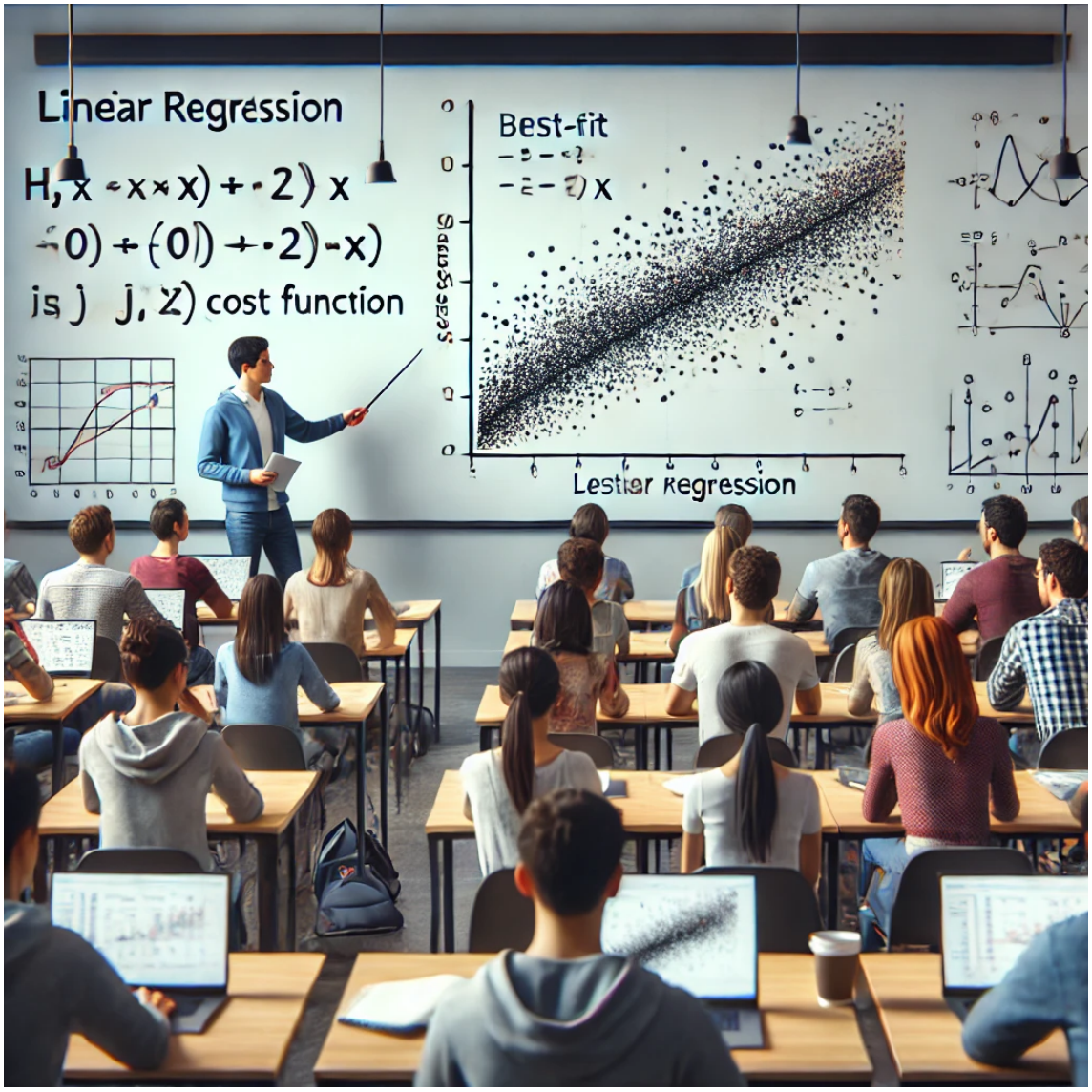
رگرسیون خطی در یادگیری ماشین: مفاهیم، کاربردها و معادلات

دانشگاه تهران
درس مبانی یادگیری الکترونیکی
۱۴۰۴-۱۴۰۳

چکیده

رگرسیون خطی یکی از مهم‌ترین تکنیک‌های یادگیری ماشین در حوزه مدل‌سازی داده‌ها است.
در این گزارش، مفاهیم پایه، کاربردها و معادلات مربوط به رگرسیون خطی بررسی خواهند شد.



فهرست مطالب

۱ مقدمه

رگرسیون خطی linear regression یکی از روش‌های پایه در یادگیری ماشین machine learning است که برای مدل‌سازی روابط بین متغیرهای مستقل و وابسته به کار می‌رود. این روش به ویژه در تحلیل داده‌ها و پیش‌بینی مقادیر عددی اهمیت دارد.

۲ رگرسیون خطی چیست؟

رگرسیون خطی یک روش آماری است که برای مدل‌سازی رابطه بین یک متغیر وابسته و یک یا چند متغیر مستقل استفاده می‌شود. این روش برای پیش‌بینی و تحلیل داده‌ها بسیار مفید است. در یادگیری ماشین، رگرسیون خطی به عنوان یک الگوریتم نظارت‌شده عمل می‌کند که از داده‌های برچسب‌دار یاد می‌گیرد و بهترین تابع خطی را برای پیش‌بینی داده‌های جدید پیدا می‌کند.

۳ اهمیت رگرسیون خطی

یکی از مهم‌ترین ویژگی‌های رگرسیون خطی، تفسیرپذیری آن است. معادله مدل ضرایبی را ارائه می‌دهد که تأثیر هر متغیر مستقل را بر متغیر وابسته نشان می‌دهند. همچنین این روش به دلیل سادگی و شفافیت آن، پایه‌ای برای بسیاری از الگوریتم‌های پیشرفته یادگیری ماشین محسوب می‌شود.

۴ تابع فرضیه رگرسیون خطی (hypothesis function in linear regression)

برای تضمین صحت نتایج مدل، چند فرض اساسی در رگرسیون خطی در نظر گرفته می‌شود:

- خطی بودن: بین متغیرهای مستقل و وابسته، رابطه‌ای خطی وجود دارد.
- استقلال داده‌ها: مشاهدات از یکدیگر مستقل هستند و خطاهای یک مشاهده بر دیگری تأثیر نمی‌گذارند.

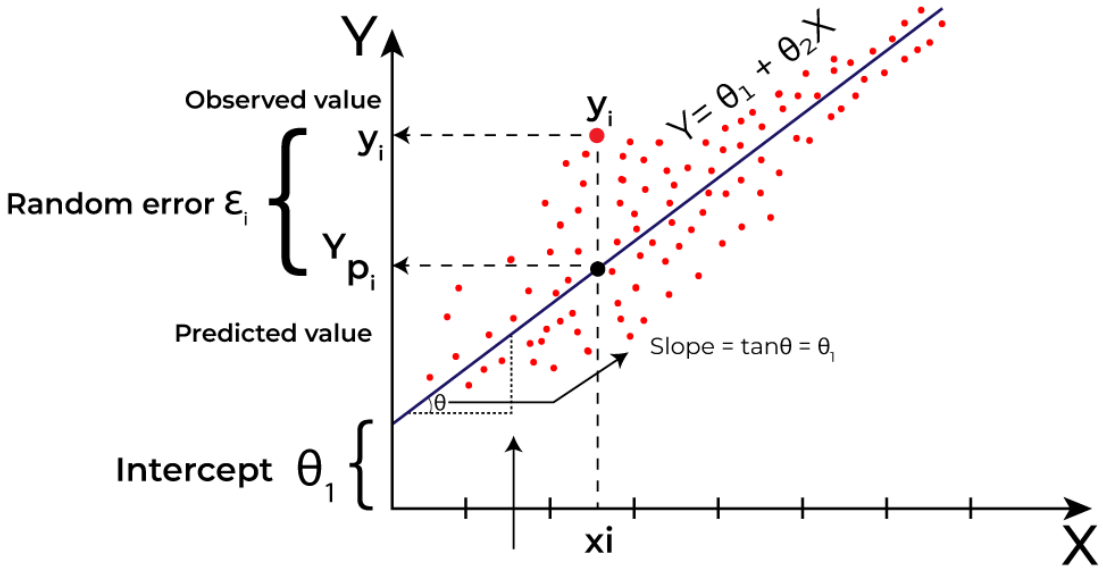
۵ خط برازش بهینه (Best Fit Line)

هدف اصلی در رگرسیون خطی، یافتن بهترین خط برازش است، به طوری که خطای بین مقادیر پیش‌بینی‌شده و مقادیر واقعی به حداقل برسد. معادله خط بهینه به شکل زیر است:

$$\hat{y} = \beta_0 + \beta_1 x \quad (۱)$$

که در آن:

- β_0 عرض از مبدأ است(بایاس هم گفته می‌شود).
- β_1 ضریب متغیر مستقل است.
- $\hat{y}$ مقدار پیش‌بینی شده است.



شکل ۱: رگرسیون خطی

شکل اصلی تابع رگرسیون خطی ساده (نحوه نمایش فرمول می‌تواند متفاوت باشد!):

$$f(x) = \alpha x + \beta$$

در اینجا می‌خواهیم بایاس (α) و شیب (β) را با به حداقل رساندن مشتق تابع مجموع مربعات باقی‌مانده (residual sum of squares(RSS)) پیدا کنیم:
مرحله ۱: RSS داده های آموزشی را محاسبه می‌کنیم:

$$RSS = \Sigma(y_i - (\hat{\beta} + \hat{\alpha} * x_i))^2$$

مرحله ۲: مشتقات تابع RSS را بر حسب α و β محاسبه می‌کنیم و آنها را برابر 0 قرار می‌دهیم تا پارامترهای مورد نظر را پیدا می‌کنیم.

$$\frac{\partial RSS}{\partial \beta} = \Sigma(-f(x_i) + \hat{\beta} + \hat{\alpha} * x_i) = 0$$

$$\rightarrow \beta = \hat{y} - \hat{\alpha}\hat{x} \rightarrow (1)$$

$$\frac{\partial RSS}{\partial \alpha} = \Sigma(-2x_i y_i + 2\hat{\beta}x_i + 2\hat{\alpha}x_i^2) = 0 \rightarrow (2)$$

$$(1), (2) \rightarrow \hat{\alpha} = \frac{\Sigma(x_i - \hat{x})(y_i - \hat{y})}{\Sigma(x_i - \hat{x})^2}$$

$$\hat{\beta} = y - \hat{a}x$$

۶ رگرسیون غیرخطی

علاوه بر رگرسیون خطی، در برخی موارد که رابطه میان متغیرهای مستقل و وابسته پیچیده‌تر است، از رگرسیون غیرخطی استفاده می‌شود. این روش شامل مدل‌هایی مانند رگرسیون چندجمله‌ای، نمایی و لگاریتمی است که امکان مدل‌سازی روابط غیرخطی را فراهم می‌کند.

۷ مثال‌های کاربردی

به عنوان مثال، برای پیش‌بینی قیمت خانه می‌توان از عوامل مختلفی مانند سن ساختمان، فاصله از جاده اصلی، موقعیت مکانی، متراژ و تعداد اتاق‌ها استفاده کرد. رگرسیون خطی رابطه این ویژگی‌ها را با قیمت خانه مدل‌سازی کرده و قیمت‌های آینده را پیش‌بینی می‌کند.

۸ انواع رگرسیون خطی

- رگرسیون خطی ساده: وقتی فقط یک متغیر مستقل در مدل وجود داشته باشد.
- رگرسیون خطی چندمتغیره: وقتی بیش از یک متغیر مستقل در مدل در نظر گرفته شود.

۱.۸ رگرسیون خطی ساده

رگرسیون خطی ساده ساده‌ترین شکل رگرسیون خطی است و تنها شامل یک متغیر مستقل و یک متغیر وابسته می‌شود. معادله رگرسیون خطی ساده به صورت زیر است:

$$\hat{y} = \beta_0 + \beta_1 x \quad (۲)$$

که در آن:

- y متغیر وابسته است.
- x متغیر مستقل است.
- β_0 عرض از مبدأ (intercept) است.
- β_1 شیب (slope) است.

۲.۸ فرضیات رگرسیون خطی ساده

رگرسیون خطی ابزاری قدرتمند برای درک و پیش‌بینی رفتار یک متغیر است، اما برای اینکه دقیق و قابل اعتماد باشد، باید چند شرط اساسی را برآورده کند:

- خطی بودن: رابطه بین متغیر مستقل و وابسته باید خطی باشد. یعنی تغییرات در متغیر وابسته به صورت خطی از تغییرات متغیر مستقل پیروی کند. اگر رابطه خطی نباشد، مدل رگرسیون خطی دقیق نخواهد بود.
- استقلال: مشاهدات در داده‌ها باید مستقل از یکدیگر باشند. یعنی مقدار متغیر وابسته برای یک مشاهده نباید به مقدار متغیر وابسته برای مشاهده دیگر وابسته باشد. اگر مشاهدات مستقل نباشند، مدل رگرسیون خطی دقیق نخواهد بود.
- همسانی واریانس (Homoscedasticity): واریانس خطاها باید در تمام سطوح متغیر مستقل ثابت باشد. این بدان معناست که مقدار متغیر مستقل نباید بر واریانس خطاها تأثیر بگذارد. اگر واریانس باقی‌مانده‌ها ثابت نباشد، مدل رگرسیون خطی دقیق نخواهد بود.
- نرمال بودن باقی‌مانده‌ها: باقی‌مانده‌ها باید به صورت نرمال توزیع شده باشند، یعنی از یک منحنی زنگ‌وله‌ای (نرمال) پیروی کنند. اگر باقی‌مانده‌ها نرمال نباشند، مدل رگرسیون خطی دقیق نخواهد بود.

این فرضیات برای اطمینان از دقت و اعتبار مدل رگرسیون خطی ضروری هستند.

۳.۸ رگرسیون خطی چندگانه

رگرسیون خطی چندگانه شامل بیش از یک متغیر مستقل و یک متغیر وابسته است. معادله رگرسیون خطی چندگانه به صورت زیر است:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \quad (۳)$$

که در آن:

- y متغیر وابسته است.
- $x_1, x_2, , x_n$ متغیرهای مستقل هستند.
- β_0 عرض از مبدأ (intercept) است.
- $\beta_0, \beta_1, ..., \beta_n$ شیب‌ها (slopes) هستند.

هدف الگوریتم یافتن بهترین معادله خط برازش است که بتواند مقادیر را بر اساس متغیرهای مستقل پیش‌بینی کند. در رگرسیون، مجموعه‌ای از رکوردها با مقادیر x و y وجود دارد که از این مقادیر برای یادگیری یک تابع استفاده می‌شود. اگر بخواهید y را از یک x ناشناخته پیش‌بینی کنید، می‌توان از این تابع یادگرفته‌شده استفاده کرد. در رگرسیون، هدف یافتن مقدار y است، بنابراین به یک تابع نیاز داریم که در مورد رگرسیون، y پیوسته را بر اساس x به عنوان ویژگی‌های مستقل پیش‌بینی کند.

۴.۸ فرضیات رگرسیون خطی چندگانه

در رگرسیون خطی چندگانه، تمامی چهار فرضیه رگرسیون خطی ساده (خطی بودن، استقلال، همسانی واریانس، و نرمال بودن باقی‌مانده‌ها) اعمال می‌شوند. علاوه بر این، فرضیات زیر نیز باید رعایت شوند:

- عدم هم‌خطی (No Multicollinearity): بین متغیرهای مستقل نباید همبستگی بالایی وجود داشته باشد. هم‌خطی زمانی رخ می‌دهد که دو یا چند متغیر مستقل به شدت با یکدیگر همبستگی داشته باشند. این موضوع می‌تواند باعث شود که اثر جداگانه هر متغیر بر متغیر وابسته به‌سختی قابل تشخیص باشد. اگر هم‌خطی وجود داشته باشد، مدل رگرسیون خطی چندگانه دقیق نخواهد بود.
- جمع‌پذیری (Additivity): مدل فرض می‌کند که اثر تغییرات در یک متغیر پیش‌بین بر متغیر پاسخ، مستقل از مقادیر سایر متغیرها است. این فرضیه به این معناست که هیچ تعاملی بین متغیرها در تأثیرشان بر متغیر وابسته وجود ندارد.
- انتخاب ویژگی‌ها (Feature Selection): در رگرسیون خطی چندگانه، انتخاب دقیق متغیرهای مستقل برای مدل بسیار مهم است. اضافه کردن متغیرهای نامرتبط یا تکراری می‌تواند منجر به بیش‌برازش (Overfitting) شود و تفسیر مدل را پیچیده کند.
- بیش‌برازش (Overfitting): بیش‌برازش زمانی اتفاق می‌افتد که مدل بیش از حد به داده‌های آموزشی نزدیک شود و نویز یا نوسانات تصادفی را به جای رابطه واقعی بین متغیرها یاد بگیرد. این موضوع می‌تواند باعث کاهش عملکرد مدل در داده‌های جدید و دیده‌نشده شود.

۹ تابع هزینه

تابع هزینه یا تابع زیان به توسعه‌دهندگان کمک می‌کند تا بهترین مقادیر برای پارامترهای مدل (β_0 و β_1) را پیدا کنند و بهترین خط برازش را برای داده‌ها ایجاد کنند. این تابع، خطای بین مقادیر واقعی (y_i)

و مقادیر پیش‌بینی‌شده ($pred_i$) را محاسبه می‌کند و هدف، کمینه‌سازی این خطا است. تابع هزینه به صورت زیر تعریف می‌شود:

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \quad (۴)$$

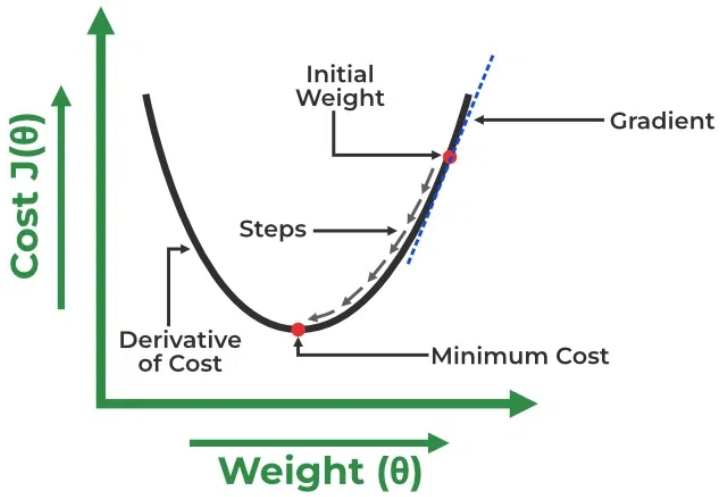
این تابع به عنوان میانگین مربع خطا (Mean Squared Error | MSE) نیز شناخته می‌شود. در این تابع:

- $pred_i$ مقدار پیش‌بینی‌شده توسط مدل است.
- y_i مقدار واقعی است.
- n تعداد نقاط داده است.

هدف این است که با تغییر مقادیر a_0 و a_1، مقدار تابع هزینه (J) به حداقل برسد.

۱۰ گرادیان کاهشی (Gradient Descent)

گرادیان کاهشی یک روش بهینه‌سازی است که برای به‌روزرسانی پارامترهای مدل a۰a۰) و (a۱a۱ و کمینه‌سازی تابع هزینه استفاده می‌شود. ایده اصلی این است که با شروع از مقادیر اولیه برای a_0 و a_1، به‌تدریج این مقادیر را تغییر دهیم تا تابع هزینه کاهش یابد.



شکل ۲: گرادیان کاهشی

۱.۱۰ نرخ یادگیری (Learning Rate):

- نرخ یادگیری (α) اندازه گام‌هایی است که در هر تکرار برداشته می‌شود.
- اگر نرخ یادگیری کوچک باشد، همگرایی به سمت مینیمم کندتر است، اما دقیق‌تر خواهد بود.
- اگر نرخ یادگیری بزرگ باشد، همگرایی سریع‌تر است، اما ممکن است از نقطه مینیمم عبور کند.

فرمول به‌روزرسانی پارامترها به صورت زیر است:

$$\beta_0 = \beta_0 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \quad (۵)$$

$$\beta_1 = \beta_1 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \cdot x_i \quad (۶)$$

۱۱ معیارهای ارزیابی رگرسیون خطی

برای ارزیابی عملکرد مدل‌های رگرسیون خطی، از معیارهای مختلفی استفاده می‌شود. این معیارها نشان می‌دهند که مدل چقدر خوب می‌تواند مقادیر واقعی را پیش‌بینی کند. برخی از رایج‌ترین معیارها عبارتند از:

۱.۱۱ میانگین مربعات خطا (Mean Squared Error | MSE)

این معیار میانگین مربعات اختلاف بین مقادیر واقعی و مقادیر پیش‌بینی‌شده را محاسبه می‌کند. فرمول آن به صورت زیر است:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (۷)$$

- n : تعداد نقاط داده.
- y_i : مقدار واقعی.
- $\hat{y}_i$: مقدار پیش‌بینی‌شده.

ویژگی‌ها:

- به داده‌های پرت (Outliers) حساس است، زیرا خطاهای بزرگ تأثیر زیادی روی نتیجه دارند.
- هرچه مقدار MSE کمتر باشد، مدل دقیق‌تر است.

۲.۱۱ میانگین خطای مطلق (Mean Absolute Error | MAE)

این معیار میانگین اختلاف مطلق بین مقادیر واقعی و مقادیر پیش‌بینی‌شده را محاسبه می‌کند. فرمول آن به صورت زیر است:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (۸)$$

ویژگی‌ها:

- به داده‌های پرت حساس نیست، زیرا از قدر مطلق استفاده می‌کند.
- هرچه مقدار MAE کمتر باشد، مدل دقیق‌تر است.

۳.۱۱ ریشه میانگین مربعات خطا (Root Mean Squared Error | RMSE)

این معیار ریشه دوم میانگین مربعات خطا (MSE) است و نشان می‌دهد که مدل چقدر خوب می‌تواند داده‌ها را پیش‌بینی کند. فرمول آن به صورت زیر است:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (۹)$$

ویژگی‌ها:

- به واحد داده‌ها وابسته است (یک معیار نرمال‌شده نیست).
- هرچه مقدار RMSE کمتر باشد، مدل دقیق‌تر است.

۴.۱۱ ضریب تعیین (R^2 | R-squared)

این معیار نشان می‌دهد که چه مقدار از واریانس متغیر وابسته توسط مدل توضیح داده می‌شود. مقدار آن بین ۰ و ۱ است. فرمول آن به صورت زیر است:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (۱۰)$$

مجموع مربعات باقی‌مانده (RSS): مجموع مربعات اختلاف بین مقادیر واقعی و پیش‌بینی‌شده.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (۱۱)$$

مجموع مربعات کل (TSS): مجموع مربعات اختلاف بین مقادیر واقعی و میانگین آن‌ها.

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad (۱۲)$$

ویژگی‌ها:

- هرچه R^2 به ۱ نزدیک‌تر باشد، مدل بهتر است.
- اگر $R = 1$ باشد، مدل تمام واریانس داده‌ها را توضیح می‌دهد.

۵.۱۱ نتیجه‌گیری

- MSE و RMSE برای اندازه‌گیری خطای مدل استفاده می‌شوند و به داده‌های پرت حساس هستند.
- MAE به داده‌های پرت حساس نیست و خطای مطلق را اندازه‌گیری می‌کند.
- R2 نشان می‌دهد که مدل چقدر خوب واریانس داده‌ها را توضیح می‌دهد.

۱۲ تکنیک‌های منظم‌سازی (Regularization) برای مدل‌های خطی

هدف از تکنیک‌های منظم‌سازی، جلوگیری از بیش‌برازش (Overfitting) در مدل‌های رگرسیون خطی است. این تکنیک‌ها با اضافه کردن یک جمله جریمه (Penalty) به تابع هدف، ضرایب مدل را محدود می‌کنند. سه روش رایج منظم‌سازی عبارتند از:

۱.۱۲ رگرسیون لاسو (Lasso Regression | L1 Regularization)

رگرسیون لاسو با اضافه کردن یک جمله جریمه مبتنی بر مجموع قدر مطلق ضرایب ($L1$)، مدل را تنظیم می‌کند. این روش برای انتخاب ویژگی‌ها (Feature Selection) مفید است، زیرا برخی از ضرایب را دقیقاً صفر می‌کند. تابع هدف:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{n}|\theta_j| \quad (۱۳)$$

- جمله اول: خطای مربعات (Least Squares Loss).
- جمله دوم: جریمه $L1$ (مجموع قدر مطلق ضرایب).
- λ: قدرت تنظیم (Regularization Strength).

ویژگی‌ها:

- برای داده‌هایی با تعداد زیادی ویژگی مفید است.
- برخی از ضرایب را صفر می‌کند و باعث انتخاب ویژگی می‌شود.

۲.۱۲ رگرسیون ریج (Ridge Regression | L2 Regularization)

رگرسیون ریج با اضافه کردن یک جمله جریمه مبتنی بر مجموع مربعات ضرایب ($L2$)، مدل را تنظیم می‌کند. این روش برای داده‌هایی با هم‌خطی (Multicollinearity) بالا مفید است. تابع هدف:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{n}\theta_j^2 \quad (۱۴)$$

- جمله اول: خطای مربعات (Least Squares Loss).
- جمله دوم: جریمه $L2$ (مجموع مربعات ضرایب).
- λ: قدرت تنظیم (Regularization Strength).

ویژگی‌ها:

- برای داده‌هایی با هم‌خطی بالا مناسب است.
- ضرایب را به صفر نزدیک می‌کند، اما دقیقاً صفر نمی‌کند.

۳.۱۲ رگرسیون الاستیک‌نت (Elastic Net Regression)

رگرسیون الاستیک‌نت ترکیبی از تنظیم‌های $L1$ و $L2$ است و مزایای هر دو روش را دارد. این روش برای داده‌هایی که هم‌خطی دارند و همچنین نیاز به انتخاب ویژگی وجود دارد، مناسب است.
تابع هدف:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \alpha\lambda\sum_{j=1}^{n}|\theta_j| + \frac{1}{2}(1-\alpha)\lambda\sum_{j=1}^{n}\theta_j^2 \quad (۱۵)$$

- جمله اول: خطای مربعات (Least Squares Loss).

- جمله دوم: جریمه $L1$ (مجموع قدر مطلق ضرایب).
- جمله سوم: جریمه $L2$ (مجموع مربعات ضرایب).
- λ: قدرت تنظیم (Regularization Strength).
- α: پارامتر ترکیبی که نسبت $L1$ به $L2$ را کنترل می‌کند.

ویژگی‌ها:

- ترکیبی از مزایای لاسو و ریج.
- برای داده‌های با هم‌خطی و نیاز به انتخاب ویژگی مناسب است.

۴.۱۲ نتیجه‌گیری

- لاسو ($L1$): برای انتخاب ویژگی و کاهش ضرایب به صفر مناسب است.
- ریج ($L2$): برای داده‌های با هم‌خطی بالا و کاهش ضرایب به مقادیر کوچک مناسب است.
- الاستیک‌نت: ترکیبی از لاسو و ریج که برای شرایط پیچیده‌تر مناسب است.

این تکنیک‌ها به بهبود عملکرد مدل و جلوگیری از بیش‌برازش کمک می‌کنند.

۱۳ نتیجه‌گیری

رگرسیون خطی یک الگوریتم پایه‌ای در یادگیری ماشین است که به دلیل سادگی، تفسیرپذیری و کارایی بالا، سال‌ها مورد استفاده قرار گرفته است. این روش ابزاری ارزشمند برای درک روابط بین متغیرها و انجام پیش‌بینی‌ها در کاربردهای مختلف محسوب می‌شود. با این حال، آگاهی از محدودیت‌های آن نیز مهم است، از جمله فرض خطی بودن رابطه بین متغیرها و حساسیت به هم‌خطی. در صورتی که این محدودیت‌ها به دقت در نظر گرفته شوند، رگرسیون خطی می‌تواند به عنوان یک ابزار قدرتمند در تحلیل داده‌ها و پیش‌بینی مورد استفاده قرار گیرد.