BI-ML1.21 přednáška 5

Daniel Vašata

FIT ČVUT

2. 11. 2023

Autoři: Karel Klouda, Daniel Vašata.
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 2. listopadu 2023 06:31.

Co bude v dnešní přednášce

- Představení modelu lineární regrese
- Úvod do problematiky hledání extrémů funkce více proměnných
- Odhad parametrů lineárního modelu pomocí metody nejmenších čtverců
- Diskuse výsledků

Motivační příklad

- Chceme prodat nemovitost (řekněme v Praze) a netušíme, za kolik ji máme potenciálním zájemcům nabídnout.
- Zároveň si nechceme platit žádného „realitního“ odborníka, který by nám se stanovením ceny poradil.
- Zkusíme tedy udělat vlastní průzkum realitního trhu a vytvořit model, který nám pomůže cenu stanovit.
- Napíšeme skript, který stáhne data z realitních serverů a uloží je v nějaké strukturované podobě (ideálně tabulce či více tabulkách).
- Pro jednoduchost uvažujme, že budeme znát u každé nabídky toto:
  - Y – cenu, za kterou se prodává,
  - X_1 – užitnou plochu,
  - X_2 – počet místností,
  - X_3 – vzdálenost od nejbližší zastávky metra.
- Jako vysvětlovanou proměnnou Y jsme označili veličinu, kterou pro naši nemovitost neznáme a chceme ji predikovat.
- Příznaky $X_1, \ldots X_3$ označují veličiny, které pro naši nemovitost známe a o kterých věříme, že cenu Y ovlivňují.

Formalizace úlohy

Obecně tedy chceme na základě p příznaků $X_1, \ldots, X_p$ predikovat hodnotu vysvětlované proměnné Y.

V modelu lineární regrese předpokládáme lineární závislost vysvětlované proměnné na hodnotách příznaků.

Jelikož nedoufáme, že tato závislost je perfektní v tom smyslu, že pro stejné hodnoty $x_1, \ldots, x_p$ příznaků $X_1, \ldots, X_p$ dostaneme vždy stejnou hodnotu vysvětlované proměnné Y, modelujeme tuto závislost následovně:

$$Y = w_1 x_1 + \ldots + w_p x_p + \varepsilon,$$

kde $w_1, \ldots, w_p$ jsou nějaké neznámé koeficienty a ε je náhodná veličina.

Poznámky:

- Veličina ε odpovídá části Y, která je nevysvětlitelná pomocí hodnot příznaků a je tedy z našeho pohledu náhodná.
- Do náhodné veličiny ε se tak „schovají“ vlivy, které neznáme nebo cíleně nezahrnujeme do našeho modelu (např. stáří budovy, počet koupelen, počet oken) ale např. i chyby, nekonzistence dat a jiné podivnosti v měření příznaků.

Model lineární regrese

Obvykle ještě oddělujeme střední hodnotu náhodných vlivů a dostáváme tak:

Model lineární regrese

Hodnota vysvětlované proměnné Y v bodě $(x_1, \ldots, x_p)^T$ je

$$Y = w_0 + w_1 x_1 + \ldots + w_p x_p + \varepsilon,$$

kde $\mathrm{E}\,\varepsilon = 0$.

- Koeficient w_0 se nazývá **intercept** a odpovídá (očekávané) výchozí hodnotě Y při nulových příznacích.
- Zavedeme-li nový konstantní příznak $X_0 = x_0 = 1$ a vektorové značení

  $$\boldsymbol{x} = (x_0, x_1, \ldots, x_p)^T \quad \text{a} \quad \boldsymbol{w} = (w_0, w_1, \ldots, w_p)^T,$$

  můžeme zkráceně psát

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

- Vektor $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)^T$ koeficientů také někdy nazýváme vektor vah.

Predikce v modelu lineární regrese

Předpokládejme nyní, že už máme odhad $\hat{\boldsymbol{w}}$ vektoru koeficientů $\boldsymbol{w}$.

Hodnotu Y v konkrétním bodě $\boldsymbol{x}$ predikujeme vztahem

$$\hat{Y} = \hat{\boldsymbol{w}}^T \boldsymbol{x} = \hat{w}_0 + \hat{w}_1 x_1 + \dots \hat{w}_p x_p.$$

Skutečná hodnota Y v bodě $\boldsymbol{x}$ je přitom určena vztahem

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

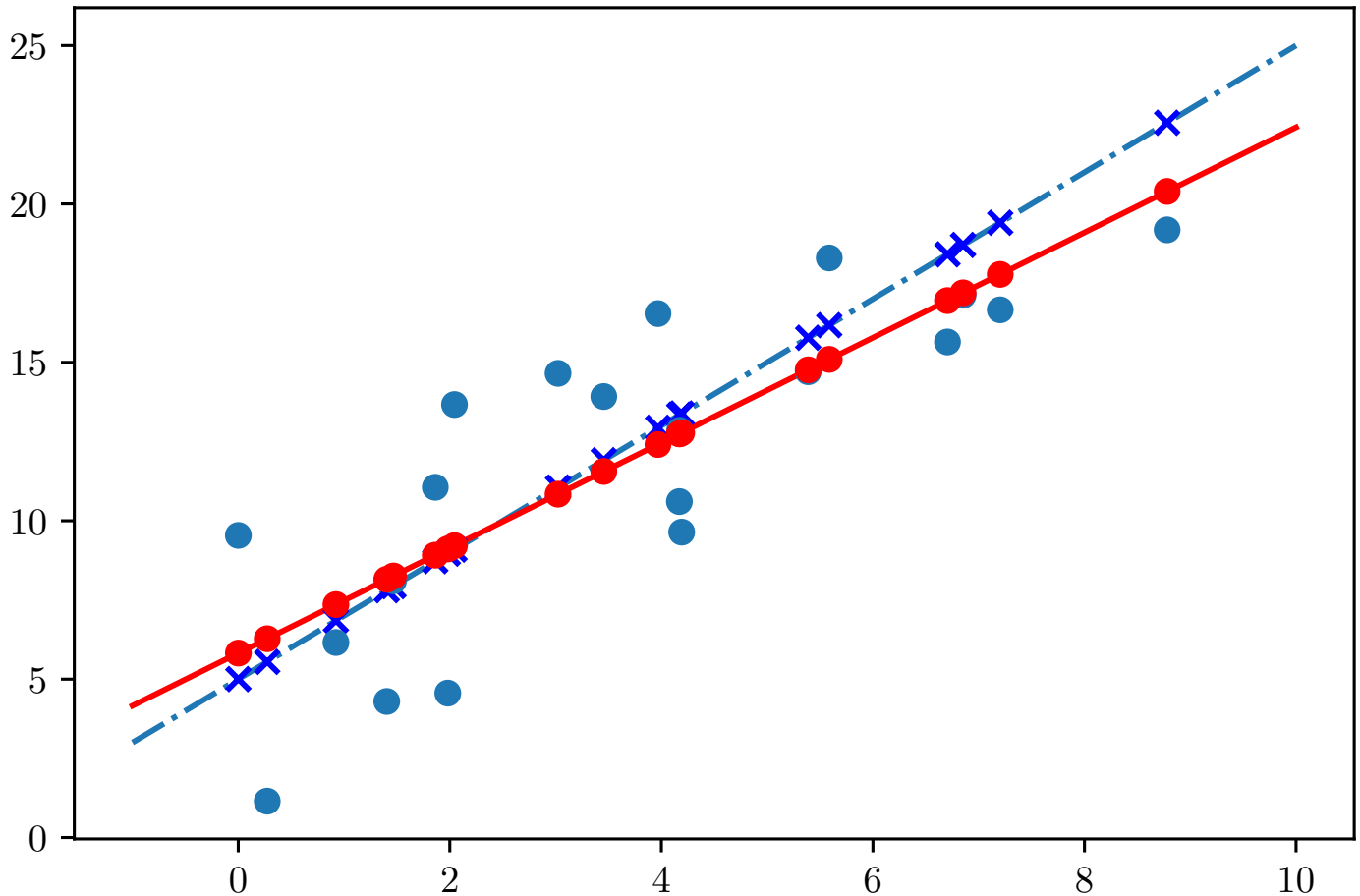
a je tedy náhodnou veličinou.

Z předpokladu $\mathrm{E}\,\varepsilon = 0$ plyne, že

$$\mathrm{E}\,Y = \boldsymbol{w}^T \boldsymbol{x}$$

a $\hat{Y}$ je tedy vlastně bodovým odhadem střední hodnoty $\mathrm{E}\,Y$ v bodě $\boldsymbol{x}$.

Vizualizace modelu lineární regrese



Modré body jsou body trénovací množiny. Červené body jsou predikce. Modré křížky odpovídají středním hodnotám bodů trénovací množiny, $(\boldsymbol{x}_i, \mathrm{E}\, Y_i)$. Modrá čerchovaná čára je skutečná regresní přímka daná rovnicí $y = \boldsymbol{w}^T \boldsymbol{x}$ a červená čára je přímka $\hat{y} = \hat{\boldsymbol{w}}^T \boldsymbol{x}$ určující naše predikce.

Měření chybovosti predikce pomocí ztrátové funkce

Zaměřme se nyní na problematiku odhadu vektoru parametrů modelu $\boldsymbol{w}$.

Následující úvahy jsou obecně platné pro supervizované učení nějakého modelu s parametry.

- Naším cílem je najít takovou hodnotu $\boldsymbol{w}$, aby chyba modelu byla co nejmenší.
- Tuto hodnotu pak použijeme jako odhad $\hat{\boldsymbol{w}}$.
- K tomu musíme specifikovat, co se myslí chybou modelu a v jakém smyslu má být nejmenší.
- Chybu modelu nejčastěji měříme pomocí nějaké nezáporné funkce $L : \mathbb{R}^2 \to \mathbb{R}$, nazývané **ztrátová funkce** (angl. **loss function**), kterou aplikujeme na skutečnou hodnotou proměnné Y a odpovídající predikci $\hat{Y}$.
- Obvyklou volbou v případě spojité vysvětlované veličiny bývá **kvadratická ztrátová funkce**,

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

Metoda nejmenších čtverců

- Velikost chyby modelu v bodě $\boldsymbol{x}$ je tedy $L(Y, \hat{Y})$, kde Y je skutečná hodnota vysvětlované Y v bodě $\boldsymbol{x}$ a $\hat{Y} = \boldsymbol{w}^T\boldsymbol{x}$ je predikce v bodě $\boldsymbol{x}$.
- Otázkou je, v jakém bodě $\boldsymbol{x}$ bychom měli hodnotu $L(Y, \hat{Y})$ vyhodnocovat a následně minimalizovat vzhledem k $\boldsymbol{w}$.
- Zřejmě to musí být bod $\boldsymbol{x}$ z trénovací množiny, protože jinak bychom neznali skutečnou hodnotu Y v tomto bodě.
- Abychom co nejvíc využili trénovací data, budeme minimalizovat součet chyb přes všechny body trénovací množiny, tj. přes všechny dvojice $(\boldsymbol{x}_i, Y_i)$ pro $i = 1, \ldots, N$.
- Součet chyb přes všechny tyto body pro kvadratickou ztrátovou funkci je

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} L(Y_i, \boldsymbol{w}^T\boldsymbol{x}_i) = \sum_{i=1}^{N} (Y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$$

  a nazýváme ho **reziduální součet čtverců** (angl. **residual sum of squares**).
- Minimalizací tohoto výrazu získáme odhad $\hat{\boldsymbol{w}}$. Tento postup se nazývá **metoda nejmenších čtverců** (angl. the method of **least squares**).

Parciální derivace

Protože $\boldsymbol{w}$ je vektor, minimalizace $\mathrm{RSS}(\boldsymbol{w})$ spadá do problematiky minimalizace funkce více proměnných.

Jak si nyní ukážeme, postupuje se analogicky jako v případě funkce jedné proměnné.

Definice

Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných. **Parciální derivaci** funkce $f(x_1, \ldots, x_d)$ podle proměnné x_i v bodě $\boldsymbol{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$ definujeme jako derivaci funkce $g(x_i) = f(a_1, \ldots, a_{i-1}, x_i, a_{i+1} \ldots, a_d)$ v bodě a_i a značíme

$$\partial_{x_i} f(\boldsymbol{a}) \quad \text{nebo} \quad \frac{\partial f}{\partial x_i}(\boldsymbol{a}).$$

Označením $\partial_{x_i} f$ nebo $\frac{\partial f}{\partial x_i}$ pak myslíme funkci, která každému bodu, kde je konečná, přiřadí hodnotu parciální derivace podle x_i v tomto bodě.

Parciální derivace je stejná jako ta „obyčejná“, akorát ostatní proměnné bereme jako konstanty.

Platí např. $\partial_x (x + y) = 1$ nebo $\partial_y \sin(y + xy) = (1 + x)\cos(y + xy)$.

Gradient funkce

Definice

Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných, která má v bodě $\boldsymbol{a} \in \mathbb{R}^d$ konečné všechny parciální derivace. **Gradient** funkce f v bodě $\boldsymbol{a}$ definujeme jako vektor

$$\nabla f(\boldsymbol{a}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \ldots, \frac{\partial f}{\partial x_d}(\boldsymbol{a}) \right).$$

Označením ∇f pak myslíme gradient funkce jakožto zobrazení, které každému bodu, kde to lze, přiřadí gradient v tomto bodě.

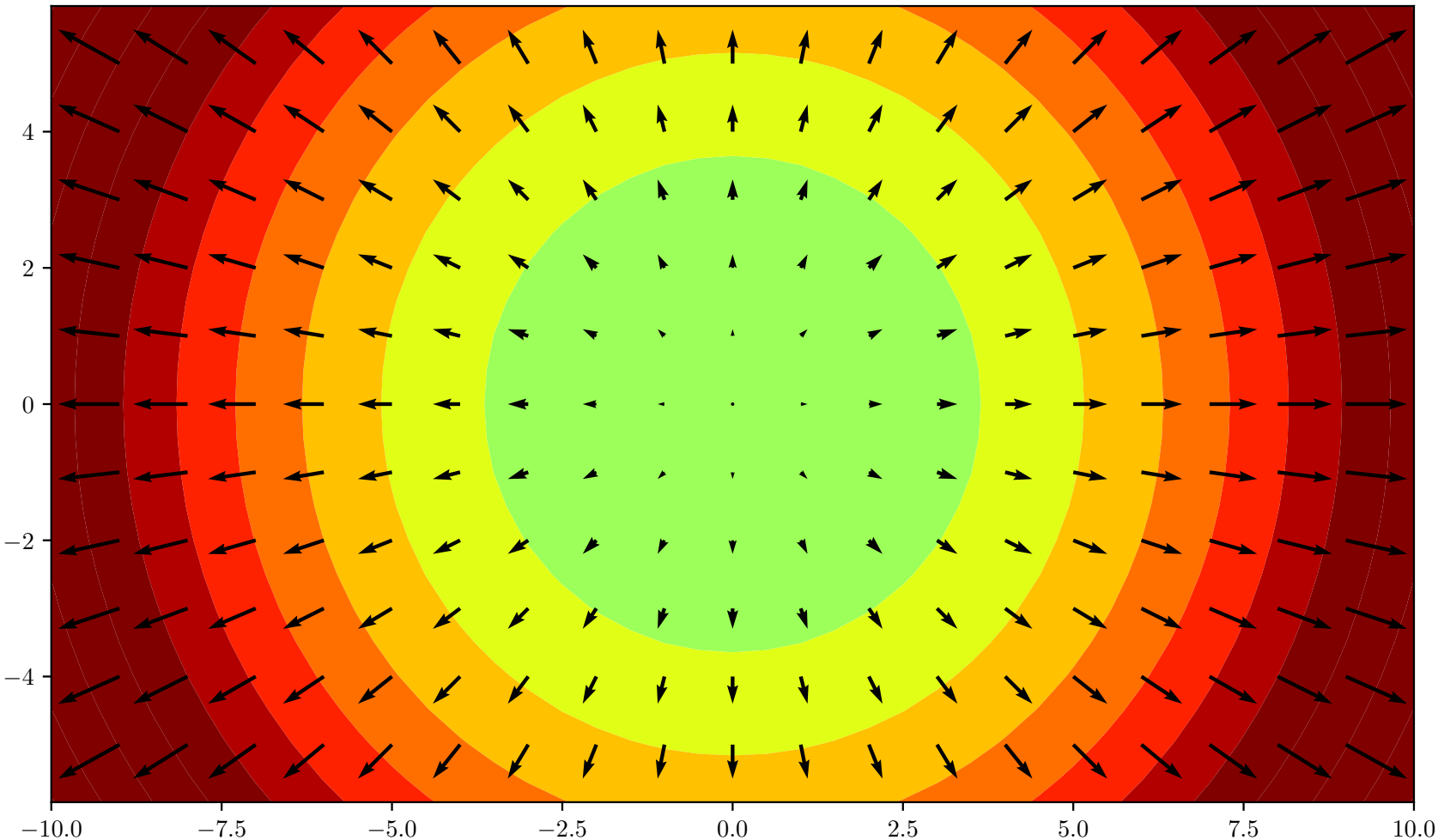
Gradient v bodě je tedy vektor z $\mathbb{R}^d$, jehož složky jsou jednotlivé parciální derivace.

Uvažujme funkci $f(x, y) = x^2 + y^2$, která odpovídá parabolické jámě s minimem v počátku $\mathbf{0}$. Platí $\nabla f = (2x, 2y)$.

Nejdůležitější vlastností gradientu[1] je, že ukazuje směr maximálního růstu funkce v daném bodě. Z toho mimo jiné plyne, že je gradient vždy kolmý na vrstevnici procházející daným bodem.

[1] Za dodatečného předpokladu spojitostí všech jeho složek na okolí daného bodu.

Vizualizace gradientu



Hessova matice

Analogicky k jednorozměrnému případu platí, že má-li funkce v nějakém bodě lokální extrém a gradient zde existuje, musí být nulový.

K získání postačující podmínky nám ještě zbývá sestrojit analog druhé derivace.

Definice

Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných. **Hessovu matici** funkce f v bodě $\boldsymbol{a} \in \mathbb{R}^d$ definujeme jako

$$\mathbf{H}_f(\boldsymbol{a}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\boldsymbol{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(\boldsymbol{a}) \end{pmatrix},$$

kde $\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{a}) = \left(\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)\right)(\boldsymbol{a})$ značí druhou parciální derivaci podle x_j a x_i.

Je to tedy matice druhých parciálních derivací podle všech proměnných.

Postačující podmínka pro existenci lokálního extrému

Věta

Buď $f : \mathbb{R}^d \to \mathbb{R}$ funkce d proměnných a bod $\boldsymbol{x}^* \in \mathbb{R}^d$ takový, že $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ a f má spojité všechny druhé parciální derivace na nějakém okolí bodu $\boldsymbol{x}^*$.

- *Jestliže*
$$\boldsymbol{s}^T \mathbf{H}_f(\boldsymbol{x}^*)\boldsymbol{s} > 0, \quad \text{pro každé} \quad \boldsymbol{s} \in \mathbb{R}^d, \boldsymbol{s} \neq \mathbf{0},$$
nabývá funkce f v bodě $\boldsymbol{x}^*$ ostrého lokálního minima.
- *Jestliže pro každé $\boldsymbol{x}$ z nějakého okolí bodu $\boldsymbol{x}^*$*
$$\boldsymbol{s}^T \mathbf{H}_f(\boldsymbol{x})\boldsymbol{s} \geq 0, \quad \text{pro každé} \quad \boldsymbol{s} \in \mathbb{R}^d,$$
nabývá funkce f v bodě $\boldsymbol{x}^*$ neostrého lokálního minima.

Tyto vlastnosti se nazývají **pozitivní definitnost** resp. **pozitivní semi-definitnost** Hessovy matice $\mathbf{H}_f$ v bodě $\boldsymbol{x}^*$ resp. $\boldsymbol{x}$.

Pro funkci $f(x, y) = x^2 + y^2$ je řešením $\nabla f = (2x, 2y) = \mathbf{0}$ bod $\boldsymbol{x}^* = \mathbf{0}$ a Hessova matice je
$$\mathbf{H}_f(\boldsymbol{x}^*) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2\mathbf{I}.$$
Protože $\boldsymbol{s}^T 2\mathbf{I}\boldsymbol{s} = 2\boldsymbol{s}^T\boldsymbol{s} = 2\|\boldsymbol{s}\|^2 > 0$ pro každé $\boldsymbol{s} \neq \mathbf{0}$, nastává v bodě $\boldsymbol{x}^* = \mathbf{0}$ ostré lokální minimum. To pro náš paraboloid skutečně platí.

Přepis modelu trénovacích dat

Než aplikujeme předchozí metodu na minimalizaci RSS, přepišme si celkový model pro naše trénovací data do maticového tvaru.

Trénovací data považujeme za náhodný výběr z uvažovaného lineárního modelu provedený v různých bodech $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

Máme tedy N párů $(Y_i, \boldsymbol{x}_i)$, kde $Y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \varepsilon_i$.

Zaveďme náhodné vektory $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$ a body $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ zapišme v řádcích do matice $\mathbf{X} \in \mathbb{R}^{N,p+1}$,

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1;1} & x_{1;2} & \cdots & x_{1;p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N;1} & x_{N;2} & \cdots & x_{N;p} \end{pmatrix}.$$

Při tomto značení můžeme celkový model trénovací množiny zapsat jako

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon},$$

kde $\mathrm{E}\,\boldsymbol{\varepsilon} = (\mathrm{E}\,\varepsilon_1, \ldots, \mathrm{E}\,\varepsilon_N)^T = \mathbf{0}$.

Minimalizace RSS (1/3)

RSS můžeme vyjádřit jako

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

Aplikujme předchozí teorii na minimalizaci $\mathrm{RSS}(\boldsymbol{w})$. Začneme parciálními derivacemi podle $w_0, \ldots, w_p$,

$$\frac{\partial \mathrm{RSS}}{\partial w_j} = \sum_{i=1}^{N} 2(Y_i - \boldsymbol{w}^T \boldsymbol{x}_i)(-x_{i;j}).$$

Pro gradient tedy dostáváme

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

Položíme-li $\nabla \mathrm{RSS} = \mathbf{0}$ získáme tzv. **normální rovnici** (angl. **normal equation**),

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

Minimalizace RSS (2/3)

Při výpočtu Hessovy matice použijeme

$$\frac{\partial^2 \mathrm{RSS}}{\partial w_k \partial w_j} = \sum_{i=1}^{N} 2(-x_{i;k})(-x_{i;j}).$$

Hessova matice je tedy

$$\mathbf{H}_{\mathrm{RSS}}(\boldsymbol{w}) = 2\mathbf{X}^T\mathbf{X}$$

bez ohledu na konkrétní hodnotu $\boldsymbol{w}$.

Dále pro každé $\boldsymbol{s} \in \mathbb{R}^{p+1}$ platí

$$\boldsymbol{s}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{s} = (\mathbf{X}\boldsymbol{s})^T(\mathbf{X}\boldsymbol{s}) = \|\mathbf{X}\boldsymbol{s}\|^2 \geq 0.$$

Hessova matice $\mathbf{H}_{\mathrm{RSS}}(\boldsymbol{w})$ je tedy vždy pozitivně semi-definitní.

Podle předchozí věty proto nastává neostré lokální minimum v jakémkoliv bodě $\boldsymbol{w}$, který řeší normální rovnici

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

Minimalizace RSS (3/3)

Předpokládejme nyní, že $\mathbf{X}^T\mathbf{X}$ je **regulární matice**.

Normální rovnice

$$\mathbf{X}^T\boldsymbol{Y} - \mathbf{X}^T\mathbf{X}\boldsymbol{w} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}^T\mathbf{X}\boldsymbol{w} = \mathbf{X}^T\boldsymbol{Y}$$

má potom jednoznačné řešení

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

kde značení vychází z anglického názvu **ordinary least squares** solution.

Pro matici $\mathbf{X}$ a libovolný vektor $\boldsymbol{s} \in \mathbb{R}^{p+1}$ snadno vidíme řetěz implikací

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

Z regularity $\mathbf{X}^T\mathbf{X}$ tedy plyne, že pro nenulové $\boldsymbol{s}$ nikdy nemůže platit $\boldsymbol{s}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{s} = 0$, a tedy Hessova matice je pozitivně definitní.

To znamená, že $\hat{\boldsymbol{w}}_{\mathsf{OLS}}$ je bodem ostrého lokálního minima.

Jak uvidíme později, v tomto bodě RSS nabývá dokonce globálního minima.

Shrnutí metody nejmenších čtverců

- Model pro vysvětlovanou proměnnou Y v bodě $\boldsymbol{x}$ je $Y = \boldsymbol{w}^T\boldsymbol{x} + \varepsilon$.
- Model pro trénovací množinu je $\boldsymbol{Y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$.
- Při trénování minimalizujeme residuální součet čtverců

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

- 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 $\hat{Y}$ v bodě $\boldsymbol{x}$ je potom

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

Různé poznámky

- Lineární regrese je krásným příkladem diskriminativní metody, kdy přímo odhadujeme $\mathrm{P}(Y|\boldsymbol{X} = \boldsymbol{x})$. Resp. přímo $\mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.
- Lineární regrese je ve vztahu ke problémům dimenzionality poměrně rezistentní.
- Důvodem je, že se jedná o parametrickou metodu. V ideálním případě nám tedy pro p příznaků $+1$ intercept může k určení přesného modelu stačit přesně $p + 1$ bodů trénovací množiny.
- Problémy nastávají, když jsou v důsledku malé trénovací množiny nebo špatných příznaků, které jsou např. silně korelované, sloupce matice $\mathbf{X}$ (skoro) lineárně závislé.
- Potom již nelze snadno provést inverzi matice $\mathbf{X}^T\mathbf{X}$ a případně je numericky nestabilní.
- V důsledku se tento problém typicky projeví přeučením modelu, které znamená, že se model příliš přizpůsobí trénovací množině a nebude schopen dobře predikovat nové body.