Malá velká data

Petr Kolář*, Daniel Sůva**, Bohdana Prchlíková***, Patrik Mitterpach****
Gymnázium Milevsko*, Gymnázium Stříbro**, Gymnázium Děčín***, Stredná športová škola Banská Bystrica****
petr.kolar@gymnazium-milevsko.cz, daniel.suva@goas.cz, bohdanka.prchlikova@seznam.cz, patrikmitterpach@gmail.com

Abstrakt:
Naším cílem bylo seznámit se se základní metodou strojového učení a poté aplikovat náš model na data ze serveru Kaggle.com, týkající se analýzy materiálů, potenciálně využitelných pro konstrukci solárních panelů.

1 Úvod

Při strojovém učení vyžadujeme, aby náš model dával užitečné předpovědi o datech, se kterými nepřišel do styku. Toho se snažíme dosáhnout na základě informací z dat, která mu poskytneme a která mají labely, to znamená, že známe jejich správné výstupy. My jsme použili základní metodu učení s učitelem (supervised learning), regresi. Obecně se snažíme minimalizovat nějakou funkci, udávající chybu v našem odhadu, v našem případě jsme použili metodu nejmenších čtverců (least squares). Na to je potřeba určitý matematický aparát, především z lineární algebry.

2 Kaggle

Kaggle.com je platforma, aktuálně vlastněna společností Google, sloužící pro prediktivní modelování a analytické soutěže. Uživatelé se v těchto soutěžích snaží vytvořit co nejefektivnější model pro předvídání a popis datových souborů. Datové soubory a zadání soutěže jsou nahrávány uživateli a společnostmi, které nabízejí nejúspěšnějším řešitelům finanční ocenění. Na základě vzorku s výsledky se řešitelé snaží o sestavení co nejefektivnějšího modelu pro zbytek poskytnutých dat. Tyto předpovědi jsou porovnány se skutečnými výsledky a na základě velikosti odchylky je sestaven žebříček. Server také vytváří komunitu, kde si uživatelé navzájem mohou radit, diskutovat a rozvíjet své stávající schopnosti.

3 Metody

Vstupní data uspořádáme do matice X, výstupní data, která známe do vektoru y. Nyní hledáme vektor α, který má tvar jako y, obsahuje koeficienty (váhy) a když ho budeme násobit maticí X,

dostaneme vektor opět stejného tvaru, který bude mít co možná nejmenší euklidovskou vzdálenost od vektoru y.

Minimalizace euklidovské vzdálenosti je ekvivalentní s použitím metody nejmenších čtverců.

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

Vztah pro optimální α, když ho zjistíme, umíme pro nové inputy jednoduše předpovědět outputy.

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

O něco rafinovanější metodou je tkzv. Tichonovova regularizace, při jejím použití penalizujeme i velikost α. Ukazuje se, že velké α se nechovají dobře na testovacích datech, ačkoliv při tréninku může toto větší α dávat o něco málo lepší výsledky.

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

Vztah pro α se nepatrně liší.

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

Při obou zmíněných metodách předpokládáme lineární závislost výstupních dat na vstupních. Toho se zbavíme například použitím metody Kernel Ridge Regression, která při násobení matic X a X^T nahrazuje skalární součin $\langle x, x' \rangle$ nelineární funkcí.

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

4 Aplikace

Námi zpracovávaná data se týkala průhledných polovodičů, vydaných firmou NOMAD, která hledá materiál pro nové solární panely, který musí být co nejvíce energeticky výhodné a mít co nejdelší životnost. Jednotlivé materiály byly popsány 12 parametry, které se týkaly chemického složení (obsah Al, Ga, In, O v molekule) a typu krystalické struktury (vzdálenost a vzájemná poloha atomů). Pomocí metody Kernel Ridge Regression jsme nalezli vztah mezi zadanými

hodnotami a požadovanými parametry, díky kterému jsme pak byli schopni přesněji předpovídat parametry dalších materiálů.

5 Shrnutí

Naši původní chybovost 0,1205 jsme postupnou optimalizací našeho postupu zredukovali na 0.0625, nejlepším výsledkem je 0,0368. Náš algoritmus dosahoval poměrně dobrých výsledků, nicméně hlavním přínosem pro nás bylo rozšíření znalostí z lineární algebry, v základech strojového učení a v práci v Matlabu.

Poděkování

Rádi bychom poděkovali doc. RNDr. Janu Vybíralovi, Ph.D. za výuku a pomoc při projektu, dále Bc. Anně Doležalové a Fakultě jaderné a fyzikálně inženýrské za organizaci Týdne vědy.

Reference:

https://www.kaggle.com/deepak2873/nomad-transparent-conductor-ensembled-regressors

ERNHARD SCHOLKOPF, ALEXANDR SMOLA, KLAUS-ROBERT MULLER
K*ernelprincipal component analysis* 1997 pp 583 - 588