question 1 ;

Can gradient descent get stuck in a local minimum when training a

linear regression model? Why?

answer 1;

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

در یک مدل رگرسیون خطی، تابع هزینه معمولاً به صورت میانگین مربعات خطا تعریف میشود .و در ایت حالت تابع هزینه یک نوع تابع محدب است .و محدب بودن به این معنا است که تابع هزینه دارای یک نقطه مینیمم منحصر به فرد است و درواقع اصلا مینیمم محلی وجود ندارد .در نتیجه گیر کردن در یک مینیمم محلی برای رگرسیون خطی وجود ندارد .زیرا تنها مینیمم آن مینیمم جهانی یا به اصطلاح گلوبال است .

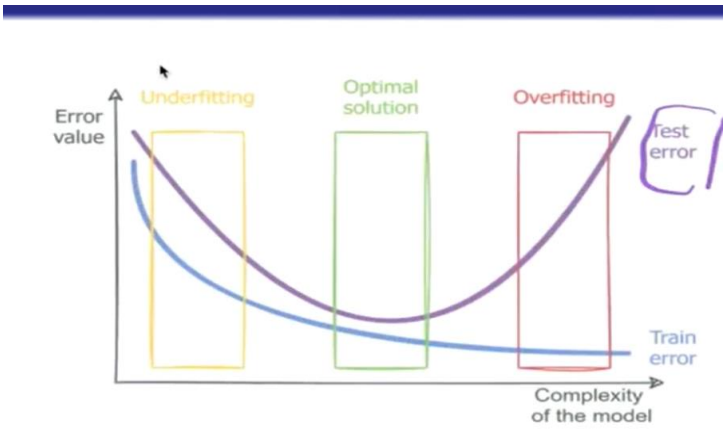
به دلیل شکل محدب تابع هزینه در رگرسیون خطی، هر قدمی که گرادیان نزولی به سمت پایین انجام می‌دهد، همواره به سمت مینیمم جهانی حرکت می‌کند .این رفتار به این دلیل است که گرادیان تابع در همه نقاط به سمت همان مینیمم واحد اشاره دارد.

question 2;

Suppose you are using polynomial regression. You plot the learning

curves and you notice that there is a large gap between the training error and the

validation error. What is happening? What are three ways to solve this?



Adapted from Towards Data Science

. وقتی این اتفاق رخ داده یعنی دچار اورفیتیگ شدیم

در حالت اورفیت، مدل ما توانسته است داده‌های آموزشی را به خوبی یاد بگیرد و خطای آموزشی پایینی داشته باشد، اما در مواجهه .با داده‌های جدید)داده های ولیدشن یا تست(، نمی‌تواند به خوبی عمل کند و خطای بالایی دارد

:راه های رفع مشکل

.با استفاده از متود کرنل سعی در کاهش درجه ی چند جمله ای کنیم .عموما از درجه ی 2 یا 3 استفاده میشود

. داده های آموزشی را افزایش دهیم

. داده ها را قبل از فیت کردن مدل به صورت رندم شافل کنیم

استفاده از منظم سازی

Ridge یا Lasso

دو روش رایج برای کنترل اورفیتینگ هستند .این تکنیک‌ها به مدل جریمه‌ای اعمال می‌کنند که مانع از بزرگ شدن بیش از حد ضرایب می‌شود .این کار به مدل کمک می‌کند که الگوهای عمومی را یاد بگیرد و از یادگیری نویزهای داده‌ها جلوگیری کند

question 3;

In a dataset where the number of predictors exceeds the number of

observations, what problems arise when applying linear regression, and how might

you resolve them?

answer 3;

هنگامی که تعداد ویژگی‌ها)متغیرهای مستقل (در یک مجموعه داده از تعداد نمونه‌ها بیشتر می‌شود، درواقع وارد مسئله ای به اسم <داده هایی با ابعاد بالا میشویم

مشکل داشتن تعداد پیش بینی ها بیشتر از تعداد نمونه ها که معمولا با

P>n

نشان داده میشود این است که در این حالت

راه‌حل منحصربه‌فردی برای مسئله رگرسیون خطی استاندارد وجود ندارد .اگر ردیف‌های ماتریس داده‌ها نشان‌دهنده نمونه‌ها و ستون‌ها نشان‌دهنده پیش‌بینی‌ها باشند، لزوماً بین ستون‌های ماتریس وابستگی‌های خطی وجود دارد .بنابراین، زمانی که ضرایب

n

پیشبینی را پیدا کردیم .ضرایب سایر

p—n

پیش‌بینی ها را می‌توان به صورت ترکیبات خطی دلخواه از آن

n

.پیش‌بینی اول بیان کرد

در شرایط عادی (زمانی که $n > p$ باشد)، هر ستون از ماتریس داده‌ها بیانگر یک ویژگی مستقل است، اما زمانی که $n < p$ باشد، وابستگی‌های خطی بین ویژگی‌ها به وجود می‌آید. به این معنا که برخی از ستون‌های ماتریس را می‌توان با ترکیب خطی ستون‌های دیگر بدست آورد، زیرا تعداد ستون‌ها بیشتر از تعداد ردیف‌ها است. این وضعیت باعث می‌شود که نتوان ضرایب مشخص و منحصربه‌فردی برای همه ویژگی‌ها پیدا کرد.

چالش‌های اصلی

1. **وجود بی‌نهایت جواب:** در این حالت، معادلات نرمال رگرسیون خطی دارای بی‌نهایت جواب هستند. این بدان معناست که می‌توانیم بی‌نهایت خط رگرسیونی را پیدا کنیم که داده‌ها را به خوبی برازش کنند، اما نمی‌توانیم یک جواب واحد و دقیق را انتخاب کنیم.
2. **اورفیتینگ شدید:** مدل‌های رگرسیونی در این شرایط تمایل به یادگیری نویزهای تصادفی در داده‌های آموزشی دارند. این امر منجر به ایجاد مدل‌هایی می‌شود که در مجموعه داده آموزشی عملکرد بسیار خوبی دارند، اما در مواجهه با داده‌های جدید عملکرد ضعیفی از خود نشان می‌دهند.
3. **ناپایداری ضرایب:** تغییرات کوچک در داده‌ها می‌تواند منجر به تغییرات بزرگی در ضرایب مدل شود. این ناپایداری نشان می‌دهد که مدل به شدت به داده‌های آموزشی وابسته است و نمی‌توان به نتایج آن اعتماد کرد.

برای حل مشکل میتوان از رگرسیون لاسو و همینطور

ridge

. استفاده کرد

این روش‌ها با افزودن یک جریمه به ضرایب مدل، از وابستگی‌های خطی زیاد جلوگیری می‌کنند و به مدل کمک می‌کنند تا ضرایب پایدارتری پیدا کند .در این روش‌ها، مدل تلاش می‌کند تا پیش‌بین‌های کم‌اهمیت‌تر را با ضرایب کوچکتر)یا حتی صفر (وزن‌گذاری .کند و در نتیجه مدل ساده‌تر و تعمیم‌پذیرتری ایجاد شود

به علاوه، سایر روش‌های یادگیری ماشین مانند کاهش ابعاد یا انتخاب ویژگی نیز می‌توانند در شرایطی که

p>n

است به مدل کمک کنند تا از پیچیدگی زیاد جلوگیری کند و پاسخی پایدارتر و کارآمدتر ارائه دهد.

question 4;

How do outliers in the response variable (target variable) affect the model in least squares regression? Which alternative techniques could help make the model more robust to outliers?

answer;

، در رگرسیون حدلاقل مربعات

هدف این است که مجموع مربعات اختلافات بین مقادیر پیش‌بینی‌شده و مقادیر واقعی متغیر هدف را کم کنیم .وجود مقادیر پرت می‌تواند تأثیر زیادی بر مدل داشته باشد، زیرا در این روش، خطاها به توان دو می‌رسند و مقادیر پرت به دلیل مقادیر بزرگی که ایجاد می‌کنند، وزن بسیار زیادی در تابع خطا خواهند داشت .این باعث می‌شود که مدل به‌جای یادگیری الگوهای اصلی، بیشتر به سمت .این داده‌های پرت متمایل شود و نتایج پیش‌بینی نادرست و باثبات‌کمتری ایجاد کند

:تاثیرات به طور کلی به این صورت میشود

.داده‌های پرت می‌توانند خط رگرسیون را به سمت خود بکشند و باعث تغییر قابل توجه در شیب خط شوند

.مجموع مربعات خطا را افزایش می‌دهند

.مدلی که تحت تأثیر داده‌های پرت قرار گرفته است، دقت کمتری در پیش‌بینی داده‌های جدید خواهد داشت

:راهکار

استفاده از روش های آماری مثل شاخص پراکندگی یا آی کیو آر

یا

z-score

یا استفاده از قضیه چبیشف

و ..

روش‌های مبتنی بر فاصله :محاسبه فاصله هر داده تا نزدیک‌ترین همسایه و شناسایی داده‌هایی که فاصله زیادی دارند .مثلا با استفاده از متر اقلیدسی

یا

حذف داده های پرت

. یا میتوان از روش رگرسیون کمترین قدر مطلق استفاده کرد

و حتی میتوانیم به داده ها وزن بدهیم و توزیع انها را نرمال کنیم

... و

روش‌های مبتنی بر

SUPPORT VECTOR MACHINE :

.این روش‌ها با ایجاد یک حاشیه بین داده‌ها و خط رگرسیون، کمتر تحت تأثیر داده‌های پرت قرار می‌گیرند

question 6;

How can we use statistical significance tests to determine if one model consistently performs better than another, rather than any differences being due to random chance? Could you give examples of these tests and explain when each one is appropriate for comparing models on metrics like accuracy, F1 score, or error rate?

answer;

:استفاده از آزمون‌های آماری برای مقایسه مدل‌های یادگیری ماشین

هنگامی که چندین مدل یادگیری ماشین داریم، اصولا می‌خواهیم بدانیم آیا تفاوت عملکرد بین آن‌ها به دلیل تصادف است یا یکی از مدل‌ها واقعاً بهتر از دیگری است .در اینجاست که آزمون‌های آماری به کمک ما می‌آیند

1. آزمون t-student:

- **مناسب برای:** مقایسه میانگین دو گروه مستقل (مثلاً میانگین دقت دو مدل)
- **فرضیه‌های صفر و جایگزین:**
  - H_0: میانگین گروه 1 = میانگین گروه 2
  - H_1: میانگین گروه 1 ≠ میانگین گروه 2
- **شرایط استفاده:** داده‌ها باید نرمال توزیع شده باشند و واریانس دو گروه برابر باشد.

2. آزمون Wilcoxon Signed-Rank Test:

- **مناسب برای:** مقایسه میانه دو گروه وابسته (مثلاً میانه دقت یک مدل قبل و بعد از یک تغییر)
- **فرضیه‌های صفر و جایگزین:**
  - H_0: میانه گروه 1 = میانه گروه 2
  - H_1: میانه گروه 1 ≠ میانه گروه 2
- **شرایط استفاده:** داده‌ها باید رتبه‌ای باشند یا می‌توان آن‌ها را به رتبه تبدیل کرد.

3. آزمون McNemar's Test:

- **مناسب برای:** مقایسه دو مدل طبقه‌بندی روی یک مجموعه داده، با توجه به ماتریس درهم‌ریختگی (confusion matrix)
- **فرضیه‌های صفر و جایگزین:**
  - H_0: دو مدل عملکرد مشابهی دارند
  - H_1: دو مدل عملکرد متفاوتی دارند
- **شرایط استفاده:** داده‌ها باید دودویی باشند.

4. آزمون Friedman Test:

- **مناسب برای:** مقایسه چندین مدل روی یک مجموعه داده، با توجه به یک متریک خاص (مثلاً دقت، F1-score)
- **فرضیه‌های صفر و جایگزین:**
  - H_0: همه مدل‌ها عملکرد مشابهی دارند
  - H_1: حداقل یک مدل عملکرد متفاوتی دارد
- **شرایط استفاده:** داده‌ها باید رتبه‌ای باشند یا می‌توان آن‌ها را به رتبه تبدیل کرد.

4. **آزمون‌های بوت‌استرپ (Bootstrap Tests)**:

- **مناسب برای:** شرایطی که داده‌های زیادی دارید و می‌خواهید توزیع‌های نمونه‌ای از معیارهای عملکرد مدل را بسازید.
- **کاربرد:** بوت‌استرپ کردن به این معناست که داده‌ها را چندین بار با جایگزینی بازنمونه‌گیری می‌کنیم و در هر تکرار معیارهایی مانند **دقت** یا **نرخ خطا** را محاسبه می‌کنیم. سپس با استفاده از این توزیع نمونه‌ای از نتایج، **فاصله اطمینان** معیارهای مورد نظر را به دست می‌آوریم. اگر فاصله‌های اطمینان دو مدل هم‌پوشانی نداشته باشند، می‌توان نتیجه گرفت که تفاوت معناداری بین عملکرد مدل‌ها وجود دارد.
- **مزیت‌ها:** این روش هیچ فرض خاصی درباره توزیع داده‌ها نمی‌کند و بنابراین در بسیاری از شرایط مختلف قابل استفاده است.

مراحل انجام آزمون‌های آماری

1. **انتخاب آزمون مناسب:** بر اساس نوع داده‌ها و هدف مقایسه، آزمون مناسب را انتخاب کنید.
2. **تعیین سطح معنی‌داری (α):** معمولاً $\alpha = 0.05$ استفاده می‌شود.
3. **محاسبه آماره آزمون:** با استفاده از فرمول‌های مربوط به آزمون انتخابی، آماره آزمون را محاسبه کنید.
4. **تعیین مقدار بحرانی یا p-value:** با استفاده از جداول آماری یا نرم‌افزارهای آماری، مقدار بحرانی یا p-value را تعیین کنید.
5. **تصمیم‌گیری:** اگر مقدار آماره آزمون از مقدار بحرانی بیشتر باشد یا p-value کمتر از α باشد، فرضیه صفر را رد می‌کنیم و نتیجه می‌گیریم که تفاوت بین مدل‌ها معنی‌دار است.