VZD - Lineární regrese, regularizace pomocí hřebenové regrese.

Vastl Martin

May 27, 2019

1 Lineární regrese

Cílem lineární regrese je predikovat hodnotu Y na základě příznaků $X_1, X_2, \ldots, X_p$. Při lineární regresi předpokládáme lineární závislost vysvětlované proměnné na příznacích.

Z důvodu toho, že tato závislost není perfektní tedy, že nečekáme, že pro stejné příznaky $X_1, X_2, \ldots, X_p$ nedostaneme stejné Y modelujeme tuto závislost následovně.

$$Y = w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \varepsilon, \tag{1}$$

kde $w_1, w_2, \ldots, w_p$ jsou nějaké neznáme koeficienty a ε je náhodná veličina, která není vysvětlitelná za pomocí hodnot příznaků nebo příznaky neznámé nebo cíleně nezahrnované a je tedy z našeho pohledu náhodná.

Obvykle ještě oddělujeme střední hodnotu náhodných vlivů a dostáváme tak:

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \varepsilon, \tag{2}$$

kde $E\varepsilon = 0$ a w_0 se nazývá intercept a odpovídá očekávané výchozí hodnotě při nulových příznacích.

V případě zavedení $x_0 = 1$ a značení $x = (x_0, x_1, \ldots, x_p)^T$ a $w = (w_0, w_1, \ldots, w_p)^T$ lze zkráceně psát

$$Y = w^T x + \varepsilon \tag{3}$$

Pro konkrétní bod x je skutečná hodnota Y určena vztahem

$$Y = w^T x + \varepsilon \tag{4}$$

a je tedy náhodnou veličinou. Z předpoklad $\mathrm{E}\varepsilon = 0$ plyne, že $\mathrm{E}\, Y = w^T x$ a $\hat{Y}$ je tedy bodovým odhadem střední hodnoty $\mathrm{E}Y$ v bodě x.

2 Odhad parametrů

Cílem je nalézt hodnotu vektoru w, tak aby byla chyba modelu co nejmenší. Tuto hodnotu pak použijeme jako odhad $\hat{w}$. Chybovou funkci modelu měříme za pomocí nějaké nezáporné funkce $\mathrm{L} : \mathbb{R}^2 \to \mathbb{R}$, kterou nazýváme ztrátovou funkcí, kterou aplikujeme na skutečnou hodnotu proměnné Y a odpovídající predikci $\hat{Y}$. Obvyklou ztrátovou funkcí je kvadratická ztrátová funkce $\mathrm{L}(Y, \hat{Y}) = (Y - \hat{Y})^2$.

Součtem chyb přes všechny body trénovací množiny je tedy:

$$\text{RSS(w)} = \sum_{i=1}^{N} L(Y_i, w^T x_i) = \sum_{i=1}^{N} (Y_i - w^T x_i)^2, \tag{5}$$

který se nazývá reziduální součet čtverců. Minimalizaci tohoto výrazu získáme odhad $\hat{w}$. Tento postup se nazývá metoda nejmenších čtverců.

Vstupní data lze přepsat jako:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,p} \end{bmatrix} \tag{6}$$

$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)^T$ a $Y = (Y_1, Y_2, \dots, Y_N)$ Při tomto značení můžeme celkový model trénovací množiny zapsat jako:

$$Y = Xw + \varepsilon \tag{7}$$

2.1 Minimalizace RSS

RSS lze vyjádřit jako

$$\text{RSS}(w) = \sum_{i=1}^{N} (Y_i - w^T x_i)^2 = ||Y - Xw||^2 \tag{8}$$

Nejdříve je nutné začít parciálními derivacemi podle $w_0, w_1, \dots, w_p$

$$\frac{\partial \text{RSS}}{\partial w_j} = \sum_{i=1}^{N} 2(Y_i - w^T x_i)(-x_{i,j}) \tag{9}$$

Pro gradient se tedy získá

$$\nabla \text{RSS} = -\sum_{i=1}^{N} 2(Y_i - w^T x_i) x_i = -2X^T(Y - Xw) \tag{10}$$

a položíme-li $\nabla \text{RSS} = 0$ získáme tzv. normální rovnici

$$\begin{aligned} -2X^T(Y - Xw) &= 0 \\ X^T(Y - Xw) &= 0 \\ X^T Y - X^T X w &= 0 \end{aligned} \tag{11}$$

Při výpočtu Hessovi matice použijeme

$$\frac{\partial^2 \text{RSS}}{\partial w_k \partial w_j} = \sum_{i=1}^{N} 2(-x_{i,k})(-x_{i,j}) = 2X^T X, \tag{12}$$

dále pro každé $s \in \mathbb{R}^{p+1}$ platí

$$s^T(X^T X)s = (Xs)^T(Xs) = ||Xs||^2 \geq 0, \tag{13}$$

tedy je semi-definitní. Dle 1 proto nastává minimum v jakémkoliv bodě, který řeší normální rovnici $X^T Y - X^T X w = 0$.

Předpokládejme nyní, že X^TX je regulární matice. Normální rovnici lze upravit na $X^TY = X^TXw$, potom je jednoznačné řešení:

$$\hat{w}_{OLS} = (X^TX)^{-1}X^TY \tag{14}$$

V případě, že jsou sloupce skoro lineárně závislé nastávají problémy s výpočtem X^TX, které jsou numericky nestabilní. Tomuto se lze vyhnout pomocí trénování s využitím gradientního sestupu.

$$w^{(i+1)} = w^{(i)} - \alpha \cdot \nabla \mathrm{RSS}(w^{(i)}) = w^{(i)} + \alpha \cdot 2X^T(Y - Xw^{(i)}) \tag{15}$$

3 Hřebenová regrese

V případě, kdy matice X není lineární závislá, pak ani součin X^TX není regulární. Pokud součin není regulární, pak normální rovnice $\hat{w}_{OLS} = (X^TX)^{-1}X^TY$ nemá jednoznačné řešení resp. má nekonečně mnoho řešení. Pro libovolné dvě řešení w a w' platí $X(w - w') = 0$. Stejný problém nastává i v případě kolinearity, tedy, že jsou skoro lineárně závislé. Z toho důvodu byl navržen způsob regularizace za pomocí přidání nového členu do ztrátové funkce, která se nazývá regularizovaný reziduální součet čtverců.

$$\mathrm{RSS}_\alpha(w) = ||Y - Xw||^2 + \alpha \sum_{i=1}^{p} w_i^2, \tag{16}$$

kde α je parametr. Pro $\alpha = 0$ dostáváme klasický RSS. Pro $\alpha > 0$ je vidět, že se snaží cílit aby hodnoty w, byly co nejmenší. Hodnota w_0 se nepenalizuje, protože se jedná pouze o posun.

Po této úpravě získáme

$$\mathrm{RSS}_\alpha(w) = ||Y - Xw||^2 + \alpha w^T I' w, \tag{17}$$

kde prime značí diagonální matici, která má na pozici $x_{0,0}$ hodnotu 0. Gradient je tedy $\nabla \mathrm{RSS}_\alpha(w) = -2X^T(Y - Xw) + 2\alpha I'w$ a Hessova matice $H_{\mathrm{RSS}_\alpha}(w) = 2X^TX + 2\alpha I'$. Ekvivalentem normální rovnice je

$$X^TY - X^TXw - \alpha I'w = 0 \tag{18}$$

a řešením je tedy

$$\hat{w}_\alpha = (X^TX + \alpha I')^{-1}X^TY, \tag{19}$$

které má pro $\alpha > 0$ jednoznačné řešení.

3.1 Bias-variance tradeoff

Jelikož $Y = Xw + \varepsilon$ z trénovací množiny je v důsledku náhodnosti ε náhodný vektor, dostáváme, že i $\hat{w}_\alpha = (X^TX + \alpha I')^{-1}X^TY$ je jakožto funkce Y náhodný vektor. Uvažujme nějaký pevný bod $x = (1, x_1, x_2, \ldots, x_p)$ a zkoumejme očekávanou chybu. Z předpokladu nezávislosti trénovacích a testovacích dat, tj. nezávislost $\hat{Y}$ a Y. Z toho plyne

$$\begin{aligned} E((Y - EY)(EY - \hat{Y})) &= E(Y(EY) - (Y\hat{Y}) - (EY)^2 + (EY)\hat{Y}) = \\ &= (EY)^2 - E(Y\hat{Y}) - (EY)^2 + EYE\hat{Y} = -E(Y\hat{Y}) + EYE\hat{Y} = 0 \end{aligned} \tag{20}$$

Pro očekávanou chybu tedy platí

$$EL(Y, \hat{Y}) = E(Y - \hat{Y})^2 = E(Y - EY + EY - \hat{Y})^2 = E(Y - EY)^2 + E(\hat{Y} - EY)^2 \tag{21}$$

označíme-li $varY = var\varepsilon = \sigma^2$ dostáváme $EL(Y, \hat{Y}) = \sigma^2 + E(\hat{Y} - EY)^2$ První člen je chyba, kterou nelze odstranit, která je dána náhodností v modelu. Druhý člen je MSE a nazývá se střední kvadratická chyba odhadu Y parametru EY.

$$
\begin{aligned}
\mathsf{MSE}(\hat{Y}) &= \mathrm{E}(\hat{Y} - \mathrm{E}\,Y)^2 = \mathrm{E}(\mathrm{E}\,\hat{Y} - \mathrm{E}\,Y + \hat{Y} - \mathrm{E}\,\hat{Y})^2 \\
&= \mathrm{E}(\mathrm{E}\,\hat{Y} - \mathrm{E}\,Y)^2 + \mathrm{E}(\hat{Y} - \mathrm{E}\,\hat{Y})^2 + 2\,\mathrm{E}(\hat{Y} - \mathrm{E}\,\hat{Y})(\mathrm{E}\,\hat{Y} - \mathrm{E}\,Y) \\
&= (\mathrm{E}\,\hat{Y} - \mathrm{E}\,Y)^2 + \mathrm{E}(\hat{Y} - \mathrm{E}\,\hat{Y})^2 + 2 \cdot 0 \cdot (\mathrm{E}\,\hat{Y} - \mathrm{E}\,Y) \\
&= (\mathrm{E}\,\hat{Y} - \mathrm{E}\,Y)^2 + \operatorname{var}\hat{Y} = (\operatorname{bias}\hat{Y})^2 + \operatorname{var}\hat{Y},
\end{aligned}
$$

kde $\operatorname{bias}\hat{Y} = \mathrm{E}\,\hat{Y} - \mathrm{E}\,Y$ značí vychýlení odhadu (angl. bias).

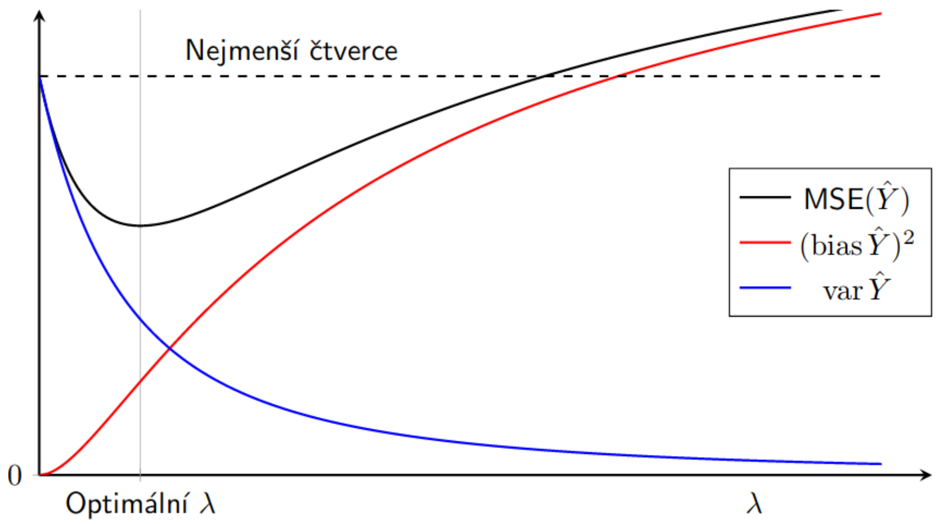
Dohromady tedy máme finální dekompozici očekávané chyby jako

$$
\mathrm{E}\,L(Y, \hat{Y}) = \sigma^2 + (\operatorname{bias}\hat{Y})^2 + \operatorname{var}\hat{Y}.
$$

U hřebenové regrese lze ukázat, že (hodně zjednodušeně) platí

$$
(\operatorname{bias}\hat{Y})^2 \sim \left(1 - \frac{1}{1+\lambda}\right)^2 \quad \text{a} \quad \operatorname{var}\hat{Y} \sim \left(\frac{1}{1+\lambda}\right)^2.
$$

To znamená, že s rostoucím λ vychýlení roste a rozptyl klesá. Takovéto chování v závislosti na hyperparametrech modelu je typické a nazývá se **bias-variance tradeoff**.



3.2 Extrémy funkce více proměnných

Gradient

Definice 1 *Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných, která má v bodě $a \in \mathbb{R}^d$ konečné všechny parciální derivace. Gredient funkce f v bodě a definujeme jako vektor*

$$\nabla f(a) = \Big(\frac{\partial f}{\partial x_1}(a), \dots, \frac{\partial f}{\partial x_d}(a)\Big). \tag{22}$$

Označením ∇f pak myslíme gradient funkce jakožto zobrazení, které každému bodu, kde to lze, přiřadí gradient v tomto bodě.

Důležitou vlastností gradientu je, že ukazuje směr maximálního růstu funkce v daném bodě.

Hessova matice

Definice 2 *Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných. Hessovu matici funkce f v bodě $a \in \mathbb{R}^d$ definujeme jako:*

$$H_f(a) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \dots \\ \frac{\partial^2 f}{\partial x_1 \partial x_d} & \dots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}, \tag{23}$$

kde $\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \big(\frac{\partial}{\partial x_i}\big(\frac{\partial}{\partial x_j}\big)\big)(a)$ značí druhou parciální derivaci podle x_j a x_i.

Věta

Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných a bod $\boldsymbol{x}^* \in \mathbb{R}^d$ takový, že $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

- *Jestliže*
  $$\boldsymbol{s}^T \mathbf{H}_f(\boldsymbol{x}^*)\boldsymbol{s} > 0, \quad \textit{pro každé} \quad \boldsymbol{s} \in \mathbb{R}^d, \boldsymbol{s} \neq \mathbf{0},$$
  nabývá funkce f v bodě $\boldsymbol{x}^*$ ostrého lokálního minima.
- *Jestliže pro každé $\boldsymbol{x}$ z nějakého okolí bodu $\boldsymbol{x}^*$*
  $$\boldsymbol{s}^T \mathbf{H}_f(\boldsymbol{x})\boldsymbol{s} \geq 0, \quad \textit{pro každé} \quad \boldsymbol{s} \in \mathbb{R}^d,$$
  nabývá funkce f v bodě $\boldsymbol{x}^*$ neostrého lokálního minima.

Figure 1: Existence minima