w_m$$

الف) به ازای مقادیر مختلف m در بازه $m = \{۱،۲،۴،۸،۱۰،۱۳،۱۵،۱۷،۱۹\}$، مقدار پارامترهای $\{w_0, w_1, w_2, \dots, w_m\}$ را محاسبه و ذخیره کنید و به سوالات زیر پاسخ دهید:

- منحنی چندجمله‌ای به ازای مقادیر مختلف m را روی ۲۰ داده آموزشی رسم کنید.
- نمودار نرم۲[1] بردار پارامترها ($||\boldsymbol{w}||_2^2$) را به ازای مقادیر مختلف m رسم کنید و تحلیل کنید نمودار نرم۲ پارامترها چگونه تغییر می‌کند.

توضیحات: $||\boldsymbol{w}||_2^2 = ||w_1||^2 + ||w_2||^2 + \cdots + ||w_m||^2$

ب) خطای MSE روی داده‌های آموزشی و داده‌های آزمون را به ازای مقادیر مختلف m همانند شکل (۱) در قالب یک نمودار رسم کنید و به منظور تحلیل آن به سوالات زیر پاسخ دهید:

- چرا خطا به ازای m کوچک روی داده‌های آموزش و آزمون زیاد است؟
- چرا با افزایش مقدار m خطا روی داده‌های آموزش کم می‌شود؟
- چرا خطا به ازای m بزرگ روی داده‌های آموزش کم ولی روی داده‌های آزمون زیاد است؟
- با توجه به نمودار بهترین چند جمله‌ای چه درجه‌ای دارد؟ دلیل خود را توضیح دهید.

---

[1] $L_2\ norm$

ج) بهترین بردار پارامترها را به ازای بهترین درجه چند جمله‌ای پیدا شده در قسمت قبل گزارش کنید.

د) منحنی چندجمله‌ای نهایی با بهترین درجه m را رسم کرده و تحلیل کنید:

- آیا منحنی به دست آمده خوب برازش شده است؟
- خطا بیش‌تر کجا وجود دارد؟
- آیا نشانه‌ای از بیش‌برازش مشاهده می‌شود؟

بخش۳,۳. بررسی تاثیر تعداد داده‌های آموزشی (۴۰ نمره)

در این بخش هدف حل مسئله رگراسیون چند جمله‌ای با درجه $m = ۱۹$ با تعداد داده‌های آموزشی متفاوت است، تا بدین وسیله تاثیر تعداد داده‌های آموزشی را بررسی کنیم.

الف) به ازای تعداد نمونه‌های آموزشی $N \in \{۱۰، ۲۰، ۴۰، ۸۰، ۱۶۰، ۳۲۰\}$ پارامترهای $\{w_0, w_1, w_2, \dots, w_{19}\}$ را پیدا کرده و ذخیره کنید. همچنین مقادیر پارامترها به ازای مقادیر مختلف تعداد نمونه‌های آموزشی را در گزارش خود داخل یک جدول درج کنید و بررسی کنید:

- با افزایش تعداد داده‌های آموزشی چه پارامترهایی مقادیر نزدیک به صفر دارند و مدل چند جمله‌ای نهایی چه درجه‌ای دارد؟

ب) خطای MSE روی داده‌های آموزشی و داده‌های آزمون را به ازای مقادیر مختلف N همانند شکل (۱) در قالب یک نمودار رسم کنید.

ج) با توجه به نمودار حاصل از قسمت قبل تاثیر تعداد داده‌های آموزشی روی مسئله بیش‌برازش را بررسی کنید.

نکات پایانی

- برای پیاده‌سازی تنها از Python استفاده کنید.
- گزارش شما باید کامل و جامع بوده و تمامی فعالیت‌های شما را پوشش دهد. نمایش نتایج و تحلیل‌ها نقش مهمی در ارزیابی فعالیت شما دارند.
- نتایج، گزارش و کدهای خود را در قالب یک فایل فشرده با فرمت AI_Phase2_[stdNumber].zip در کوئرا بارگذاری کنید (در صورت دسترسی نداشتن به کوئرا به دستیاران آموزشی خود میل کنید).
- در صورت کشف هر گونه تقلب (شباهت بالای ۶۰ درصد با کدهای موجود در اینترنت و دوستان خود و یا استفاده از ابزارهای هوش مصنوعی) در هر یک از قسمت‌های پروژه، نمره آن قسمت را از دست خواهید داد.
- امکان ارسال پروژه با تاخیر وجود ندارد.
- در صورت سوال از هر فاز پروژه با مدرس درس و یا دستیاران آموزشی در ارتباط باشید.

موفق باشید