Malá velká data

A. Horák*, F. Svoboda**, M. Štefaňák***, Jakub Šuráň****
Střední průmyslová škola Třebíč, **Gymnázium Velké Meziříčí,
***Gymnázium Jána Chalupku Brezno, ****Purkyňovo gymnázium Strážnice
*horak.alda@seznam.cz, **svobofilip@seznam.cz,
***michalstefanak22@gmail.com, ****suranjakub@email.cz

Abstrakt

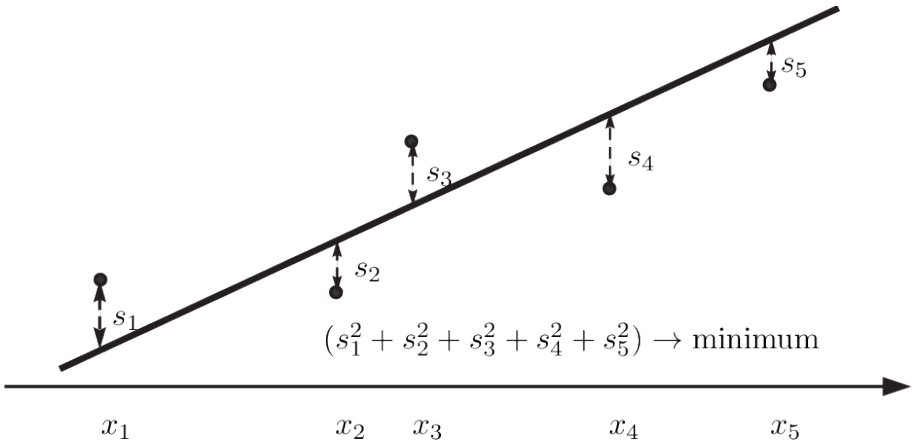
Cílem našeho projektu bylo seznámení se s jednoduchými metodami strojového učení a jejich aplikace na data ze serveru Kaggle.com. Metody jsme modifikovali pro dosáhnutí nejlepšího možného výsledku.

1 Úvod

Data jsou v dnešní době součástí každého oboru a tyto data je potřeba zpracovat. K tomu se často uplatňují rozmanité metody strojového učení. Strojové učení je odvětví počítačových věd zabývající se učením se strojů ze zkušeností. Algoritmy automatizovaného učení uchopí existující data, pročešou je, pokusí se rozpoznat vzory, a z těch potom počítají odhady budoucích událostí. Na data jsme prvně aplikovali jednoduchou metodu nejmenších čtverců. Metodou lze řešit soustavy rovnic, které nemají konrétní řešení. Nejmenší čtverce znamenají, že výsledné řešení má minimalizovat součet čtverců tedy odchylek vůči každé rovnici, a tak nalézt nejoptimálnější řešení. Naše výsledky jsme porovnávali s ostatními, což umožňuje server Kaggle.com. Postupným upravováním našich metod jsme se snažili dosáhnout co nejvyšších příček.

2 Kaggle

Kaggle je platforma pro prediktivní modelování a analytické soutěže. V nich mohou uživatelé využít svoje vědomosti na vytváření modelů pro předvídání a popis datových souborů. Soubory jsou nahrány uživateli a společnostmi, které nabízí odměnu nejlepším řešitelům daného problému. Celý server, data na něm a soutěže jsou volně přístupné komukoli. Účastníci soutěže musí předpovědět výsledky na základě vstupních dat. Výsledky zaslané soutěžícími jsou porovnány se skutečnými výsledky a je vypočítána jejich vzájemná odchylka. Na základě odchylky jsou zařazeni do žebříčku a ti nejlepší získají pěněžní odměnu. Kaggle zároveň vytváří komunitu, kde si uživatelé můžou navzájem radit, diskutovat o svých metodách a rozvíjet svoje znalosti. Server je momentálně vlastněn společností Google.



Obrázek 1: Metoda nejmenších čtverců [1]

3 Metody

Vstupní data jsme převedli na matici X, kde je 12 zadaných vlastností materiálů, d = 12. Matice y reprezentuje hledané parametry, oba se vyhodnocují samostatně.

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^d \\ x_2^1 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^d \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

V množině reálných čísel jsme hledali koeficient α, tak aby

$$\min_{\alpha \in \mathbb{R}^d} \|y - X\alpha\|^2 \tag{1}$$

bylo co nejmenší (zde $\|z\|^2 = z_1^2 + \dots + z_d^2$ je norma vektoru z). Derivováním (1) podle všech komponent vektoru α a položením rovno nule, získáme soustavu lineárních rovnich, jejichž řešení nalezneme ve tvaru

$$\alpha = (X^T X)^{-1} X^T y.$$

Pokud máme poté předpovědět výstupní parametry pro nové, testovací materiály, načteme novou tabulku hodnot, kterou reprezentujeme maticí

$$Z = \begin{pmatrix} z_1^1 & z_1^2 & \dots & z_1^d \\ z_2^1 & z_2^2 & \dots & z_2^d \\ \vdots & \vdots & & \vdots \\ z_m^1 & z_m^2 & \dots & z_m^d \end{pmatrix}.$$

Výsledné odpovědi pak získáme jako prostý součin

$$w = Z\alpha.$$

Pro zvýšení numerické stability je vhodné se vyhnout příliš velkým vektorům α. To je cílem metody zvané Tichonovova reguralizace, která zavádí kladný reálný parametr λ a minimalizuje

$$\min_{\alpha \in \mathbb{R}^d} \|y - X\alpha\|^2 + \lambda \|\alpha\|^2.$$

Stejným způsobem jako pro rovnici (1) najdeme řešení ve tvaru

$$\alpha = (X^T X + \lambda I)^{-1} X^T y.$$

I tato metoda ovšem pracuje s předpokladem, že výstupní data závisí lineárně na vstupních datech. Toto omezení do značné míry eliminuje metoda Kernel Ridge Regression, a to velice jednoduchým tzv. Kernel Trickem. Tento trik spočívá v tom, že při výpočtu součinu matice X a X^T nahrazuje skalární součin $\langle x, x' \rangle$ nelineární funkci

$$e^{-\|x-x'\|^2/2}.$$

4 Aplikace

Data, které jsme zpracovávali, se týkaly průhledných polovodičů. Byly předmětem soutěže od firmy NOMAD hledající nejvhodnější materiály pro solární panely. Jednotlivé materiály byly popsány 12 vlastnostmi, které se týkaly jejich složení (Al, Si, In), krystalických struktur a vzdálenosti molekul, a 2 parametry určující účinnost a odolnost z nich vyrobených solárních panelů. Z toho jsem měli odvodit vztah mezi 12 zadanými vlastnostmi materiálu a 2 parametry. Pomocí nalezeného vztahu jsme ze zadaných vlastností předpovídali parametry u dalších materiálů. Daný vztah jsme nalezli pomocí metody Kernel Ridge Regression.

5 Shrnutí

S postupnými modifikacemi algoritmů jsme dosáhli poměrně dobrých výsledků ve srovnání s lidmi z celého světa. Za pomocí relativně jednoduché metody se nám podařilo dostat vysoko v žebříčku. Původní chybovost 0,1205 jsme zredukovali na 0,0558, vítěz soutěže měl chybovost 0,0368. Naučili jsme se základy strojového učení, využití metody Kernel Ridge Regression a práce v programu MatLab.

Poděkování

Chtěli bychom poděkovat za podporu při práci doc. Janu Vybíralovi, PhD.. Děkujeme i FJFI za zorganizování Týdne vědy.

Reference

[1] . Mařík. *Metoda nejmenších čtverců*. http://user.mendelu.cz/marik/mat-web/mat-webse24.html. 2008.