BI-ML1.21 přednáška 7

Daniel Vašata

FIT ČVUT

16. 11. 2023

Autoři: Karel Klouda, Daniel Vašata.
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 16. listopadu 2023 14:18.

Co bude v dnešní přednášce

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

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 $\boldsymbol{w}$ s vynecháním interceptu.

Minimalizujeme tedy **regularizovaný reziduální součet čtverců**

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

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

- Pro $\lambda = 0$ dostáváme $\mathrm{RSS}_0(\boldsymbol{w}) = \mathrm{RSS}(\boldsymbol{w})$ a máme tedy obyčejnou metodu nejmenších čtverců.
- Pro $\lambda > 0$ je vidět, že v minimu se bude cílit na takové vektory $\boldsymbol{w}$, které mají co nejmenší složky.
- Hodnotu w_0 interceptu nijak nepenalizujeme. Jedná se pouze o vertikální posun, který zajišťuje předpoklad $\mathrm{E}\,\varepsilon = 0$ modelu a je tedy vhodné ho neomezovat.
- Stále platí, že model pro Y v bodě $\boldsymbol{x}$ je $Y = w_0 + w_1 x_1 + \ldots w_p x_p + \varepsilon$.

Hřebenová regrese - normální rovnice

Zavedeme-li matici

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

můžeme psát

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

Gradient je

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

Ekvivalent **normální rovnice** je tedy

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

Hřebenová regrese - odhad parametrů

Hessova matice je dále

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

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

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

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

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

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

a odpovídá globálnímu minimu RSS_λ.

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

Očekávaná chyba modelu

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

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

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

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

Z toho plyne

$$\begin{aligned}
\mathrm{E}\left((Y - \mathrm{E}\,Y)(\mathrm{E}\,Y - \hat{Y})\right) &= \mathrm{E}\left(Y(\mathrm{E}\,Y) - (Y\hat{Y}) - (\mathrm{E}\,Y)^2 + (\mathrm{E}\,Y)\hat{Y}\right) \\
&= (\mathrm{E}\,Y)^2 - \mathrm{E}(Y\hat{Y}) - (\mathrm{E}\,Y)^2 + \mathrm{E}\,Y\,\mathrm{E}\,\hat{Y} \\
&= -\mathrm{E}(Y\hat{Y}) + \mathrm{E}\,Y\,\mathrm{E}\,\hat{Y} = 0.
\end{aligned}$$

Rozklad očekávané chyby modelu

Pro očekávanou chybu tedy platí

$$\begin{aligned}\mathrm{E}\, L(Y,\hat{Y}) &= \mathrm{E}(Y-\hat{Y})^2 = \mathrm{E}(Y-\mathrm{E}\,Y+\mathrm{E}\,Y-\hat{Y})^2 \\ &= \mathrm{E}(Y-\mathrm{E}\,Y)^2 + 2\,\mathrm{E}\left((Y-\mathrm{E}\,Y)(\mathrm{E}\,Y-\hat{Y})\right) + \mathrm{E}(\hat{Y}-\mathrm{E}\,Y)^2 \\ &= \mathrm{E}(Y-\mathrm{E}\,Y)^2 + \mathrm{E}(\hat{Y}-\mathrm{E}\,Y)^2.\end{aligned}$$

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

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

První člen odpovídá neodstranitelné chybě, která je dána náhodností v modelu. Tato chyba se nazývá Bayesovská (**Bayes error**).

Druhý člen se značí $\mathrm{MSE}(\hat{Y})$ a nazývá střední kvadratická chyba odhadu $\hat{Y}$ parametru $\mathrm{E}\,Y$ (angl. **mean squared error**).

Rozklad očekávané chyby modelu

Pro MSE($\hat{Y}$) dále platí:

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

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

Dohromady tedy máme finální dekompozici očekávané chyby jako

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

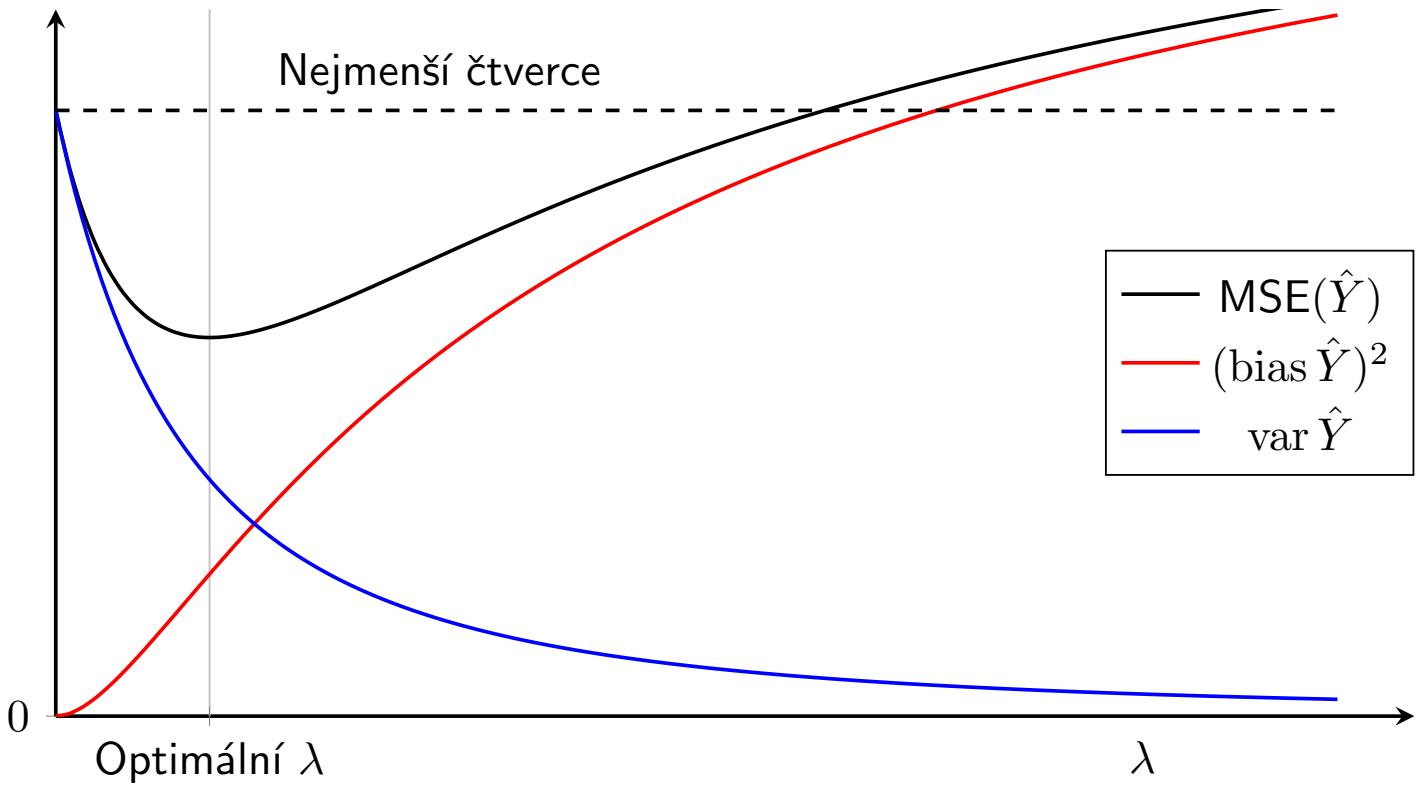
Očekávaná chyba modelu je tedy součtem **neodstranitelné chyby**, kvadrátu **vychýlení** odhadu a **rozptylu** odhadu.

Vztah vychýlení a rozptylu

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

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

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**.



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 $(Y'_i, \boldsymbol{x}'_i)$ velikosti n počítá jako

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y'_i - \boldsymbol{x}'^T_i \hat{\boldsymbol{w}}_\lambda\right)^2.$$

- 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

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

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

Modely bázových funkcí (1/2)

- Doposud jsme uvažovali pouze jednoduchý lineární model ve tvaru

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

- 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 $\varphi_1, \ldots, \varphi_M$ z $\mathbb{R}^p$ do $\mathbb{R}$ reprezentujících transformace X a nazvěme je **bázové funkce** (angl. **basis functions**).
- K těmto funkcím přidáme $\phi_0(\boldsymbol{x}) = 1$ a poskládáme je do vektorové funkce $\boldsymbol{\varphi} : \mathbb{R}^p \to \mathbb{R}^{M+1}$ vztahem $\boldsymbol{\varphi}(\boldsymbol{x}) = (1, \varphi_1(\boldsymbol{x}), \ldots, \varphi_M(\boldsymbol{x}))^T$.
- Jako model vztahu Y a $\boldsymbol{x}$ budeme uvažovat lineární model

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

- Další postup je nyní zcela analogický jako před tím.

Modely bázových funkcí (2/2)

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

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

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

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = \|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 + \lambda \boldsymbol{w}^T \mathbf{I}' \boldsymbol{w}.$$

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

$$\hat{\boldsymbol{w}}_\lambda = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \mathbf{I}'\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{Y}.$$

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

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

Bázové funkce

Mezi obvyklé volby bázových funkcí patří:

- $\varphi(\boldsymbol{x}) = x_i$ – přímo jednotlivé příznaky.
- $\varphi(\boldsymbol{x}) = x_i^2$, $\varphi(\boldsymbol{x}) = x_k x_\ell$ – mocniny příznaků a jejich různé součiny, odpovídá polynomiální regresi.
- $\varphi(\boldsymbol{x}) = \log(x_i), \sqrt{x_i}, \sin(x_i)$ atd. – nelineární transformace jednotlivých příznaků.
- $\varphi(\boldsymbol{x}) = \mathbb{1}_{(a,b)}(x_i)$, 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.
- $\varphi(\boldsymbol{x}) = h(\|\boldsymbol{x} - \boldsymbol{x}_i\|)$, kde $\boldsymbol{x}_i$ 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.

Nestrannost metody nejmenších čtverců

Jak jsme již zmínili, odhad vektoru parametrů $\hat{\boldsymbol{w}}_\lambda = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ je náhodný vektor.

Je to tedy **bodový odhad** vektoru parametrů $\boldsymbol{w}$ a je příkladem tzv. statistiky[1].

Věta

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

Důkaz.

Z linearity střední hodnoty plyne:

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

Dále platí:

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

□

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

Nestrannost metody nejmenších čtverců

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

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

Z předchozí věty tak plyne

$$\mathrm{E}\,\hat{Y} = \mathrm{E}\,\boldsymbol{x}^T\hat{\boldsymbol{w}}_{\text{OLS}} = \boldsymbol{x}^T\,\mathrm{E}\,\hat{\boldsymbol{w}}_{\text{OLS}} = \boldsymbol{x}^T\boldsymbol{w} = \mathrm{E}\,Y$$

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

To samozřejmě znamená, že

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

t.j. vychýlení je 0.