

# BI-ML1.21 přednáška 6

Daniel Vašata

FIT ČVUT

9. 11. 2023

Autoři: Karel Klouda, Daniel Vašata.  
Problémy, návrhy apod. hlaste v [GitLabu](#).  
Verze souboru: 9. listopadu 2023 15:29.

## Co bude v dnešní přednášce

- Připomenutí lineární regrese
- Geometrická interpretace metody nejmenších čtverců
- Problém lineárně závislých sloupců
- Metoda gradientního sestupu

## 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 přidáním umělého 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

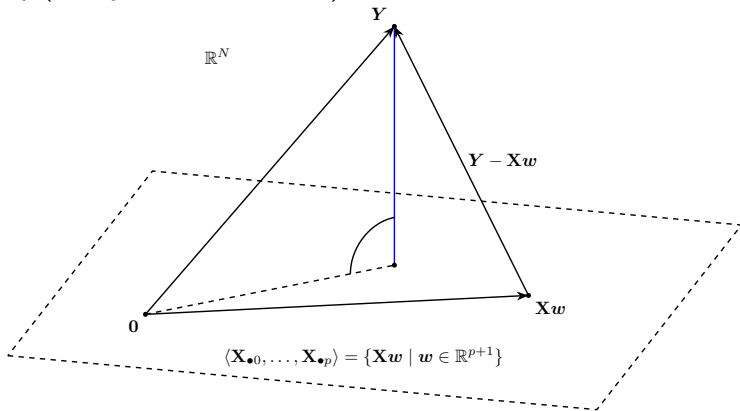
$$\mathbf{X}\mathbf{w} = w_0\mathbf{X}_{\bullet 0} + w_1\mathbf{X}_{\bullet 1} + \dots + w_p\mathbf{X}_{\bullet p}.$$

- 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  $Y$  a  $Xw$ , hledáme bod  $Xw$  v podprostoru sloupců matice  $X$ , který je k  $Y$  nejbližší.
- Bod  $Xw$  je k bodu  $Y$  nejbližší, jestliže je vektor  $Y - Xw$  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  $\mathbf{s}$ .

- 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  $\mathbf{X}^T \mathbf{X}$  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.

## Regularita versus lineární nezávislost sloupců (3/3)

- Podívejme se, co to znamená pro řešení úlohy minimalizace  $\text{RSS}(\mathbf{w})$ .
- 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

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

## Regularita versus lineární nezávislost sloupců (4/4)

- V opačném případě existuje nekonečně mnoho řešení tak, že pro každé dvě řešení  $\mathbf{w}$  a  $\mathbf{w}'$  platí  $\mathbf{X}^T \mathbf{X}(\mathbf{w} - \mathbf{w}') = \mathbf{0}$ .
- To, jak víme z předchozích úvah, implikuje  $\mathbf{X}(\mathbf{w} - \mathbf{w}') = \mathbf{0}$ .
- Pro každou dvojici řešení tedy 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 tudíž 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.<sup>1</sup>
- Pokud  $\mathbf{X}^T \mathbf{X}$  není regulární, vrátí (v rámci numerických možností) takový vektor  $\hat{\mathbf{w}}$ , který řeší normální rovnici a zároveň má mezi všemi řešeními nejmenší normu  $\|\hat{\mathbf{w}}\|$ .
- Jen pro zajímavost uveďme, 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  $\mathbf{X}^T \mathbf{X}$ .

---

<sup>1</sup>Jen pro zajímavost uveďme, 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, kdy 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  $\mathbf{Y}$ .

## Důsledky kolinearity

- 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  $\mathbf{Y}$  nebo  $\mathbf{X}$ .
- 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  $\hat{Y}$ , které pak mají v některých bodech velký rozptyl, což znamená, že jim **nemůžeme příliš důvěřovat**.

## Řešení problému kolinearity

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 příštím semestru.

- 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.

## Metoda gradientního sestupu

- K minimalizaci  $\text{RSS}(\mathbf{w})$  a potažmo k trénování lineární regrese můžeme použít metody gradientního sestupu (angl. **gradient descent**).
- V takovém případě začneme s nějakou počáteční hodnotou  $\mathbf{w}^{(0)}$  a pomocí rekurentního vztahu

$$\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} - \alpha \cdot \nabla \text{RSS}(\mathbf{w}^{(i)}) = \mathbf{w}^{(i)} + \alpha \cdot 2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{w}^{(i)})$$

postupně konstruujeme posloupnost vektorů o které doufáme, že konverguje ke skutečnému řešení  $\hat{\mathbf{w}}_{\text{OLS}}$ .

- Protože gradient ukazuje směrem nejvyššího růstu, díky zápornému znaménku děláme kroky ve směru nejvyššího poklesu.
- Koeficient  $\alpha$  je tzv. učicí parametr (angl. **learning rate**) a může záviset na  $i$  (v tzv. adaptivní verzi).
- V našem případě platí, že pro vhodné  $\alpha$  tato metoda konverguje do globálního optima daného  $\hat{\mathbf{w}}_{\text{OLS}}$ . Konvergence může být ale velmi pomalá a pro nevhodné  $\alpha$  ani konvergovat nemusí.