

Co bude v dnešní přednášce

- Hřebenová regrese
- Vztah vychýlení a rozptylu
- Modely bazových funkcí

27 Hřebenová regrese

Hřebenová regrese - úvod

Hřebenová regrese (angl. *ridge regression*) [Hoerl, Kennard (1970)] nebo taky L_2 regularizace se k problému kolinearity staví zavedením penalizačního členu úměrného kvadrátu normy vektoru koeficientů \mathbf{w} s vynecháním interceptu.

Minimalizujeme tedy *regularizovaný reziduální součet čtverců*

$$\text{RSS}_\lambda(\mathbf{w}) = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{i=1}^p w_i^2,$$

který závisí na parametru $\lambda \geq 0$.

- Pro $\lambda = 0$ dostáváme $\text{RSS}_0(\mathbf{w}) = \text{RSS}(\mathbf{w})$ a máme tedy obyčejnou metodu nejmenších čtverců.
- Pro $\lambda > 0$ je vidět, že v minimu se bude cílit na takové vektory \mathbf{w} , které mají co nejmenší složky.
- Hodnotu w_0 interceptu nijak nepenalizujeme. Jedná se pouze o vertikální posun, který zajišťuje předpoklad $E\varepsilon = 0$ modelu a je tedy vhodné ho neomezovat.
- Stále platí, že model pro Y v bodě \mathbf{x} je $Y = w_0 + w_1x_1 + \dots + w_px_p + \varepsilon$.

Hřebenová regrese - normální rovnice

Zavedeme-li matici

$$\mathbf{I}' = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{p+1, p+1},$$

můžeme psát

$$\text{RSS}_\lambda(\mathbf{w}) = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{i=1}^p w_i^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \mathbf{w}^T \mathbf{I}' \mathbf{w}.$$

Gradient je

$$\nabla \text{RSS}_\lambda(\mathbf{w}) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{w}) + 2\lambda \mathbf{I}' \mathbf{w}.$$

Ekvivalent *normální rovnice* je tedy

$$\mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \mathbf{w} - \lambda \mathbf{I}' \mathbf{w} = \mathbf{0}.$$

Hřebenová regrese - odhad parametrů

Hessova matice je dále

$$\mathbf{H}_{\text{RSS}_\lambda}(\mathbf{w}) = 2\mathbf{X}^T\mathbf{X} + 2\lambda\mathbf{I}' = 2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}').$$

Pro každé $\mathbf{s} \in \mathbb{R}^{p+1}$, $\mathbf{s} \neq \mathbf{0}$ a $\lambda > 0$ platí

$$\mathbf{s}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}')\mathbf{s} = (\mathbf{X}\mathbf{s})^T(\mathbf{X}\mathbf{s}) + \lambda\mathbf{s}^T\mathbf{I}'\mathbf{s} = \|\mathbf{X}\mathbf{s}\|^2 + \lambda\sum_{i=1}^p s_i^2 > 0,$$

protože pro $\mathbf{s} = (s_0, 0, \dots, 0)^T \neq \mathbf{0}$ máme $\mathbf{X}\mathbf{s} = (s_0, \dots, s_0)^T \neq \mathbf{0}$.

Hessova matice je tedy pozitivně definitní a matice $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}'$ je regulární.

Pro $\lambda > 0$ tak vždy existuje jednoznačné řešení normální rovnice

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}')^{-1}\mathbf{X}^T\mathbf{Y}$$

a odpovídá globálnímu minimu RSS_λ .

Predikce v bodě \mathbf{x} je potom opět $\hat{Y} = \mathbf{x}^T\hat{\mathbf{w}}_\lambda$.

28 Vztah vychýlení a rozptylu

Očekávaná chyba modelu

V modelu pro trénovací množinu, $\mathbf{Y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, je $\boldsymbol{\varepsilon}$ náhodný vektor, pro který předpokládáme $E\boldsymbol{\varepsilon} = \mathbf{0}$.

Z toho plyne, že i \mathbf{Y} je náhodný vektor a tedy i odhad vektoru parametrů $\hat{\mathbf{w}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}')^{-1}\mathbf{X}^T\mathbf{Y}$ je náhodný vektor.

Uvažujme nyní nějaký pevný bod $\mathbf{x} = (1, x_1, \dots, x_p) \in \mathbb{R}^{p+1}$ a zkoumejme očekávanou chybu měřenou pomocí kvadratické ztrátové funkce při predikci $Y = \mathbf{x}^T\mathbf{w} + \varepsilon$ pomocí $\hat{Y} = \mathbf{x}^T\hat{\mathbf{w}}_\lambda$.

Budeme předpokládat nezávislost trénovacích a testovacích dat, tj. nezávislost \mathbf{Y} a Y , a v důsledku tedy nezávislost \hat{Y} a Y .

Z toho plyne

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

Rozklad očekávané chyby modelu

Pro očekávanou chybu tedy platí

$$\begin{aligned} EL(Y, \hat{Y}) &= E(Y - \hat{Y})^2 = E(Y - EY + EY - \hat{Y})^2 \\ &= E(Y - EY)^2 + 2E((Y - EY)(EY - \hat{Y})) + E(\hat{Y} - EY)^2 \\ &= E(Y - EY)^2 + E(\hat{Y} - EY)^2. \end{aligned}$$

Označíme-li $\text{var } Y = \text{var } \varepsilon = \sigma^2$ dostáváme

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

První člen odpovídá neodstranitelné chybě, která je dána náhodností v modelu. Tato chyba se nazývá Bayesovská (*Bayes error*). Druhý člen se značí $\text{MSE}(\hat{Y})$ a nazývá střední kvadratická chyba odhadu \hat{Y} parametru EY (angl. *mean squared error*).

Rozklad očekávané chyby modelu

Pro $\text{MSE}(\hat{Y})$ dále platí:

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

kde $\text{bias } \hat{Y} = E\hat{Y} - EY$ značí **vychýlení odhadu** (angl. **bias**). Dohromady tedy máme finální dekompozici očekávané chyby jako

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

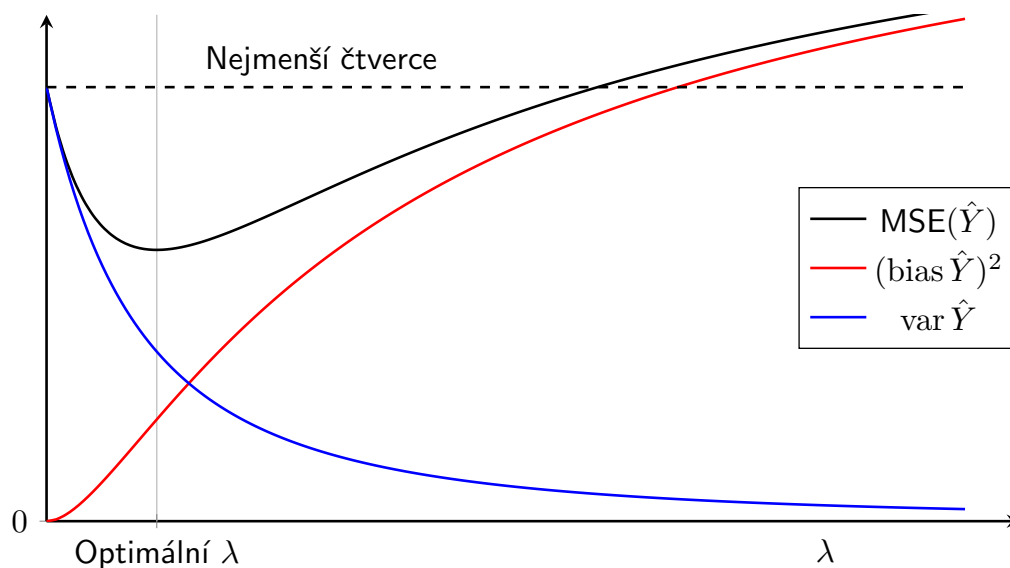
Očekávaná chyba modelu je tedy součtem *neodstranitelné chyby*, kvadrátu *vychýlení odhadu* a *rozptylu odhadu*.

Vztah vychýlení a rozptylu

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

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

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*.



Různé poznámky

- Hledáme tedy optimální hodnotu parametru λ , pro kterou je chyba modelu nejmenší.
- Obvykle se snažíme minimalizovat odhad MSE validační množiny dat případně odhad MSE pomocí cross-validace. Odhad MSE se pro validační množinu (Y'_i, \mathbf{x}'_i) velikosti n počítá jako

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y'_i - \mathbf{x}'_i{}^T \hat{\mathbf{w}}_\lambda)^2.$$

- Při použití hřebenové regrese bývá obvyklé nejprve jednotlivé příznaky standardizovat, aby se staly rozsahově porovnatelné a tedy, aby byly penalizovány všechny stejně. Tj. místo příznaku X_i použijeme příznak

$$X'_i = \frac{X_i - \bar{X}_i}{\sqrt{s_{X_i}^2}}, \quad \text{kde } \bar{X}_i = \frac{1}{N} \sum_{j=1}^N x_{j;i} \quad \text{a} \quad s_{X_i}^2 = \frac{1}{N-1} \sum_{j=1}^N (x_{j;i} - \bar{X}_i)^2$$

- Existují i další možnosti regularizace jako např. $\lambda \sum_{i=1}^p |w_i|$ (Lasso).

29 Modely bazových funkcí

Modely bazových funkcí (1/2)

- Doposud jsme uvažovali pouze jednoduchý lineární model ve tvaru

$$Y = \mathbf{x}^T \mathbf{w} + \varepsilon.$$

- Principiálně jsme tak schopni modelovat pouze lineární funkci ve vstupních proměnných. Ukažme si, jak rozšířit naše možnosti za obzor linearity.
- Základní rozšíření spočívá v nahrazení původních příznaků jejich transformovanými variantami.
- Pro $M \in \mathbb{N}$ vezměme M funkcí $\varphi_1, \dots, \varphi_M$ z \mathbb{R}^p do \mathbb{R} reprezentujících transformace X a nazvěme je *bazové funkce* (angl. *basis functions*).
- K těmto funkcím přidáme $\phi_0(\mathbf{x}) = 1$ a poskládáme je do vektorové funkce $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^{M+1}$ vztahem $\varphi(\mathbf{x}) = (1, \varphi_1(\mathbf{x}), \dots, \varphi_M(\mathbf{x}))^T$.
- Jako model vztahu Y a \mathbf{x} budeme uvažovat lineární model

$$Y = \sum_{j=0}^M w_j \varphi_j(\mathbf{x}) + \varepsilon = \varphi(\mathbf{x})^T \mathbf{w} + \varepsilon,$$

- Další postup je nyní zcela analogický jako před tím.

Modely bazových funkcí (2/2)

- Mějme tedy trénovací množinu jako náhodný výběr z výše uvedeného modelu určený N páry typu (Y_i, \mathbf{x}_i) .
- Maticově můžeme zapsat model pro trénovací data jako $\mathbf{Y} = \Phi \mathbf{w} + \varepsilon$, kde

$$\Phi = \begin{pmatrix} \varphi(\mathbf{x}_1)^T \\ \vdots \\ \varphi(\mathbf{x}_N)^T \end{pmatrix} = \begin{pmatrix} 1 & \phi_1(\mathbf{x}_1) & \cdots & \phi_M(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathbf{x}_N) & \cdots & \phi_M(\mathbf{x}_N) \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

- Obecně budeme minimalizovat (pro $\lambda = 0$ máme metodu nejmenších čtverců)

$$\text{RSS}_\lambda(\mathbf{w}) = \|\mathbf{Y} - \Phi \mathbf{w}\|^2 + \lambda \mathbf{w}^T \mathbf{\Gamma} \mathbf{w}.$$

- Řešením je (pro $\lambda = 0$ značíme také $\hat{\mathbf{w}}_{\text{OLS}}$)

$$\hat{\mathbf{w}}_\lambda = \left(\Phi^T \Phi + \lambda \mathbf{\Gamma} \right)^{-1} \Phi^T \mathbf{Y}.$$

- Predikce hodnoty Y v bodě \mathbf{x} je potom určena vztahem

$$\hat{Y} = \varphi(\mathbf{x})^T \hat{\mathbf{w}}_\lambda.$$

Bázové funkce

Mezi obvyklé volby bázových funkcí patří:

- $\varphi(\mathbf{x}) = x_i$ – přímo jednotlivé příznaky.
- $\varphi(\mathbf{x}) = x_i^2, \varphi(\mathbf{x}) = x_k x_\ell$ – mocniny příznaků a jejich různé součiny, odpovídá polynomiální regresi.
- $\varphi(\mathbf{x}) = \log(x_i), \sqrt{x_i}, \sin(x_i)$ atd. – nelineární transformace jednotlivých příznaků.
- $\varphi(\mathbf{x}) = \mathbb{1}_{(a,b)}(x_i)$, kde $\mathbb{1}_A(x) = 1$ pokud $x \in A$ a $\mathbb{1}_A(x) = 0$ pokud $x \notin A$ – indikátory množin. Umožňují rozdělení prostoru příznaků na kousky a následné modelování v každém kousku zvlášť.
- $\varphi(\mathbf{x}) = h(\|\mathbf{x} - \mathbf{x}_i\|)$, kde \mathbf{x}_i je i -tý trénovací bod a h je nějaká funkce – tzv. *radiální bázové funkce* centrované v bodech trénovací množiny.

Pokud nemáme žádné speciální znalosti o systému, který modelujeme, typicky na počátku volíme velké množství bázových funkcí a používáme hřebenovou regresi, případně jinou formu regularizace.

30 Statistické vlastnosti lineární regrese

Nestrannost metody nejmenších čtverců

Jak jsme již zmínili, odhad vektoru parametrů $\hat{\mathbf{w}}_\lambda = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ je náhodný vektor. Je to tedy *bodový odhad* vektoru parametrů \mathbf{w} a je příkladem tzv. statistiky¹.

Věta 30.1

Odhad $\hat{\mathbf{w}}_{\text{OLS}}$ získaný metodou nejmenších čtverců je za předpokladu $E \boldsymbol{\varepsilon} = \mathbf{0}$ *nestranný*, tj. $E \hat{\mathbf{w}}_{\text{OLS}} = \mathbf{w}$.

Důkaz. Z linearitě střední hodnoty plyne:

$$E \mathbf{Y} = E(\mathbf{X} \mathbf{w} + \boldsymbol{\varepsilon}) = \mathbf{X} \mathbf{w} + E \boldsymbol{\varepsilon} = \mathbf{X} \mathbf{w}.$$

Dále platí:

$$\begin{aligned} E \hat{\mathbf{w}}_{\text{OLS}} &= E (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E \mathbf{Y} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{w}. \end{aligned}$$

Nestrannost metody nejmenších čtverců

Z nestrannosti odhadu vektoru parametrů $\hat{\mathbf{w}}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ dále plyne nestrannost odhadu \hat{Y} v bodě \mathbf{x} .

K tomu je třeba si uvědomit, že pomocí $\hat{Y} = \mathbf{x}^T \hat{\mathbf{w}}_{\text{OLS}}$ se sice snažíme predikovat skutečnou hodnotu Y , ale ze statistického pohledu se jedná o *bodový odhad* střední hodnoty $E Y = \mathbf{x}^T \mathbf{w} + E \boldsymbol{\varepsilon} = \mathbf{x}^T \mathbf{w}$.

Z předchozí věty tak plyne

$$E \hat{Y} = E \mathbf{x}^T \hat{\mathbf{w}}_{\text{OLS}} = \mathbf{x}^T E \hat{\mathbf{w}}_{\text{OLS}} = \mathbf{x}^T \mathbf{w} = E Y$$

a \hat{Y} je tedy *nestranným odhadem* $E Y$.

To samozřejmě znamená, že

$$\text{bias } \hat{Y} = E \hat{Y} - E Y = 0,$$

t.j. vychýlení je 0.

ChangeLog

¹Funkce od náhodného výběru (v našem případě trénovací množiny), viz BI-PST.

Verze	Datum	Autor	Log
1.0	27.10.2022	DV	Výchozí verze pro rok 2022/2023.
1.0	16.11.2022	DV	Výchozí verze pro rok 2023/2024.