BI-ML1.21 přednáška 6 handout 9. listopadu 2023
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 9. listopadu 2023 15:31.

Co bude v dnešní přednášce

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

23 Geometrická interpretace metody nejmenších čtverců

Lineární model

- Model pro vysvětlovanou proměnnou Y v bodě $\boldsymbol{x}$ je
$$Y = \boldsymbol{w}^T\boldsymbol{x} + \varepsilon = \boldsymbol{x}^T\boldsymbol{w} + \varepsilon = w_0 + w_1x_1 + \dots w_px_p + \varepsilon.$$
- Model pro trénovací množinu tvořenou N páry $(Y_i, \boldsymbol{x}_i)$ je $Y_i = \boldsymbol{x}_i^T\boldsymbol{w} + \varepsilon_i$.
- Dohromady při značení $\boldsymbol{x}_i = (1, x_{i;1}, \dots, x_{i;p})^T$ 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 $\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$, kde
$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1;1} & \cdots & x_{1;p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N;1} & \cdots & x_{N;p} \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$
- 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
$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2 = \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{w}\|^2.$$
- Minimum je určeno řešením *normální rovnice*
$$\mathbf{X}^T\boldsymbol{Y} - \mathbf{X}^T\mathbf{X}\boldsymbol{w} = \mathbf{0},$$
která odpovídá podmínce $\nabla\,\mathrm{RSS}(\boldsymbol{w}) = \mathbf{0}$.
- Za předpokladu, že je matice $\mathbf{X}^T\mathbf{X}$ regulární, existuje jediné řešení minimalizující $\mathrm{RSS}(\boldsymbol{w})$,
$$\hat{\boldsymbol{w}}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}.$$
- Predikce v bodě $\boldsymbol{x}$ je potom $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{w}}_{\mathrm{OLS}}$.

Geometrická interpretace metody nejmenších čtverců (1/3)

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

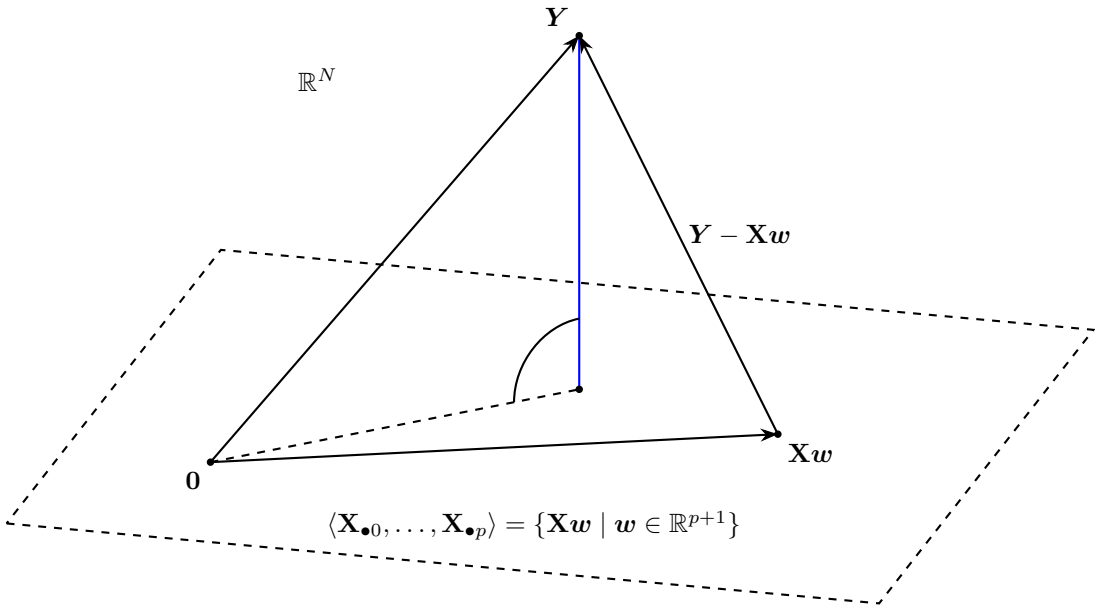
$$\mathbf{X}\boldsymbol{w} = w_0\mathbf{X}_{\bullet 0} + w_1\mathbf{X}_{\bullet 1} + \ldots + w_p\mathbf{X}_{\bullet p}.$$

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

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

Geometrická interpretace metody nejmenších čtverců (2/3)

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



Geometrická interpretace metody nejmenších čtverců (3/3)

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

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

- To lze maticově zapsat jako

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

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

24 Regularita versus lineární nezávislost

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

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

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

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

- 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 $\mathbf{X}_{\bullet 0}, \ldots, \mathbf{X}_{\bullet p}$ matice $\mathbf{X}$ nemohou být lineárně nezávislé.
- I v situaci $N \geq p + 1$ se ale může stát, že sloupce $\mathbf{X}_{\bullet 0}, \ldots, \mathbf{X}_{\bullet p}$ 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 $\mathrm{RSS}(\boldsymbol{w})$.
- Normální rovnice $\mathbf{X}^T\boldsymbol{Y} - \mathbf{X}^T\mathbf{X}\boldsymbol{w} = \mathbf{0}$ je z pohledu složek vektoru $\boldsymbol{w}$ 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

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

Regularita versus lineární nezávislost sloupců (4/4)

- V opačném případě existuje nekonečně mnoho řešení tak, že pro každé dvě řešení $\boldsymbol{w}$ a $\boldsymbol{w}'$ platí $\mathbf{X}^T\mathbf{X}(\boldsymbol{w}-\boldsymbol{w}') = \mathbf{0}$.
- To, jak víme z předchozích úvah, implikuje $\mathbf{X}(\boldsymbol{w} - \boldsymbol{w}') = \mathbf{0}$.
- Pro každou dvojici řešení tedy platí

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

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

25 Problém kolinearity

Řešení při lineární závislých sloupcích

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

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

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

Problém kolinearity

- Problémem nejsou pouze případy, kdy jsou sloupce matice $\boldsymbol{X}$ 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.

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

- 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 $\hat{\boldsymbol{w}}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ velmi citlivý na malé nevhodné změny $\boldsymbol{Y}$.

[1] Jen pro zajímavost uveďme, že uvnitř se využívá funkce `dgelsd` z knihovny LAPACK.

Důsledky kolinearity

- Především - a to je to hlavní jádro problému - je získaný odhad $\hat{\boldsymbol{w}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ velmi citlivý na malé nevhodné změny $\boldsymbol{Y}$ nebo $\mathbf{X}$.
- To znamená, že kdybychom náhodný výběr trénovací množiny zopakovali, může se hodnota odhadu $\hat{\boldsymbol{w}}_{\text{OLS}}$ radikálně změnit.
- Z pravděpodobnostního pohledu lze ukázat, že $\hat{\boldsymbol{w}}_{\text{OLS}}$ 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.

26 Metoda gradientního sestupu

Metoda gradientního sestupu

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

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

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

ChangeLog

Verze	Datum	Autor	Log
1.0	20.10.2022	DV	Výchozí verze pro rok 2021/2022.
1.0	9.11.2023	DV	Výchozí verze pro rok 2023/2024.