0 Lineární regrese

V modelu lineární regrese předpokládáme lineární závislost vysvětlované proměnné na hodnotách příznaků.

0.1 Model

Pro data s hodnotami příznaků $X_1, \ldots, X_p$ a hodnotou vysvětlované proměnné Y předpokládáme lineární model

$$Y = w_0 + w_1 X_1 + \ldots + w_p X_p + \varepsilon$$

kde w_i jsou neznámé koeficienty a ε představuje chybu nebo nekonzistenci výsledné hodnoty Y, kterou model nezachycuje a pro kterou platí $\mathrm{E}\,\varepsilon = 0$.

V bodě $(x_1, \ldots, x_p)^T$ tohoto modelu platí vztah

$$Y = w_0 + w_1 x_1 + \ldots + w_p x_p + \varepsilon = \boldsymbol{w}^T \boldsymbol{x} + \varepsilon$$

kde zavádíme následující vektorovou notaci pro vstup $\boldsymbol{x}$ a vektor vah $\boldsymbol{w}$

$$\boldsymbol{x} = (1, x_1, \ldots, x_p)^T$$
$$\boldsymbol{w} = (w_0, w_1, \ldots, w_p)^T$$

0.2 Predikce

S odhadnutými váhami $\hat{\boldsymbol{w}}$ predikujeme vztahem

$$\hat{Y} = \hat{\boldsymbol{w}}^T \boldsymbol{x} = \hat{w}_0 + \hat{w}_1 x_1 + \ldots + \hat{w}_p x_p$$

Pro skutečnou hodnoty $Y = \boldsymbol{w}^T \boldsymbol{x} + \varepsilon$ platí z předpokladu $\mathrm{E}\,\varepsilon = 0$

$$\mathrm{E}\,Y = \mathrm{E}\,\boldsymbol{w}^T \boldsymbol{x} + \mathrm{E}\,\varepsilon = \boldsymbol{w}^T \boldsymbol{x}$$

$\hat{Y}$ je tedy bodovým odhadem EY v bodě $\boldsymbol{x}$.

0.3 Metoda nejmenších čtverců

Metoda nejmenších čtverců nalézá hodnotu $\hat{\boldsymbol{w}}$ odhadu $\boldsymbol{w}$ minimalizací následující kvadratické ztrátové funkce:

$$L(Y, \hat{Y}) = (Y - \hat{Y})^2$$

Pro trénovací množinou $(\boldsymbol{x}_i, Y_i), i = 1, \ldots, N$ minimalizujeme reziduální součet čtverců

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} L(Y_i, \hat{Y}_i) = \sum_{i=1}^{N} (Y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

0.3.1 Maticový zápis trénovací množiny

Zavádíme náhodné vektory $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T, \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ a body $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ zapisujeme do řádků matice

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,p} \end{pmatrix} \in \mathbb{R}^{N,p+1}$$

Při tomto značení platí rovnice $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$ kde $\mathrm{E}\,\boldsymbol{\varepsilon} = \boldsymbol{0}$.

0.3.2 Minimalizace RSS

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} (Y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

Začneme nalezením gradientu této funkce

$$\begin{aligned} \frac{\partial\,\mathrm{RSS}(\boldsymbol{w})}{\partial w_j} &= \sum_{i=1}^{N} 2(Y_i - \boldsymbol{w}^T \boldsymbol{x}_i)(-x_{i,j}) \\ \nabla\,\mathrm{RSS}(\boldsymbol{w}) &= \sum_{i=1}^{N} 2(Y_i - \boldsymbol{w}^T \boldsymbol{x}_i)(-\boldsymbol{x}_i) \\ &= -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) \end{aligned}$$

Položením $\nabla\,\mathrm{RSS}(\boldsymbol{w}) = \boldsymbol{0}$ obdržíme normální rovnici

$$\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{0}.$$

Předpokládáme-li, že $\boldsymbol{X}^T\boldsymbol{X}$ je regulární, pak lze $\boldsymbol{w}$ odhadnout následovně

$$\hat{\boldsymbol{w}}_{\mathrm{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

V případě regulární matice $\boldsymbol{X}^T\boldsymbol{X}$ je $\hat{\boldsymbol{w}}_{\mathrm{OLS}}$ jediným kritickým bodem. Z geometrické interpretace, kterou dosáhneme stejného výsledku, bude plynout, že se jedná o globální minimum $\mathrm{RSS}(\boldsymbol{w})$.

Predikce $\hat{Y}$ v bodě $\boldsymbol{x}$ je $\hat{Y} = \hat{\boldsymbol{w}}_{\mathrm{OLS}}^T\boldsymbol{x}$.

0.3.3 Geometrická interpretace

Minimalizace $\mathrm{RSS}(\boldsymbol{w}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}\|^2$ je problém ekvivalentní minimalizaci $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}\|$. Hledáme tedy $\boldsymbol{w} \in \mathbb{R}^{p+1}$ takové, že vektory $\boldsymbol{Y}$ a $\boldsymbol{X}\boldsymbol{w}$ jsou si v prostoru $\mathbb{R}^N$ co nejblíže.

$\boldsymbol{X}\boldsymbol{w}$ je lineární kombinací sloupců $\boldsymbol{X}$:

$$\boldsymbol{X}\boldsymbol{w} = \sum_{i=0}^{p} \boldsymbol{w}_i \boldsymbol{X}_{:,i} \in \langle \boldsymbol{X}_{:,0}, \boldsymbol{X}_{:,1}, \ldots, \boldsymbol{X}_{:,p} \rangle = \boldsymbol{P}$$

Hledaný bod $\boldsymbol{X}\boldsymbol{w}$ je proto nejblíže k bodu $\boldsymbol{Y}$, právě pokud je vektor $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}$ ortogonální na podprostor $\boldsymbol{P}$:

$$\boldsymbol{X}_{:,i}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) = 0 \text{ pro všechna } i = 0, 1, \ldots, p$$

Což po přepsání do maticového tvaru dá opět normální rovnici

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{0}$$

Z úvah také navíc plyne, že každé $\hat{\boldsymbol{w}}$, které splňuje normální rovnici, RSS($\boldsymbol{w}$) minimalizuje (jedná se o globální minimum).

0.4 Regularita versus lineární nezávislost sloupců matice X

Pro libovolné $\boldsymbol{s} \in \mathbb{R}^{p+1}$ platí následující posloupnost implikací:

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{s} = \boldsymbol{0} \Rightarrow \boldsymbol{s}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{s} = \|\boldsymbol{X}\boldsymbol{s}\|^2 = 0 \Rightarrow \boldsymbol{X}\boldsymbol{s} = \boldsymbol{0} \Rightarrow \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{s} = \boldsymbol{0}$$

Z toho vyplývá, že $\boldsymbol{X}^T\boldsymbol{X}$ je regulární právě tehdy, když $\boldsymbol{X}\boldsymbol{s} = \boldsymbol{0}$ pouze pro $\boldsymbol{s} = \boldsymbol{0}$, což platí právě tehdy, když jsou sloupce matice $\boldsymbol{X}$ lineárně nezávislé. To zřejmě nemusí vždy platit: např. když $N < p + 1$ nebo pokud je nějaký příznak lineární kombinací ostatních.

Normální rovnice má v případě regulární $\boldsymbol{X}^T\boldsymbol{X}$ právě jedno řešení.

$$\hat{\boldsymbol{w}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

V opačném případě má jich má nekonečně mnoho.

0.5 Problém kolinearity

Pokud je nějaký příznak "skoro" lineární kombinací ostatních příznaků, pak můžeme narazit na problém kolinearity. V případě silně korelovaného příznaku je příslušná proměnná do jisté míry zbytečná, protože ve svém rozměru (dimenzi) nepřidává žádnou informaci navíc a zároveň může mít ve výsledném vektoru $\hat{\boldsymbol{w}}_{\text{OLS}}$ vysoký koeficient. Jako příklad lze uvést 3-dimenzionální prostor (X_1, X_2, Y) ve kterém řešení degeneruje do skoro-přímky, kterou lze protnou rovinou více způsoby.

0.5.1 Důsledky kolinearity

Ve výsledku kolinearita způsobuje vysoký rozptyl $\hat{\boldsymbol{w}}_{\text{OLS}}$, který je tak velmi citlivý na data. Pro různé realizace stejného náhodného výběru se řešení normální rovnice může značně lišit. Vysoký rozptyl se následovně přenáší na predikce, které jsou pak méně spolehlivé.

0.5.2 Řešení problému kolinearity

Řešením by bylo odstranit příznaky, které kolinearity způsobují. To však není vždy jednoduché. S vyšším počtem příznaků je velmi obtížné takové sloupce identifikovat. Také se může stát, že kolinearita je mezi velkým počtem příznaků, které jsou dohromady klíčové pro správnou predikci. V takovém případě není ideální takové příznaky zahodit. Existují však metody, které dokážou počet příznaků snížit (odstraněním, případně nahrazením menším počtem) tak, aby byly sloupce LN.

Je také možné změnit funkci, kterou minimalizujeme, abychom měli stabilnější a jednoznačnější řešení. Typicky se přidává regulační člen, který problémy kolinearity může zmírnit (hřebenová regrese v sekci **??** nebo lasso v sekci **??**).

0.6 Shrnutí

Lineární regrese je rezistentní vůči problémům spojené s vysokou dimenzí dat. Problém nastává, když je dat méně než příznaků nebo když jsou příznaky silně korelované.