Regrese nejmenších čtverců s použitím SGD

Jakub Dorian Charbulák

ČVUT–FIT

charbjak@fit.cvut.cz

17. května 2021

Strukturu sekcí reportu si pochopitelně můžete upravit, aby vám co nejlépe vyhovovala pro popis vaší práce.

1 Úvod

Report informuje o semestrální práci pro předmět BI-PYT v letním semestru B202.

Práce je na téma Ordinary least squares regression - metody nejmenších čtverců užívané ve statistice k nalezení neznámých parametrů lineárních rovnic. K tomu je v práci dále využito iterativní metody Stochastic gradient descent.

Cílem práce bylo zobecnit tyto postupy, aby bylo možno pomocí algoritmu hledat lineární funkci o libovolném počtu parametrů a tuto funkčnost ověřit pomocí bostonského datasetu cen bydlení. V řešení jsou dle zadání zahrnuty výpočty s NumPy, načítání dat s Pandas, vykreslování grafů s Matplotlib, projekce v Jupyter notebooku a PyTesty.

2 Vstupní data

Vstupní dataset boston.csv je ze stránky Kaggle [1] [1]. S .csv vstupními daty pracuje python modul csv_interface.py - metoda read_from_csv().

Metoda načte soubor dle zadané cesty, odstraní případný indexovací sloupec a řádek názvů parametrů (toto nastavení je dáno vstupními parametruy metody). Zbylá data jsou převedena na NumPy pole a rozložena na matici X - vektory parametrů jednotlivých domů a na vektor y - výsledné ceny domů. X a y jsou ještě dle zadání rozloženy na 80% trénovacích a 20% testovacích dat.

3 Metody/postupy/algoritmy

3.1 OLS, error function

Z metody nejmenších čverců nám vyplýva vzorec pro tzv. Error function každého z vah parametrů: $1/\mathrm{n}\sum((y_i - \hat{y}_i)^2)\frac{\mathrm{d}f}{\mathrm{d}x} prokaždýparametrx.$

[1] na stránce zadání SP je neplatný link na stránku tohoto datasetu na scikit-learn

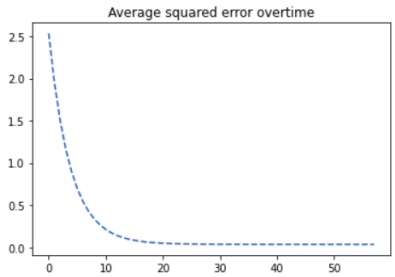
3.2 SGD

Iterativní odečítání násobku derivovaného error function.

$\mathrm{w}_i = w_i - learningRate * D_w 0$

4 Výsledky

OLS i SGD se zobecnit povedlo, pro jejich aplikaci na jiný dataset je třeba správně odhadnout počínající váhy parametrů a learning rate.



Zde je vidět historie iterací SGD - průměrná čtvercová chyba klesá, dokud nenarazí na minimum.

5 Závěr

Toto zadání nebylo tolik složité na počet řádků kódu, jako na nastudování použité matematiky a hledání algoritmů pro funkčnost a efektivitu programu. Ač výzkum použitých metod zabral více času než samotné programování, jsem velmi rád za dokončení svého prvního Machine Learning programu.

Reference

[1] Feature scaling. online. [cit. 2021–05–17] https://www.geeksforgeeks.org/ml-feature-scaling-part-2/.

[2] Statquest: Linear models pt.1 - linear regression. online. [cit. 2021–05–17] https://www.youtube.com/watch?v=nk2CQITm_eo&t=1048s.

[3] Towards data science - linear regression. online. [cit. 2021–05–17] https://towardsdatascience.com/linear-regression-using-gradient-descent-97a6c

[4] Vzorový report. online. [cit. 2021–05–17] https://courses.fit.cvut.cz/BI-PYT/semestral-project/vzor-reportu_dlabamat.pdf.

[5] https://www.kaggle.com/puxama. boston.csv z kaggle. online, 2018. [cit. 2021–05–17] https://www.kaggle.com/puxama/bostoncsv?select=Boston.csv.