دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

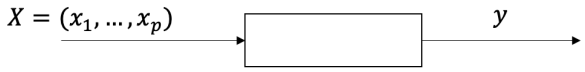
نیم‌سال دوم ۱۴۰۲-۱۴۰۱

مدرس: دکتر سید ابوالفضل مطهری

درس‌نامه چهارم

۱ رگرسیون چند بعدی

در مساله رگرسیون چندبعدی، ورودی یک بردار pـ بعدی و خروجی یک عدد میباشد.



به بعدهای ورودی:

- Input
- Predictor
- Feature
- Variable Independent

و به بعد خروجی اصطلاحات:

- Output
- Response
- Variable Dependent

اطلاق میگردد.

معمولا داده را به صورت یک ماتریس نشان میدهیم:

x_1	x_2	...	x_p	y
		...		

مدل به صورت زیر در نظر گرفته میشود:

$$\hat{y} = \beta_{۰} + \sum_{j=۱}^{p} \beta_j x_j \tag{۱}$$

تابع ریسک (یا خسارت) را نیز به صورت:

$$l(y, \hat{y}) = (y - \hat{y})^{۲} \tag{۲}$$

تعریف میکنیم. بنابراین برای کمینه کردن ریسک (یا خسارت) از تعریف زیر استفاده میکنیم:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=۱}^{n} (y_i - \hat{y}_i)^{۲} = \underset{\beta}{\operatorname{argmin}}\, \mathbf{RSS}(\beta) \tag{۳}$$

که در آن داریم:

$$\beta = (\beta_{۰}, \beta_{۱}, ..., \beta_p)^T \tag{۴}$$

و

$$\mathbf{RSS}(\beta) = \sum_{i=۱}^{n} (y_i - \beta_{۰} - \sum_{i=۱}^{p} \beta_j x_{ij})^{۲}. \tag{۵}$$

رابطه فوق را میتوان به صورت ماتریسی نیز نشان داد. اگر ماتریس X را به صورت زیر تعریف کنیم:

$$X = \begin{bmatrix} ۱ & x_{۱۱} & x_{۱۲} & x_{۱۳} & \dots & x_{۱p} \\ ۱ & x_{۲۱} & x_{۲۲} & x_{۲۳} & \dots & x_{۲p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ ۱ & x_{n۱} & x_{n۲} & x_{n۳} & \dots & x_{np} \end{bmatrix} \tag{۶}$$

و بردار y را به صورت زیر تعریف کنیم:

$$y = (y_{۱}, ..., y_n)^T \tag{۷}$$

میتوانیم عبارت RSS را به صورت زیر بازنویسی کنیم:

$$\mathbf{RSS}(\beta) = (y - X\beta)^T (y - X\beta) = ||y - X\beta||_{۲}^{۲} \tag{۸}$$

در نتیجه اگر نسبت به $\hat{\beta}$ مشتق بگیریم و در $\hat{\beta}$ برابر صفر قرار دهیم:

$$\frac{\partial RSS}{\partial \beta}|_{\beta=\hat{\beta}} = -۲X^T(y - X\hat{\beta}) = ۰ \quad (۹)$$

به دستگاه خطی زیر می‌رسیم:

$$X^T X\hat{\beta} = X^T y \quad (۱۰)$$

اگر مشتق دوم RSS را محاسبه کنیم:

$$\frac{\partial^۲ RSS}{\partial \beta \partial \beta^T} = ۲X^T X \quad (۱۱)$$

خواهیم دید مشتق دوم، به ازای تمام βـ ها یک ماتریس مثبت نیمه معین است. (چرا؟) بنابراین تابع $\mathbf{RSS}(\beta)$ یک تابع محدب است و جواب‌های ۱۰ همگی نقاط بهینه می‌باشند.

اگر X دارای رتبه ستونی کامل باشد، آنگاه $X^T X$ رتبه کامل (Full Rank) خواهد بود. پس $\hat{\beta}$ یک جواب یکتا دارد.

در این حالت:

$$\hat{\beta} = (X^T X)^{-۱} X^T y \quad (۱۲)$$

و مقدار $\hat{y}$ به صورت زیر قابل محاسبه است.

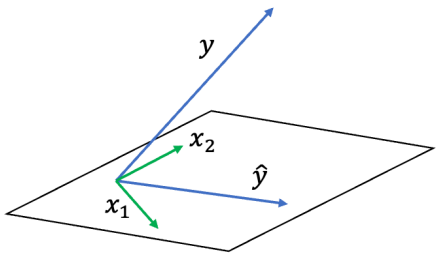
$$\hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-۱} X^T}_{H} y \quad (۱۳)$$

۲ دیدگاه هندسی

در حقیقت مساله از دیدگاه هندسی بدین صورت است که می‌خواهیم یک بردار $\hat{y}$ را که از ترکیب خطی ستونهای X حاصل میشود، طوری بیابیم که کمترین فاصله را با y داشته باشد. این بدین معنی است که باید y را روی فضایی که با ستونهای X ساخته میشود، تصویر نماییم:

$$\hat{y} = P_{col(۱,x_۱,\ldots,x_p)}(y) \quad (۱۴)$$

ماتریس H در حقیقت این عمل را انجام میدهد.



در شرایطی ممکن است ستونهای ماتریس X مستقل خطی نباشد. در این شرایط هرچند $\hat{y}$ مقداری مشخص است اما برای $\hat{\beta}$ جوابهای متعددی وجود خواهد داشت.

یکی از راههای مقابله با این مساله این است که ستونهایی که به کمک ستونهای دیگر قابل بدست آمدن میباشد (ترکیب خطی ستونهای دیگر است) را دور بریزیم تا ماتریس X دارای رتبه ستونی کامل شود.

۳ ارزیابی

اگر m نمونه را کنار گذاشته باشیم، میتوانیم تعمیم را مقایسه ریسک (یا خسارت) درون داده و بیرون داده مورد آزمون قرار دهیم.

اگر مدل واقعی به صورت زیر باشد:

$$y = \underbrace{(1, x_1, \ldots, x_p)}_{X^T} \beta^* + \epsilon \tag{۱۵}$$

خواهیم داشت:

$$\begin{aligned}
\mathbb{E}[l(y, \hat{y})] &= \mathbb{E}[(X^T(\beta^* - \hat{\beta}) + \epsilon)^2] && (۱۶)\\
&= \mathbb{E}[\{(X^T(\beta^* - \mathbb{E}[\hat{\beta}]) + X^T(\mathbb{E}[\hat{\beta}] - \hat{\beta}) + \epsilon\}^2] && (۱۷)\\
&= ||X^T(\beta^* - \mathbb{E}[\hat{\beta}])||_2^2 + \mathbb{E}[||X^T(\mathbb{E}[\hat{\beta}] - \hat{\beta})||_2^2] + \sigma_\epsilon^2 && (۱۸)\\
&= \mathbf{Bias}^2 + \mathbf{Variance} + \mathbf{Noise} && (۱۹)
\end{aligned}$$

به این صورت میتوان ریسک (خسارت) را به سه جز مذکور تجزیه کرد. (Bias-Variance Decomposition)

۴ استنتاج

اگر $\epsilon_i \sim \mathcal{N}(۰, \sigma_\epsilon^۲)$ را فرض کنیم، خواهیم داشت:

$$\hat{\beta} = (X^T X)^{-۱} X^T y \tag{۲۰}$$
$$= (X^T X)^{-۱} X^T (X\beta^* + \epsilon) \tag{۲۱}$$
$$= \beta^* + (X^T X)^{-۱} X^T \epsilon \tag{۲۲}$$

بنابراین با فرض دانستن X داریم:

$$\hat{\beta} \sim N(\beta^*, (X^t X)^{-۱} \epsilon) \tag{۲۳}$$

بنابراین، $\hat{\beta}$ دارای توزیع گوسی میباشد. تخمین $\sigma_\epsilon^۲$ را نیز میتوان به صورت زیر محاسبه کرد:

$$\hat{\sigma_\epsilon^۲} = \frac{\mathbf{RSS}}{n-p-۱} = \frac{۱}{n-p-۱} \sum_{i=۱}^{n} (y_i - \hat{y}_i)^۲ \tag{۲۴}$$