

BI-VZD přednáška 6

Alexander Kovalenko

FIT ČVUT

28. 03. 2022

Autoři: Karel Klouda, Juan Pablo Maldonado Lopez, Daniel Vašata.

Problémy, návrhy apod. hlase v [GitLabu](#).

Verze souboru: 28. března 2022 10:41.

Co bude v dnešní přednášce

- Připomenutí lineární regrese
- Geometrická interpretace metody nejmenších čtverců
- Problém lineárně závislých sloupců
- Regularizace pomocí hřebenové regrese
- Modely bázových funkcí

Lineární model

- Model pro vysvětlovanou proměnnou Y v bodě \mathbf{x} je

$$Y = \mathbf{w}^T \mathbf{x} + \varepsilon = \mathbf{x}^T \mathbf{w} + \varepsilon = w_0 + w_1 x_1 + \dots + w_p x_p + \varepsilon.$$

- Model pro trénovací množinu tvořenou N páry (Y_i, \mathbf{x}_i) je $Y_i = \mathbf{x}_i^T \mathbf{w} + \varepsilon_i$.
- Dohromady při značení $\mathbf{x}_i = (1, x_{i;1}, \dots, x_{i;p})^T$ tedy

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{1;1} & x_{1;2} & \cdots & x_{1;p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N;1} & x_{N;2} & \cdots & x_{N;p} \end{pmatrix} \begin{pmatrix} w_0 \\ \vdots \\ w_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

- Toto zapisujeme maticově jako $\mathbf{Y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, kde

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1;1} & \cdots & x_{1;p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N;1} & \cdots & x_{N;p} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

- Naměřené hodnoty jednotlivých příznaků X_1, \dots, X_p spolu s přidaným umělým příznakem $X_0 = 1$ jsou tedy uvedeny ve sloupcích matice \mathbf{X} .

Metoda nejmenších čtverců

- Při trénování minimalizujeme residuální součet čtverců

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

- Minimum je určeno řešením **normální rovnice**

$$\mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{0},$$

která odpovídá podmínce $\nabla \text{RSS}(\mathbf{w}) = \mathbf{0}$.

- Za předpokladu, že je matice $\mathbf{X}^T \mathbf{X}$ regulární, existuje jediné řešení minimalizující $\text{RSS}(\mathbf{w})$,

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

- Predikce v bodě \mathbf{x} je potom $\hat{Y} = \mathbf{x}^T \hat{\mathbf{w}}_{\text{OLS}}$.

Geometrická interpretace metody nejmenších čtverců (1/3)

- Minimalizace $\text{RSS}(\mathbf{w}) = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2$ je ekvivalentní minimalizaci $\|\mathbf{Y} - \mathbf{X}\mathbf{w}\|$.
- To znamená, že pro optimální \mathbf{w} je Eukleidovská vzdálenost bodů \mathbf{Y} a $\mathbf{X}\mathbf{w}$ v prostoru \mathbb{R}^N nejmenší možná.
- Označíme-li i -tý sloupec matice \mathbf{X} jako $\mathbf{X}_{\bullet i}$, můžeme si všimnout, že

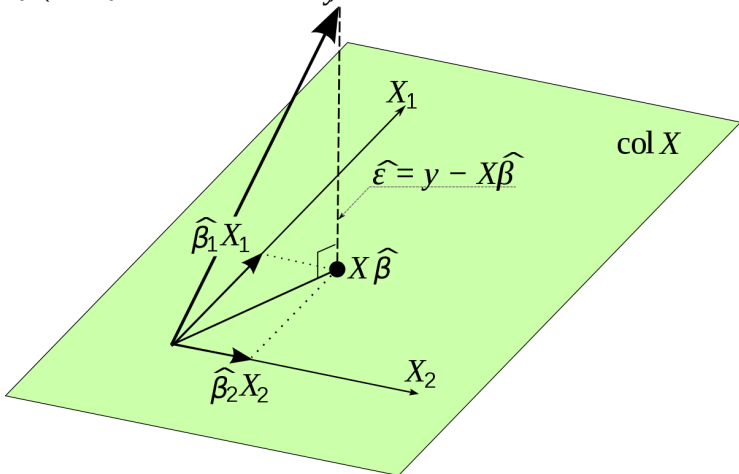
$$\mathbf{X}\mathbf{w} = w_0\mathbf{X}_{\bullet 0} + w_1\mathbf{X}_{\bullet 1} + \dots + w_p\mathbf{X}_{\bullet p}.$$

- Vektor $\mathbf{X}\mathbf{w}$ je lineární kombinací sloupců matice \mathbf{X} s koeficienty w_0, \dots, w_p .
- Leží tedy v lineárním podprostoru prostoru \mathbb{R}^N , který je lineárním obalem $p + 1$ sloupců $\mathbf{X}_{\bullet 0}, \dots, \mathbf{X}_{\bullet p}$.
- Pro různé hodnoty \mathbf{w} pak vektor $\mathbf{X}\mathbf{w}$ celý tento prostor pokrývá, tj.

$$\langle \mathbf{X}_{\bullet 0}, \dots, \mathbf{X}_{\bullet p} \rangle = \{\mathbf{X}\mathbf{w} \mid \mathbf{w} \in \mathbb{R}^{p+1}\}.$$

Geometrická interpretace metody nejmenších čtverců (2/3)

- Chceme-li minimalizovat vzdálenost \mathbf{Y} a $\mathbf{X}\mathbf{w}$, hledáme bod $\mathbf{X}\mathbf{w}$ v podprostoru sloupců matice \mathbf{X} , který je k \mathbf{Y} nejbližší.
- Bod $\mathbf{X}\mathbf{w}$ je k bodu \mathbf{Y} nejbližší, jestliže je vektor $\mathbf{Y} - \mathbf{X}\mathbf{w}$ na ten podprostor kolmý (modrý vektor na obrázku).



Geometrická interpretace metody nejmenších čtverců (3/3)

- Bod $\mathbf{X}\mathbf{w}$ je k bodu \mathbf{Y} nejbližší, jestliže je vektor $\mathbf{Y} - \mathbf{X}\mathbf{w}$ na ten podprostor kolmý.
- To znamená, že je kolmý na všechny vektory $\mathbf{X}_{\bullet 0}, \dots, \mathbf{X}_{\bullet p}$, které ho generují:

$$(\mathbf{X}_{\bullet i})^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) = 0 \quad \text{pro všechny } i = 0, \dots, p.$$

- To lze maticově zapsat jako

$$\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) = \mathbf{0} \quad \text{a tedy} \quad \mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X}\mathbf{w} = \mathbf{0}.$$

- Získali jsme tím starou známou **normální rovnici** a tedy stejné řešení.
- Z výše uvedených geometrických úvah navíc plyne, že pro jakékoliv řešení \mathbf{w} normální rovnice je $\|\mathbf{Y} - \mathbf{X}\mathbf{w}\|$ a tedy i $\text{RSS}(\mathbf{w})$ nejmenší možné.
- Jakékoliv řešení normální rovnice tedy dává globální minimum.

Regularita versus lineární nezávislost sloupců (1/3)

- Normální rovnice má jednoznačné řešení, pokud je $\mathbf{X}^T \mathbf{X}$ regulární.
- Pojďme si odvodit, jak to souvisí s lineární nezávislostí sloupců matice \mathbf{X} .
- Je-li $\mathbf{X}_{\bullet i}$ i -tý sloupec matice \mathbf{X} , platí, že vektor

$$\mathbf{X}\mathbf{s} = s_0 \mathbf{X}_{\bullet 0} + s_1 \mathbf{X}_{\bullet 1} + \dots + s_p \mathbf{X}_{\bullet p}$$

je lineární kombinací sloupců matice \mathbf{X} s koeficienty danými složkami s .

- Matice \mathbf{X} má lineárně nezávislé sloupce, právě když je $\mathbf{X}\mathbf{s} = \mathbf{0}$ pouze pro $\mathbf{s} = \mathbf{0}$.
- Obecně pro matici \mathbf{X} a libovolný vektor $\mathbf{s} \in \mathbb{R}^{p+1}$ platí

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

- Z toho plyne, že je $\mathbf{X}^T \mathbf{X}$ je regulární, právě když jsou sloupce matice \mathbf{X} lineárně nezávislé.

Regularita versus lineární nezávislost sloupců (2/3)

- Problém zcela určitě nastává, pokud $N < p + 1$. Pak totiž v N rozměrném prostoru \mathbb{R}^N nemůže existovat $p + 1$ lineárně nezávislých vektorů a tak ani sloupce $\mathbf{X}_{\bullet 0}, \dots, \mathbf{X}_{\bullet p}$ matice \mathbf{X} nemohou být lineárně nezávislé.
- I v situaci $N \geq p + 1$ se ale může stát, že sloupce $\mathbf{X}_{\bullet 0}, \dots, \mathbf{X}_{\bullet p}$ nebudou lineárně nezávislé.
- Může to být například tím, že přímo jednotlivé příznaky jsou lineárně závislé a tedy jeden z nich je lineární kombinací ostatních.
- V takovém případě nepomůže ani libovolně vysoké N a sloupce matice \mathbf{X} budou lineárně závislé vždy.
- Podívejme se, co to znamená pro řešení úlohy minimalizace $\text{RSS}(\mathbf{w})$.

Regularita versus lineární nezávislost sloupců (3/3)

- Normální rovnice $\mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{0}$ je z pohledu složek vektoru \mathbf{w} soustava $p + 1$ lineárních rovnic o $p + 1$ neznámých.
- Tato soustava má vždy alespoň jedno řešení. Pokud jsou sloupce matice \mathbf{X} lineárně nezávislé, je $\mathbf{X}^T \mathbf{X}$ regulární, řešení je právě jedno a to

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

- V opačném případě existuje nekonečně mnoho řešení tak, že pro každé dvě řešení \mathbf{w} a \mathbf{w}' platí $\mathbf{X}(\mathbf{w} - \mathbf{w}') = \mathbf{0}$.
- Poznamenejme, že pro každou dvojici řešení platí

$$\begin{aligned} \text{RSS}(\mathbf{w}) &= \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{w} + \mathbf{X}\mathbf{w}' - \mathbf{X}\mathbf{w}'\|^2 \\ &= \|\mathbf{Y} - \mathbf{X}\mathbf{w}' - \mathbf{X}(\mathbf{w} - \mathbf{w}')\|^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{w}'\|^2 = \text{RSS}(\mathbf{w}'). \end{aligned}$$

- Všechny řešení tedy odpovídají stejné hodnotě RSS, která je, jak jsme již zmínili, globálním minimem, které je v tomto případě neostré.

Řešení při lineární závislosti sloupců

- Otázkou je, jak nějaké řešení získat, když nemůžeme invertovat matici $\mathbf{X}^T \mathbf{X}$ a následně použít vzoreček $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.
- Funkce `LinearRegression` z balíčku `scikit-learn` si s takovými případy poradí a řešení vrátí.¹
- Pokud $\mathbf{X}^T \mathbf{X}$ není regulární, vrátí takový vektor $\hat{\mathbf{w}}$, který řeší normální rovnici a zároveň má mezi všemi řešeními nejmenší normu $\|\hat{\mathbf{w}}\|$.
- Jen pro zajímavost uvedme, že toto řešení lze zapsat ve tvaru

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^+ \mathbf{X}^T \mathbf{Y},$$

kde $(\mathbf{X}^T \mathbf{X})^+$ je takzvaná Moorova-Penroseova pseudoinverzní matice k matici $\mathbf{X}^T \mathbf{X}$.

¹Jen pro zajímavost uvedme, že uvnitř se využívá funkce `dge1sd` z knihovny LAPACK.

Problém kolinearity

- Problémem nejsou pouze případy, kdy jsou sloupce matice \mathbf{X} lineárně závislé, ale úplně stačí, když jsou „skoro“ lineárně závislé.
- V obou těchto případech mluvíme o problému **kolinearity** (angl. **collinearity**).
- Myslíme tím tedy, že existují lineární kombinace sloupců, které dávají téměř nulové vektory, zatímco jiné lineární kombinace vrací mnohem větší vektory, tj.

$$\|\mathbf{X}\mathbf{u}\| \gg \|\mathbf{X}\mathbf{v}\| \doteq 0 \quad \text{pro nějaké} \quad \|\mathbf{u}\| = \|\mathbf{v}\| = 1.$$

- V takovém případě sice inverze $\mathbf{X}^T\mathbf{X}$ teoreticky existuje, ale prakticky je její výpočet numericky problematický.
- Především - a to je to hlavní jádro problému - je získaný odhad $\hat{\mathbf{w}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ velmi citlivý na malé nevhodné změny \mathbf{Y} .
- To znamená, že kdybychom náhodný výběr trénovací množiny zopakovali, může se hodnota odhadu $\hat{\mathbf{w}}_{\text{OLS}}$ radikálně změnit.
- Z pravděpodobnostního pohledu lze ukázat, že $\hat{\mathbf{w}}_{\text{OLS}}$ má potom v jistých směrech velký rozptyl.
- Toto se pochopitelně přenáší i na predikce $\hat{\mathbf{Y}}$, které pak mají v některých bodech velký rozptyl, což znamená, že jim **nemůžeme příliš důvěřovat**.

Regularizace lineární regrese

Pokud narazíme na problém kolinearity, máme v zásadě tři možnosti:

- Přigenerovat další data nebo odebrat existující a doufat, že se problém vyřeší.

To se ale nestane, pokud jsou samotné příznaky (skoro) lineárně závislé.

- Snížit počet příznaků. To znamená vyhození některých příznaků, případně nahrazení příznaků menším počtem nových, které již nebudou lineárně závislé.

Jednou z metod redukce počtu příznaků, kdy ty existující nahrazujeme menším počtem jejich lineárních kombinací, se budeme zabývat v 10. přednášce.

- Změnit funkci, kterou minimalizujeme, abychom měli jednoznačné a stabilní řešení.

Typicky provedeme tak zvanou regularizaci, kdy k RSS přidáme **regularizační člen**, který problémy kolinearity odstraní nebo alespoň dostatečně zmírní.

Hřebenová regrese (1/3)

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} bez 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 $\mathbb{E} \varepsilon = 0$ modelu a je tedy vhodné ho neomezovat.

Hřebenová regrese (2/3)

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

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

Gradient a Hessova matice jsou

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

Ekvivalent normální rovnice je tedy

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

Jak za chvíli ukážeme, matice $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}'$ je pro $\lambda > 0$ regulární, a řešení je tedy jednoznačně

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

Hřebenová regrese (3/3)

Podle předchozího slajdu je Hessova matice

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

Dále 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

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

a odpovídá globálnímu minimu RSS_λ .

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

Očekávaná chyba modelu

Jelikož $\mathbf{Y} = \mathbf{X}\mathbf{w} + \varepsilon$ z trénovací množiny je v důsledku náhodnosti ε náhodný vektor, dostáváme, že i $\hat{\mathbf{w}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}')^{-1}\mathbf{X}^T\mathbf{Y}$ je jakožto funkce \mathbf{Y} náhodný vektor.

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

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

Pro očekávanou chybu tedy platí

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

Rozklad očekávané chyby modelu

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

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

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**).

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

kde $\text{bias } \hat{Y} = E\hat{Y} - EY$ značí **vychýlení odhadu** (angl. **bias**).

Dohromady tedy máme finální dekompozici očekávané chyby jako

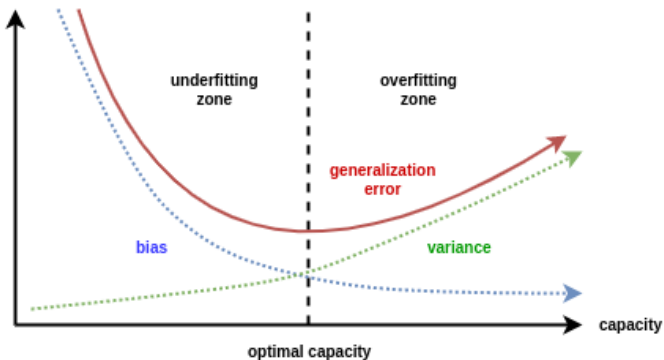
$$E L(Y, \hat{Y}) = \sigma^2 + (\text{bias } \hat{Y})^2 + \text{var } \hat{Y}.$$

Bias-variance tradeoff

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

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

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

Modely báзовých 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 **báзовé 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 báзовých 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 $\mathbf{Y} = \Phi \mathbf{w} + \boldsymbol{\varepsilon}$, kde

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

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

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

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

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

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

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

Bázové funkce

Mezi obvyklé volby báзовých 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 báзовé funkce** centrované v bodech trénovací množiny.

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