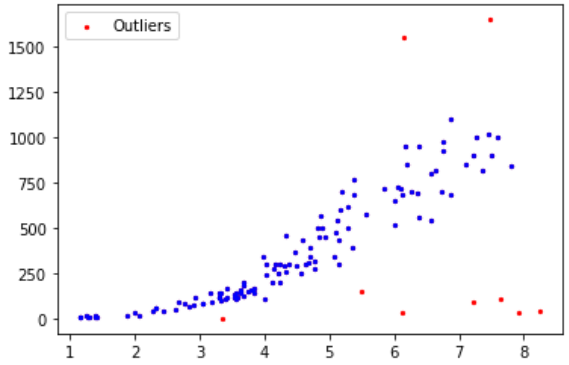
Linearna regresija

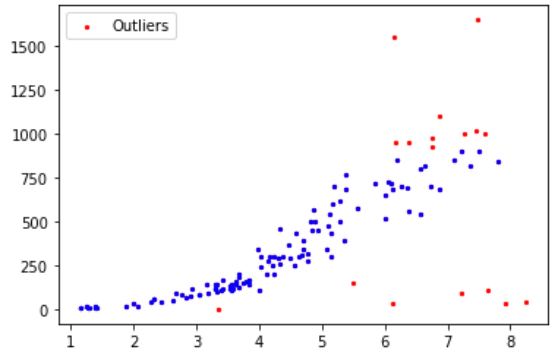
Za problem linearne regresije testirali smo sledeće algoritme *Normal Equation* i *Gradient descent*. Funkcija greške koju smo koristili je RMSE.

Uklanjanje outilier-a

Ekstreme koje smo odabrali da uklonimo su one tačke koje imaju veoma mali broj tačaka u blizini.



Ilustracija 1 Kritične tačke



Ilustracija 2 Kritične tačke sa pooštrenim kriterijumom

Testirali smo i uklanjanje tačaka sa pooštrenim kriterijumom za *y* vrednosti gde su nam se rezultati poboljšali, međutim kako podaci najviše podsećaju na ekponecijalnu funkciju odlučili smo da bi ovakvo odesecanje moglo uticati na uklanjanje važnih podataka.

Podela podataka na trening i validacioni skup

Razmera za podelu podataka na koju smo se odlučili je 80% za trening skup i 20% za test skup. Da bi podaci ostali izbalansirani podatke smo podelili u 5 kategorija tako što smo iterirali kroz sortiranu kolekciju elemenata i smeštali jedan po jedan element u odgovarajući skup. Na kraju smo na nasumičan način odabrali jedan skup da bude validacioni, dok smo ostale smo spojili u trening skup.

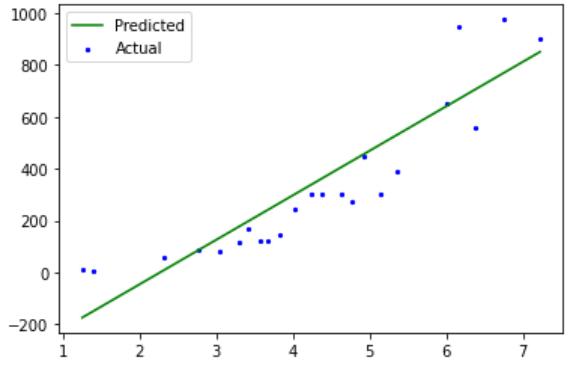
Rezultati testiranja po algoritmu

Tabela 1 RMSE za testirane algoritme kroz 5 iteracija

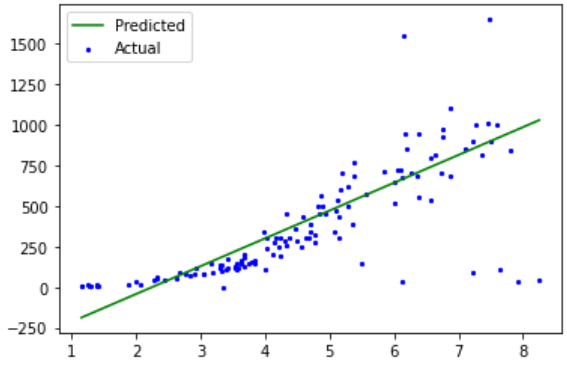
	Normal equation	**Batch gradient descent**	**Stohastic gradient descent**	**Mini batch gradient descent**
1	109.7420155551421	110.54503195914123	117.39054939470921	110.25330111496562
2	109.7420155551421	110.54503195914123	116.60988607722922	111.35543664517522
3	109.7420155551421	110.54503195914123	127.64124965584568	109.72636382233387
4	109.7420155551421	110.54503195914123	109.68965397274926	110.04056423958487
5	109.7420155551421	110.54503195914123	113.36854136138804	110.58385953007708

U tabeli iznad zelenom bojom predstavljen je rezultat sa najmanjim RMSE, dok je crvenom bojom predtsavljen rezultat čiji RMSE prekoračuje zadovoljavajući prag tolerancije.

Analizirajući podatke iz tabele možemo primetiti da *Normal equation* algoritam najkonzistentnije daje najbolji rezultat. U nekim slučajevima se može videti da najbolji rezultat daje *Stohastic gradient descent*, međutim kako u drugim iteracijama rezultat prekoračuje prag tolerancije odlučili smo da je *Normal equation* za ovaj skup.



Ilustracija 3 Poređenje dobijenog rešenja sa vrednostima testnog skupa



Ilustracija 4 Poređenje dobijenog rešenja sa vrednostima kompletnog skupa

Članovi tima

- Milovanović Miloš SW 17/2019
- Stojanov Dunja SW 30/2019