BI-ML2.21 přednáška 1

Daniel Vašata

FIT ČVUT

21. 2. 2024

Autor: Daniel Vašata.
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 21. února 2024 12:43.

Podmínky získání zápočtu

- **Zápočet bude postaven na vypracovávání domácích úkolů.**
- Úkoly budou během semestru **dva**, každý za **max. 25 bodů**.
- Domácí úkoly budete vypracovávat v jazyce Python ve formátu Jupyter notebook (`.ipynb`).
- Zadání a podrobné instrukce k vypracování a odevzdání najdete na stránkách předmětu:

  ☞ **courses.fit.cvut.cz/BI-ML2/** ☜

- **K získání zápočtu je třeba získat alespoň 25 bodů z 50.**

Podmínky složení zkoušky

- Zkouška bude **ústní**.
- Každý student dostane **dvě otázky z předem zveřejněného seznamu**.
- Z každé otázky můžete získat **až 25 bodů**.
- Celkem tedy můžete ze zkoušky získat **až 50 bodů**. Není žádný minimální nutný počet bodů, který je třeba ze zkoušky získat.
- Pokud student/ka u zkoušky prokáže zásadní neznalost, může zkoušející použít právo veta a zkoušku ukončit jako neúspěšnou.
- V případě úspěchu u zkoušky se výsledná známka odvodí ze součtu bodů ze semestru a ze zkoušky.

O čem to všechno vlastně bude?

Budeme rozšiřovat znalosti získané v kurzu **BI-ML1**.

Zejména se budeme zabývat:

- Dalšími důležitými metodami supervizovaného učení
- Metodami redukce dimenzionality
- Neuronovými sítěmi
- Posilovaným učením

Doporučená literatura:

- Hastie T., Tibshirani R., Friedman J.: **The Elements of Statistical Learning.** Springer, 2009.
- Murphy K. P.: **Machine Learning, A Probabilistic Perspective.** MIT Press, 2012.
- Goodfellow I., Bengio Y., Courville A.: **Deep Learning.** MIT Press, 2016.
- Sutton R. S., Barto A. G.: **Reinforcement Learning.** MIT Press, 2018.

Co bude v dnešní přednášce

- Opakování modelu lineární a hřebenové regrese
- Opakování lineárního modelu bázových funkcí
- Duální reprezentace optimalizační úlohy pro trénování
- Diskuse jádrového triku a ukázky jádrových funkcí

Opakování lineární regrese

Začneme připomenutím lineární regrese.

Model lineární regrese

Hodnota vysvětlované proměnné Y v bodě o hodnotách $x_1, \dots, x_p$ příznaků $X_1, \dots, X_p$ je

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

kde $w_0, \dots, w_p$ jsou neznámé parametry a ε je náhodná veličina pro kterou platí $\mathrm{E}\,\varepsilon = 0$.

Poznámky:

- Zavedeme-li nový konstantní příznak $X_0 = x_0 = 1$ a vektorové značení

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

  můžeme zkráceně psát

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

- Naším cílem je odhadnout neznámé váhy $\boldsymbol{w}$ pomocí $\hat{\boldsymbol{w}}$ a pak predikovat Y jako

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

- Ze statistického pohledu je predikce $\hat{Y}$ bodovým odhadem $\mathrm{E}\,Y = \boldsymbol{w}^T \boldsymbol{x}$.

Model pro trénovací množinu

Trénovací množina je tvořena N dvojicemi $(Y_1, \boldsymbol{x}_1), \ldots, (Y_N, \boldsymbol{x}_N)$, které jsou nezávislé a pocházejí ze stejného rozdělení, tj.

$$Y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \varepsilon_i,$$

kde $\varepsilon_1, \ldots, \varepsilon_N$ jsou nezávislé a $\mathrm{E}\,\varepsilon_i = 0$.

Toto zapisujeme maticově jako

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

kde

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

s $\mathrm{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$.

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 $\mathrm{RSS}(\boldsymbol{w})$,

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

- Predikce v bodě $\boldsymbol{x}$ je potom $\hat{Y} = \boldsymbol{x}^T \hat{\boldsymbol{w}}_{\mathsf{OLS}}$.
- Když $\mathbf{X}^T\mathbf{X}$ není regulární můžeme použít Mooreovu-Penroseovu pseudoinverzní matici $(\mathbf{X}^T\mathbf{X})^+$ místo $(\mathbf{X}^T\mathbf{X})^{-1}$ nebo přidat regularizační člen k $\mathrm{RSS}(\boldsymbol{w})$ což vede na hřebenovou regresi.

Hřebenová regrese

- Ve **hřebenové regresi** minimalizujeme **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 hyperparametru $\lambda \geq 0$.

- 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

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

- Normální rovnice je potom

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

- Protože $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}'$ je regulární pro každé $\lambda > 0$, existuje jednoznačné řešení

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

- Hřebenová regrese pro $\lambda = 0$ dává obyčejnou metodu nejmenších čtverců, $\hat{\boldsymbol{w}}_0 = \hat{\boldsymbol{w}}$.

Lineární model bázových funkcí

Doposud jsme byli schopni modelovat pouze lineární funkci ve vstupních proměnných.

Základní rozšíření spočívá v nahrazení původních příznaků jejich transformovanými variantami.

Označme jako $\mathcal{X}$ prostor všech možných hodnot vektoru příznaků $\boldsymbol{X} = (X_1, \dots, X_p)^T$. Typicky, $\mathcal{X} = \mathbb{R}^p$ nebo alespoň $\mathcal{X} \subset \mathbb{R}^p$.

Pro $M \in \mathbb{N}$ uvažujme M lineárně nezávislých funkcí $\varphi_1, \dots, \varphi_M$ z $\mathcal{X}$ do $\mathbb{R}$. Tyto funkce nazývané **bázové funkce** (angl. **basis functions**) představují transformace původních příznaků $X_1, \dots, X_p$ do nového M-rozměrného příznakového prostoru.

Jako model pro Y nyní použijme lineární model v tomto novém příznakovém prostoru.

Lineární model bázových funkcí

Vysvětlovaná proměnná Y v bodě $\boldsymbol{x} = (x_1, \dots, x_p)^T$ hodnot vektoru příznaků $\boldsymbol{X}$ je určena vztahem

$$Y = w_1\varphi_1(\boldsymbol{x}) + \dots w_M\varphi_M(\boldsymbol{x}) + \varepsilon = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}) + \varepsilon,$$

kde $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^M$ je vektorová funkce definována jako $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \dots, \varphi_M(\boldsymbol{x}))^T$ pro všechna $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{w} = (w_1, \dots, w_M)^T$ je vektor neznámých parametrů, a ε je náhodná veličina s $\mathrm{E}\,\varepsilon = 0$.

Pro jednoduchost při pozdějších úvahách nyní neuvažujeme intercept (který můžeme vytvořit vhodnou volbou jedné z bázových funkcí).

Odhad v lineárním modelu bázových funkcí

- Model pro trénovací množinu tvořenou N dvojicemi $(Y_1, \boldsymbol{x}_1), \ldots, (Y_N, \boldsymbol{x}_N)$ můžeme ve vektorovém tvaru zapsat jako

$$\boldsymbol{Y} = \boldsymbol{\Phi} \boldsymbol{w} + \boldsymbol{\varepsilon},$$

kde

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\varphi}(\boldsymbol{x}_1)^T \\ \vdots \\ \boldsymbol{\varphi}(\boldsymbol{x}_N)^T \end{pmatrix} = \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \cdots & \varphi_M(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\boldsymbol{x}_N) & \cdots & \varphi_M(\boldsymbol{x}_N) \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

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

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = \|\boldsymbol{Y} - \boldsymbol{\Phi} \boldsymbol{w}\|^2 + \lambda \boldsymbol{w}^T \boldsymbol{w}.$$

- Normální rovnice je

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

- Pro $\lambda > 0$ exituje jednoznačné řešení

$$\hat{\boldsymbol{w}}_\lambda = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{Y}.$$

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

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

Bázové funkce

Mezi obvyklé volby bázových funkcí patří:

- $\varphi(\boldsymbol{x}) = x_i$ – přímo jednotlivé příznaky, odpovídají původnímu lineárnímu modelu.
- $\varphi(\boldsymbol{x}) = x_i^2$, $\varphi(\boldsymbol{x}) = x_k x_\ell$ – 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.
- $\varphi(\boldsymbol{x}) = \mathbb{1}_{(a,b)}(x_i)$, where $\mathbb{1}_A(x) = 1$ if $x \in A$ a $\mathbb{1}_A(x) = 0$ if $x \notin A$ – 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.

Různé volby umožňují získat dostatečně flexibilní model pro Y.

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.

Duální reprezentace

Lineární modely bázových funkcí pro regresi a klasifikaci můžeme v mnoha případech přeformulovat do duální reprezentace, ve které se bázové funkce vyskytují pouze implicitně a jsou určeny pomocí tzv. jádrových funkcí.

Začneme s regresním lineárním modelem bázových funkcí, kdy minimalizujeme

$$\mathrm{RSS}_\lambda(\boldsymbol{w}) = \left\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}\right\|^2 + \lambda \boldsymbol{w}^T \boldsymbol{w}.$$

Uvažujme nyní hodnoty $\boldsymbol{w}$ ve tvaru

$$\boldsymbol{w} = \boldsymbol{\Phi}^T \boldsymbol{\alpha} \quad \text{kde} \quad \boldsymbol{\alpha} \in \mathbb{R}^N,$$

tj. omezení $\boldsymbol{w}$ na podprostor $\mathbb{R}^M$ generovaný vektory $\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_N)$.

Dosazením do $\mathrm{RSS}_\lambda(\boldsymbol{w})$ dostaneme

$$\mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \left\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T \boldsymbol{\alpha}\right\|^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{\Phi}\boldsymbol{\Phi}^T \boldsymbol{\alpha}.$$

Duální reprezentace

Jelikož je $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ určeno zúžením $\boldsymbol{w}$ na $\boldsymbol{w}(\boldsymbol{\alpha}) = \boldsymbol{\Phi}^T\boldsymbol{\alpha}$ dostáváme

$$\min_{\boldsymbol{\alpha}} \mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}} \mathrm{RSS}_\lambda(\boldsymbol{w}(\boldsymbol{\alpha})) \geq \min_{\boldsymbol{w}} \mathrm{RSS}_\lambda(\boldsymbol{w}).$$

Následující věta ukazuje, že mezi hodnotami obou minim platí rovnost.

Věta

Pro $\lambda > 0$ platí

$$\min_{\boldsymbol{\alpha}} \mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}} \mathrm{RSS}_\lambda(\boldsymbol{w}).$$

Navíc, pokud $\boldsymbol{w}^*$ minimalizuje $\mathrm{RSS}_\lambda(\boldsymbol{w})$, tak $\boldsymbol{\alpha}^* = \frac{1}{\lambda}(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}^*)$ minimalizuje $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$.

Na druhou stranu, pokud $\boldsymbol{\alpha}^*$ minimalizuje $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$, tak $\boldsymbol{w}^* = \boldsymbol{\Phi}^T\boldsymbol{\alpha}^*$ minimalizuje $\mathrm{RSS}_\lambda(\boldsymbol{w})$.

Minimalizace $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ vzhledem k $\boldsymbol{\alpha}$ je tedy ve skutečnosti ekvivalentní původní minimalizaci $\mathrm{RSS}_\lambda(\boldsymbol{w})$ vhledem $\boldsymbol{w}$.

Duální reprezentace

Důkaz.

Je-li $\boldsymbol{w}^*$ argumentem minima $\mathrm{RSS}_\lambda(\boldsymbol{w})$, pak splňuje normální rovnici odpovídající $\nabla \mathrm{RSS}_\lambda(\boldsymbol{w}^*) = \mathbf{0}$:

$$\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}^*) - \lambda \boldsymbol{w}^* = \mathbf{0}. \tag{1}$$

Tudíž $\boldsymbol{w}^* = \frac{1}{\lambda}\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}^*)$ a pro $\boldsymbol{\alpha}^* = \frac{1}{\lambda}(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}^*)$ dostáváme $\boldsymbol{w}^* = \boldsymbol{\Phi}^T\boldsymbol{\alpha}^*$. Z toho plyne $\mathrm{RSS}_\lambda(\boldsymbol{\alpha}^*) = \mathrm{RSS}_\lambda(\boldsymbol{w}^*)$ což znamená

$$\min_{\boldsymbol{\alpha}} \mathrm{RSS}_\lambda(\boldsymbol{\alpha}) \leq \mathrm{RSS}_\lambda(\boldsymbol{\alpha}^*) = \mathrm{RSS}_\lambda(\boldsymbol{w}^*) = \min_{\boldsymbol{w}} \mathrm{RSS}_\lambda(\boldsymbol{w}).$$

Protože $\min_{\boldsymbol{\alpha}} \mathrm{RSS}_\lambda(\boldsymbol{\alpha}) \geq \min_{\boldsymbol{w}} \mathrm{RSS}_\lambda(\boldsymbol{w})$, ukázali jsme $\min_{\boldsymbol{\alpha}} \mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}} \mathrm{RSS}_\lambda(\boldsymbol{w})$ a tedy i první implikaci.

Pro obrácený směr uvažujme, že $\boldsymbol{\alpha}^*$ minimalizuje $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$. Potom také splní normální rovnici odpovídající $\nabla \mathrm{RSS}_\lambda(\boldsymbol{\alpha}^*) = \mathbf{0}$:

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}^*) - \lambda\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}^* = \mathbf{0}.$$

Nyní stačí ukázat, že $\boldsymbol{w}^*$ sestrojené jako $\boldsymbol{w}^* = \boldsymbol{\Phi}^T\boldsymbol{\alpha}^*$ splní rovnici (1).

$$\begin{aligned}\left\|\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{w}^*) - \lambda\boldsymbol{w}^*\right\|^2 &= \left(\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}^*) - \lambda\boldsymbol{\Phi}^T\boldsymbol{\alpha}^*\right)^T\left(\boldsymbol{\Phi}^T(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}^*) - \lambda\boldsymbol{\Phi}^T\boldsymbol{\alpha}^*\right) \\ &= \left(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}^* - \lambda\boldsymbol{\alpha}^*\right)^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\left(\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\alpha}^* - \lambda\boldsymbol{\alpha}^*\right) = 0.\end{aligned}$$

Protože norma vektoru je nula právě tehdy, když je vektor nulový, dostáváme výsledek $\square$

Gramova matice

Definujeme **Gramovu matici** (angl. **Gram matrix**) vztahem $\mathbf{G} = \mathbf{\Phi}\mathbf{\Phi}^T \in \mathbb{R}^{N,N}$.

Gramova matice je zjevně symetrická a protože

$$\boldsymbol{a}^T \mathbf{G} \boldsymbol{a} = \boldsymbol{a}^T \mathbf{\Phi}\mathbf{\Phi}^T \boldsymbol{a} = (\mathbf{\Phi}^T \boldsymbol{a})^T (\mathbf{\Phi}^T \boldsymbol{a}) = \|\mathbf{\Phi}^T \boldsymbol{a}\|^2 \geq 0$$

pro každé $\boldsymbol{a} \in \mathbb{R}^N$, je také pozitivně semidefinitní.

$\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ můžeme vyjádřit jako

$$\mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \|\boldsymbol{Y} - \mathbf{G}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha}.$$

Definujme dále **jádrovou funkci** (angl. **kernel function**) $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ jako

$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\varphi}(\boldsymbol{x})^T \boldsymbol{\varphi}(\boldsymbol{y}) \quad \text{pro každé} \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p.$$

Pro i, j-tou složku Gramovy matice platí

$$G_{i,j} = (\mathbf{\Phi}\mathbf{\Phi}^T)_{i,j} = \sum_{\ell=1}^{M} \Phi_{i,\ell}\Phi_{j,\ell} = \sum_{\ell=1}^{M} \varphi_\ell(\boldsymbol{x}_i)\varphi_\ell(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

tj. Gramova matice je plně určena jádrovou funkcí.

Predikce v duální reprezentaci

Předpokládejme, že jsme našli $\hat{\boldsymbol{\alpha}}$ minimalizující $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$.

Odpovídající $\hat{\boldsymbol{w}}$ minimalizující $\mathrm{RSS}_\lambda(\boldsymbol{w})$ je dáno vztahem

$$\hat{\boldsymbol{w}} = \boldsymbol{\Phi}^T \hat{\boldsymbol{\alpha}}.$$

Pro predikci Y v bodě $\boldsymbol{x}$ máme

$$\hat{Y} = \hat{\boldsymbol{w}}^T \boldsymbol{\varphi}(\boldsymbol{x}) = \hat{\boldsymbol{\alpha}}^T \boldsymbol{\Phi} \boldsymbol{\varphi}(\boldsymbol{x}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\alpha}_i \Phi_{i,j} \varphi_j(\boldsymbol{x}) = \sum_{i=1}^{N} \hat{\alpha}_i k(\boldsymbol{x}_i, \boldsymbol{x}),$$

Tudíž nejenom účelovou funkci[1] $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ ale také predikce $\hat{Y}$ můžeme vytvořit s využitím pouze jádrové funkce k.

Jak ukážeme, platí to i pro explicitní vyjádření $\hat{\boldsymbol{\alpha}}$.

[1] Název pro funkci, kterou minimalizujeme.

Explicitní řešení minimalizace

Najdeme explicitní řešení duální minimalizační úlohy.

Pro gradient $\nabla \mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ platí

$$\nabla \mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = -2\mathbf{G}(\boldsymbol{Y} - \mathbf{G}\boldsymbol{\alpha}) + 2\lambda\mathbf{G}\boldsymbol{\alpha}.$$

Položíme-li ho roven nule, dostaneme normální rovnici

$$\mathbf{G}(\boldsymbol{Y} - \mathbf{G}\boldsymbol{\alpha} - \lambda\boldsymbol{\alpha}) = \mathbf{0}.$$

Protože je Gramova matice pozitivně semidefinitní, je matice $(\mathbf{G} + \lambda\mathbf{I})$ regulární a dostáváme

$$\hat{\boldsymbol{\alpha}} = (\mathbf{G} + \lambda\mathbf{I})^{-1}\boldsymbol{Y}.$$

Tedy i optimální $\hat{\boldsymbol{\alpha}}$ můžeme spočítat pouze na základě Gramovy matice a tedy jádrové funkce.

Shrnutí duální reprezentace

- Při trénování minimalizujeme $\mathrm{RSS}_\lambda(\boldsymbol{\alpha})$ jakožto duální verzi reziduálního součtu čtverců danou vztahem

$$\mathrm{RSS}_\lambda(\boldsymbol{\alpha}) = \|\boldsymbol{Y} - \mathbf{G}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha},$$

  kde $G_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

- Řešení minimalizace je pro každé $\lambda > 0$ rovno

$$\hat{\boldsymbol{\alpha}} = (\mathbf{G} + \lambda\mathbf{I})^{-1}\boldsymbol{Y}.$$

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

$$\hat{Y} = \sum_{i=1}^{N} \hat{\alpha}_i k(\boldsymbol{x}_i, \boldsymbol{x}).$$

- Zároveň vidíme, že predikce $\hat{Y}$ v bodě $\boldsymbol{x}$ je vlastně váženou lineární kombinací hodnot bodů z trénovací množiny, kdy váhy jsou spočteny pomocí jádrové funkce. To umožňuje intepretovat výsledky predikce.

Jádrový trik

- Vytvořili jsme ekvivalentní duální reprezentaci celého modelu včetně účelové funkce pro trénování.
- V této reprezentaci se body $\boldsymbol{x}, \boldsymbol{x}_1, \ldots$ z originálního příznakového prostoru $\mathcal{X}$ vyskytují pouze ve tvaru skalárních součinů jejích transformací pomocí bázových funkcí (tj. skalárních součinů v novém příznakovém prostoru), $\boldsymbol{\varphi}(\boldsymbol{x})^T \boldsymbol{\varphi}(\boldsymbol{y})$, které můžeme kompletně vyjádřit pomocí jádrové funkce $k(\boldsymbol{x}, \boldsymbol{y})$.
- Toto nahrazení skalárních součinů pomocí jádrové funkce se nazývá **jádrový trik** (angl. **kernel trick**).
- Přirozené rozšíření je rovnou začít s jádrovou funkcí bez explicitního zavádění bázových funkcí.
- Tento přístup umožňuje implicitně pracovat v příznakových prostorech vysoké (nekonečné) dimenze.
- Z pohledu výpočetní náročnosti je dobré si uvědomit, že maticová inverze potřebná pro odhad $\boldsymbol{w}$ resp. $\boldsymbol{\alpha}$ má složitost $\mathcal{O}(M^3)$ resp. $\mathcal{O}(N^3)$.

Příklad využití jádrového triku

- Uvažujme například úlohu regrese nad 1000 obrázky o rozměrech $32 \times 32 = 1024$ pixelů ve stupních šedi.

- Při použití kvadratického jádra $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^2$ je ekvivalentní lineárnímu modelu bázových funkcí v $525\,825$ rozměrném prostoru.
- Tento počet odpovídajících bázových funkcí lze snadno určit, pokud položíme $x_0 = y_0 = 1$ a všimneme si, že

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} x_i y_i + 1 \right)^2 = \left(\sum_{i=0}^{n} x_i y_i \right)^2 = \sum_{i=0}^{n} \sum_{j=0}^{n} (x_i x_j) \cdot (y_i y_j)$$

  což má $(n+1) \cdot (n+2)/2 = 525\,825$ rozdílných složek.

- Při použití jádrového triku a duální reprezentace musíme invertovat matici 1000×1000 namísto matice $525\,825 \times 525\,825$.

Lineární a polynomiální jádrové funkce

Nyní si představme nejdůležitější příklady používaných jádrových funkcí.

Když $\mathcal{X} = \mathbb{R}^p$ a $\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{x}$ pro všechna $\boldsymbol{x} \in \mathbb{R}^p$ dostáváme **lineární jádro** (angl. **linear kernel**)

$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}.$$

Zobecněním je (nehomogenní) **polynomiální jádro** (angl. **polynomial kernel**)

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^n.$$

Bázové funkce, které toto jádro implicitně definuje si ukažme na příkladu $p = 2$ a $n = 2$:

$$\begin{aligned}(\boldsymbol{x}^T \boldsymbol{y} + 1)^2 &= (x_1 y_1 + x_2 y_2 + 1)^2 \\ &= (x_1 y_1)^2 + (x_2 y_2)^2 + 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 y_1 x_2 y_2 \\ &= \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2\right)\left(1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1 y_2\right)^T.\end{aligned}$$

Což odpovídá 6 bázovým funkcím.

Gaussovské jádro

Jednou z nejpoužívanějších jádrových funkcí je **Gaussovské jádro** (angl. **Gaussian kernel** nebo **squared exponential kernel**), které je definováno jako

$$k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\gamma \|\boldsymbol{x}-\boldsymbol{y}\|^2}.$$

pro každé $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$.

Tato jádrová funkce je příkladem **radiální bázové funkce** nebo také **RBF jádra**, což znamená, že je to funkce pouze $\|\boldsymbol{x} - \boldsymbol{y}\|$.

Bývá ale také často označována pouze jako **RBF jádro** (viz např. ve `scikit-learn`).

Existuje i obecnější formulace Gaussovského jádra, kde mohou být různé škály ve směrech různých příznaků.

Existují i další jádrové funkce, z nichž některé umí pracovat i s nečíselnými vstupy (např. s dvojicemi konečných řetězců).

Obecně, aby vše „dobře fungovalo“, musí být jádrová funkce pozitivně semidefinitní symetrická funkce, typicky také nezáporná. Co to přesně znamená si zde ovšem nebudeme podrobněji vysvětlovat.

Jádrové modely

- Uvažujme, že skutečný model, ze kterého pochází vysvětlovaná proměnná Y v bodě $\boldsymbol{x}$ je
$$Y = f(\boldsymbol{x}) + \varepsilon,$$
kde f je nějaká neznámá funkce a ε je náhodná veličina s $\mathrm{E}\,\varepsilon = 0$.
- V lineárním modelu bázových funkcí hledáme f ve tvaru
$$f(\boldsymbol{x}) = w_1\varphi_1(\boldsymbol{x}) + \dots w_M\varphi_M(\boldsymbol{x}).$$
- V obecném **jádrovém modelu** (angl. **kernel machine**) s jádrovou funkcí k máme f ve tvaru
$$f(\boldsymbol{x}) = \sum_{j=1}^{K} \alpha_j k(\boldsymbol{\mu}_j, \boldsymbol{x})$$
kde $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K \in \mathcal{X}$ jsou nějaké středové body.
- Speciální případ jádrových modelů (angl. **vector machines**) nastává, když středové body jsou body trénovací množiny, tj.
$$f(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}).$$
- Takový regresní model jsme dnes přesně získali pomocí jádrového triku.