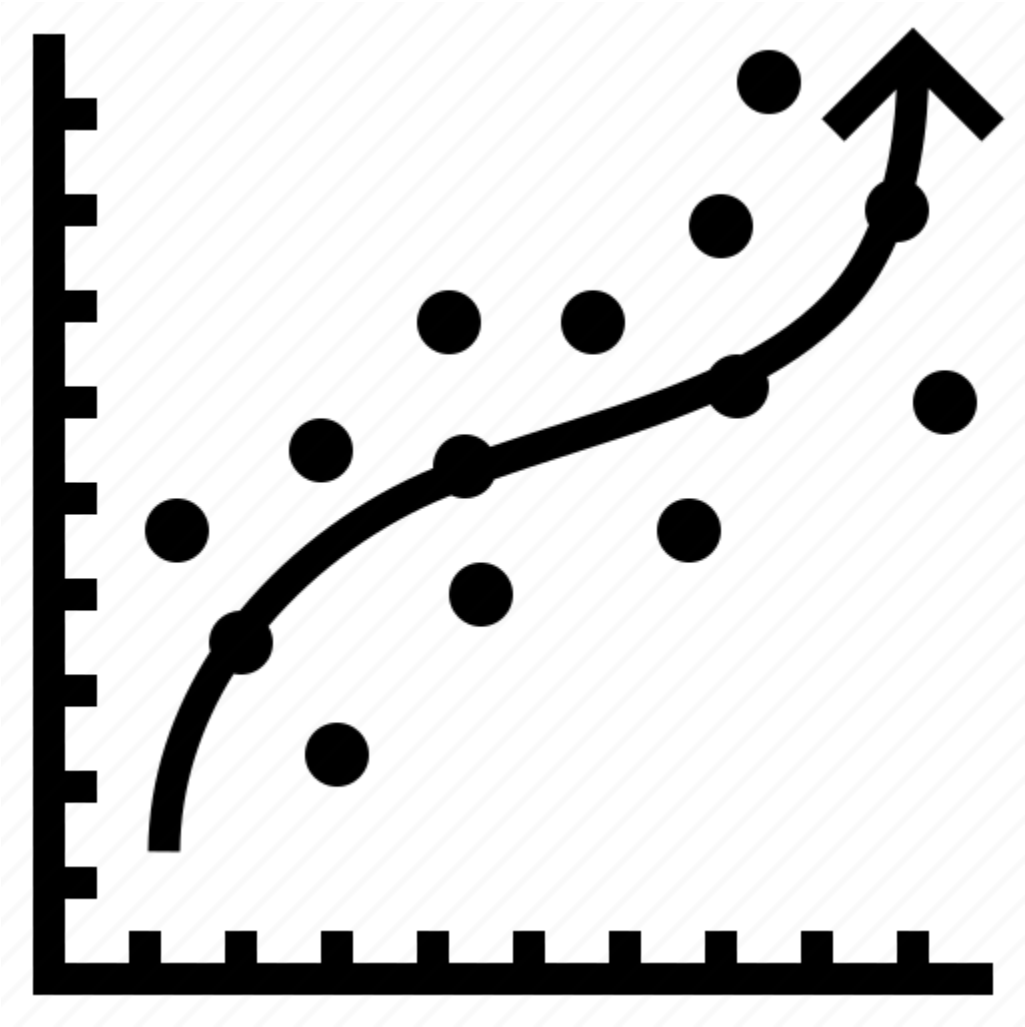
ИЗВЕШТАЈ ЗАДАТКА 1

Имплементација модела једноструке линеарне регресије



Душан Бркић, Филип Живанац

27. март 2022.

Софтверско инжењерство и информационе технологије
Факултет техничких наука
Универзитет у Новом Саду

ЗАДАТАК

Скуп података садржи зависност резидуалне отпорности брода по јединици тежине током померања (Y) у односу на његове димензије и брзину кретања (X). Пронаћи регресиони модел који се најбоље уклапа у дати скуп података. Задатак се решава употребом једноструке линеарне регресије и сматра се успешно урађеним уколико се на комплетном тестном скупу података добије RMSE (Root mean square error) мањи од 9.1. Алгоритме машинског учења имплементирате сами (забрањена употреба алгоритама из готових библиотека).

ПРИСТУП ПРОБЛЕМУ

За решење проблема најпре смо имплементирали два регресиона алгоритма и исцртали скуп тачака на X-Y оси, како би имали бољи увид у скуп података. Пошто смо приметили да функција у облику праве не би дала довољно добар резултат за дати проблем, одлучили смо се за коришћење полиномске функције. Испробавањем различитих бројева степена функције и других параметара алгоритама упоређивали смо решења и одредили се за оптималан однос резултата (RMSE) и робусности. Такође, приметили смо да дати скуп података има одређен број аутлајера (eng. outliers), тако да смо одлучили да такве тачке уклонимо из тренинг скупа, због чега два пута тренирамо модел.

ИСПРОБАНИ АЛГОРИТМИ

STOCHASTIC GRADIENT DESCENT

Овај алгоритам функционише по принципу смањивања грешке функције регресије итеративно. Пробали смо да га применимо користећи први и други степен регресионе функције али нисмо били задовољни резултатима па смо прешли на следећи метод.

RIDGE (L2 REGULARIZATION)

Ова регресија је уснови регресија најмањих квадрата с тим да има у себи и параметар α (алфа). Уколико α=0 овај модел има исто решење као и регресија најмањих квадрата. Повећавањем α постижемо то да се наш модел слабије уклапа у тренинг скуп података али се зато смањује варијанса са различитим тестним скуповима података што нам за овај проблем одговара јер нам тестни скуп није потпуно познат.

РЕЗУЛТАТИ

STOCHASTIC GRADIENT DESCENT

LEARNING RATE	БРОЈ ИТЕРАЦИЈА	СТЕПЕН ФУНКЦИЈЕ	RMSE (СА АУТЛАЈЕРИМА)	RMSE
10^{-1}	1000	4	5.790780718798237	4.368790044589453
10^{-4}	500	8	3.637916366945067	3.948707773524382
10^{-4}	200	4	3.5095420728112408	3.8780720913311466
10^{-8}	10000	8	3.56103941606852	3.4957366033126025
10^{-2}	100	4	4.653404316607075	2.7282939979682332

RIDGE (L2 regularization)

АЛФА ВРЕДНОСТ	СТЕПЕН ФУНКЦИЈЕ	RMSE (СА АУТЛАЈЕРИМА)	RMSE
1	2	5.9670448513553485	4.244652161568299
10^{-3}	4	3.0748690809896977	2.8547060960570008
10^{-6}	4	1.320065737875306	1.2028042013565907
10^{-6}	8	1.162531230243723	0.8882380332660963
10^{-7}	**4**	**1.1279266879032737**	**0.824775782707209**

10^{-7}	8	1.1531074377662796	0.7757827020421239
10^{-8}	12	1.20235984253517	0.7616155016828964
10^{-8}	4	1.1117144134895207	0.7353972542723556
10^{-10}	4	1.1121376766325863	0.7263899302572988

ОДАБРАНО РЕШЕЊЕ

Као одабрано решење одлучили смо се за Ridge регресију 4. степена α вредности 10^{-7}. Изабрали смо се за коришћење парних степена јер се тренинг подаци могу добро уклопити у параболу. Поред тога што су мање α и већи степен давали боље резултате, веће алфа и мањи степен су давале робусност моделу што је врло битно за конкретан задатак, јер нам тестни скуп података није познат. Такво решење подразумева што веће α и што мањи степен, тако да RMSE није драстично већи од мање робусних решења.