RAPORT, Mateusz Zacharecki, 313549

1 Cel

Celem pracy domowej jest implementacja dwóch różnych algorytmów optymalizacji dla regresji logistycznej i porównanie ich wydajności.

2 Dane

W pracy wykorzystano 3 różne zbiory danych dla problemu klasyfikacji binarnej pochodzące z repozytorium danych https://www.openml.org/. Wybrane zbiory danych to:

1. sa-heart (South Africa Heart Disease Dataset) - 9 zmiennych objaśniających, 462 obserwacji
2. fertility – 9 zmiennych objaśniających, 100 obserwacji
3. heart-statlog – 13 zmiennych objaśniających, 270 obserwacji

Zbiory przygotowano pod model regresji logistycznej. Sprawdzono braki danych (brak braków danych), z pierwszego zbioru usunięto zmienną V4, która jest współliniowa ze zmienną V7 i V9, przekształcono zmienną nominalną V5 z pierwszego zbioru na numeryczną oraz dane przeskalowano.

3 Implementacja algorytmów optymalizacji

W ramach pracy zaimplementowano algorytmy optymalizacji do estymacji parametrów w regresji logistycznej:

1. Gradient Descent (z domyślnym parametrem uczenia 0.01 oraz z domyślną liczbą epok 500)
2. Stochastic Gradient Descent w wersji standardowej (z domyślnym parametrem uczenia 0.01 oraz z domyślną liczbą epok 500)
3. Stochastic Gradient Descent w wersji mini batch (z domyślnym parametrem uczenia 0.01, z domyślną liczbą epok 500 oraz z domyślną wielkością batch 0.01)

4 Analiza

Podczas analizy oprócz zaimplementowanych metod z części 3, rozważono gotową implementację metody Iterative Reweighted Least Squares.

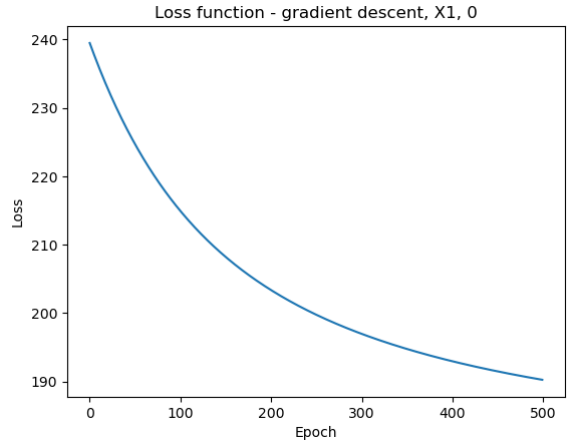
4.1 Reguła stopu

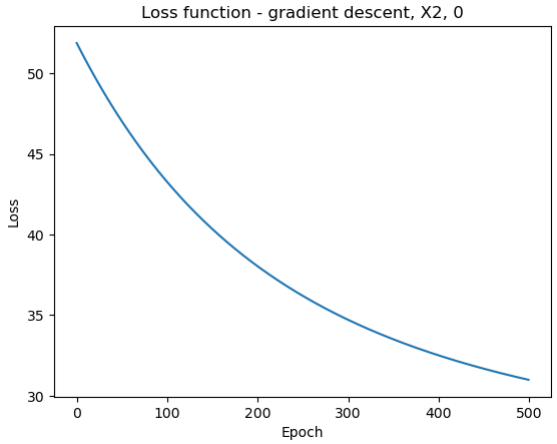
Jako regułę stopu wybrano liczbę epok. Ponadto uznaje się, że algorytm zbiega, jeżeli norma L^2 gradientu w danej iteracji jest mniejsza od 0.0001.

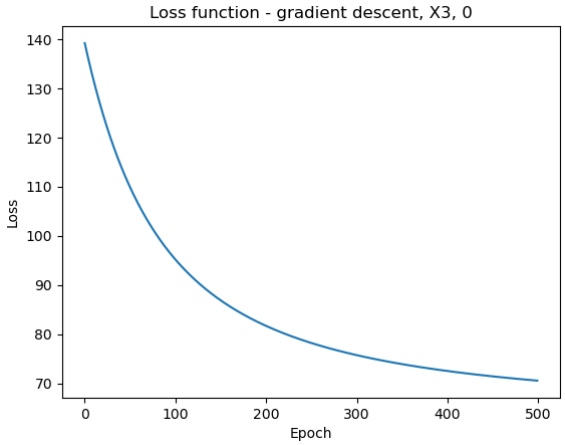
4.2 Analiza zbieżności

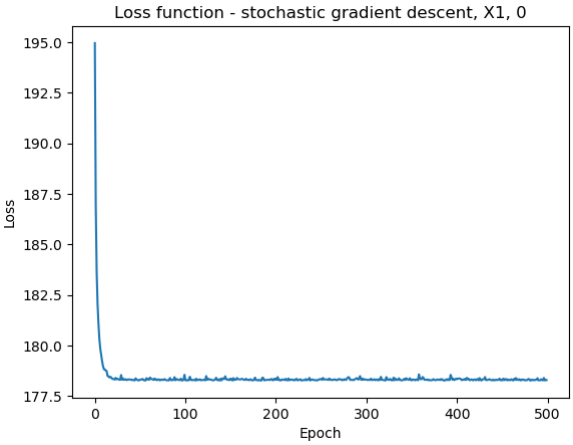
Dla rozważanych zbiorów danych przetestowano działanie rozważanych metod. Każdą metodę przetestowano 5-krotnie w zależności od wartości ziarna (0,1,2,3,4) przy podziale zbioru na zbiór

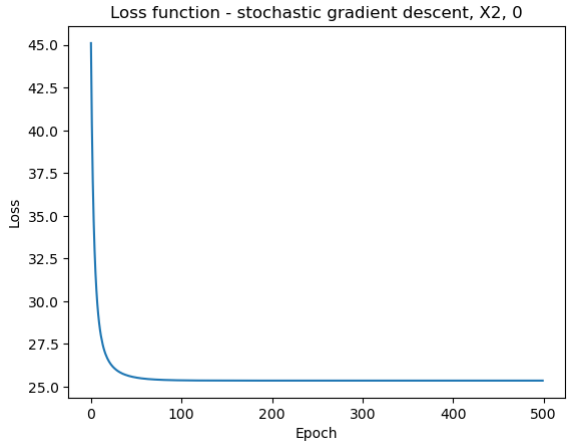
treningowy i testowy. Iteracje dla każdego algorytmu wykonywano do osiągnięcia zbieżności bądź do osiągnięcia 500 iteracji. Poniższe wykresy przedstawiają wartości funkcji log-wiarogodności w zależności od iteracji dla każdej z rozważanych metod, dla każdego z rozważanych zbiorów oraz dla ziarna równego 0 przy podziale zbioru na zbiór treningowy i testowy.

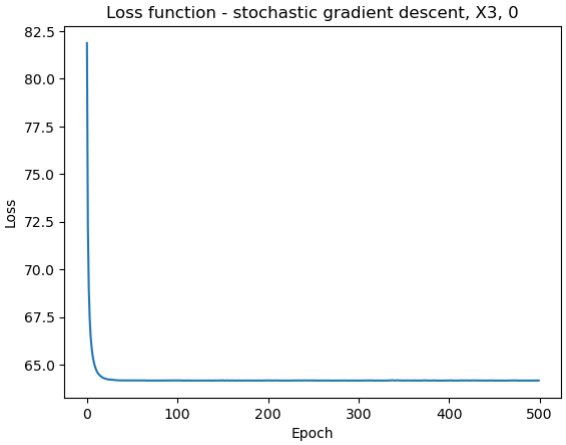


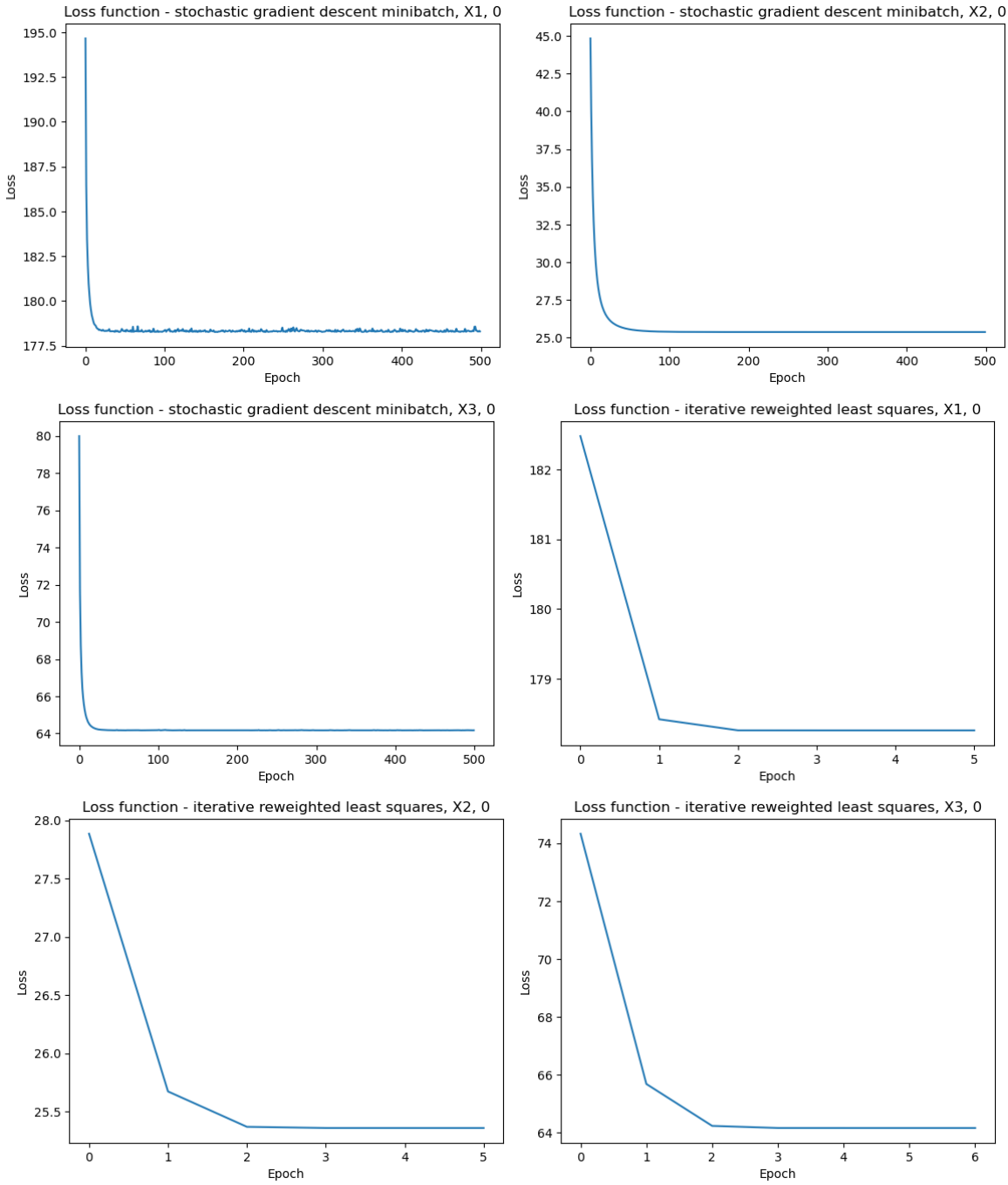












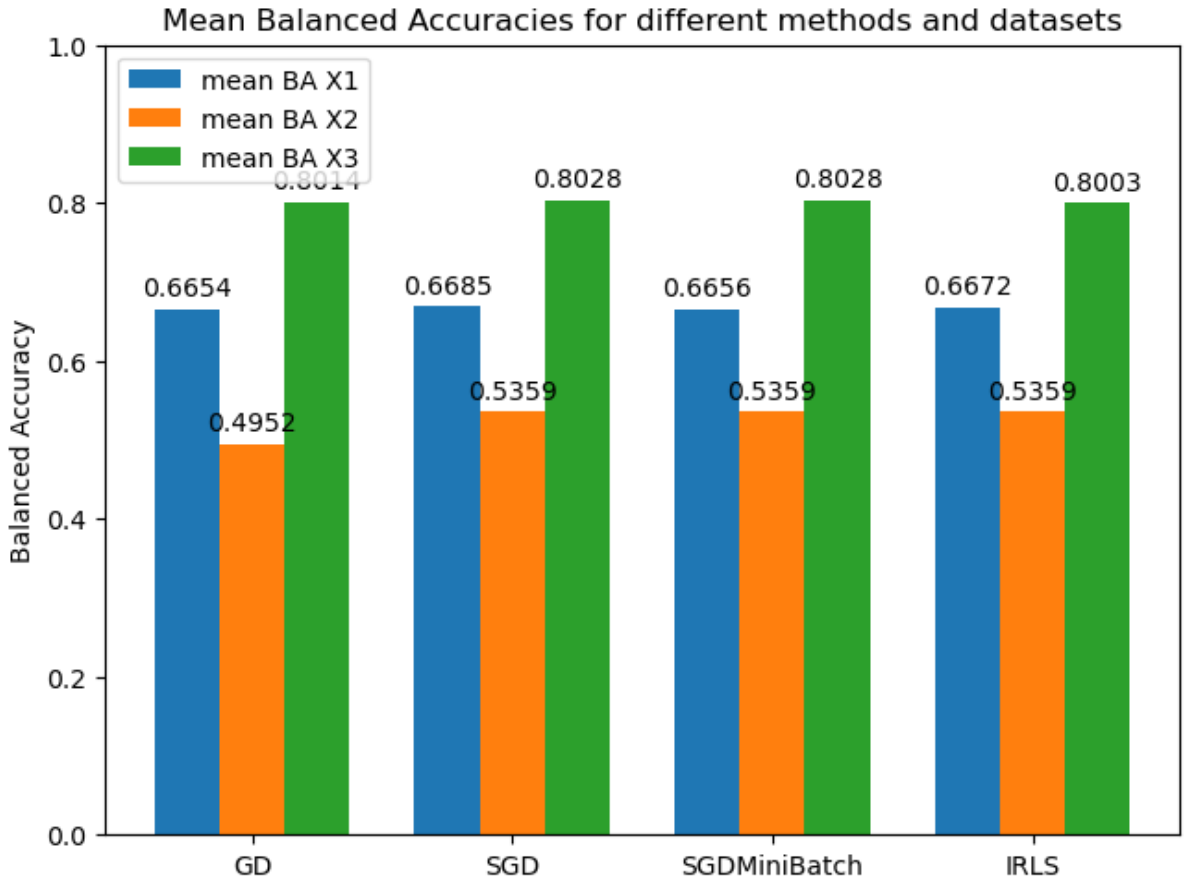
Rysunek 1. Straty dla rozważanych metod i zbiorów w zależności od iteracji

Jedyna metoda, która zbiega w mniej niż 500 krokach to IRLS. Różnica jest znacząca w porównaniu z innymi metodami – IRLS zbiega w zaledwie kilku iteracjach. Gradient Descent wydaje się zbiegać wolno – log-wiarogodność w zależności od kroków przypomina kształtem funkcję 1/x. Z kolei strata dla SGD i SGD Mini Batch dla rozmiaru batch = 20 zachowuje się podobnie - wartość straty w pierwszych krokach szybko spada, po czym wydaje się oscylować wokół jednej wartości.

4.3 Analiza jakości modeli

Jakość modeli oceniono za pomocą miary zrównoważonej dokładności. Modele trenowano na zbiorze uczącym, natomiast miarę obliczano na danych testowych. Wyniki uśredniono z 5 podziałów trening-test. Jeśli dany algorytm nie osiągnął zbieżności w ciągu 500 iteracji, użyto rozwiązania z

ostatniej iteracji. Poniższy wykres przestawia średnią miarę balanced accuracy w zależności od użytego algorytmu dla każdego z rozważanych zbiorów.



Rysunek 2. Wartości balanced accuracy w zależności od rozważanych metod dla rozważanych zbiorów

Średnie wartości balanced accuracy dla odpowiednich zbiorów nie różnią się znacząco między metodami. Jedynie istotna różnica jest dla zbioru 2 – metoda gradient descent daje wówczas wartość miary zrównoważonej dokładności o około 0.04 niższą w porównaniu z pozostałymi algorytmami.

5 Wnioski

Metody wydają się nie różnić między sobą pod względem jakości wyników. Istotne różnice są jednak w czasie działania algorytmów. Z rozważanych metod pod tym względem zdecydowanym liderem wydaje się być IRLS, który dla wszystkich zbiorów osiągał zbieżność w zaledwie kilka kroków. Z kolei sądząc po wykresach straty w iteracjach można się spodziewać, że najwolniejszą (a tym samym najgorszą) z metod jest Gradient Descent.

Oczywiście dla uzyskania dokładniejszych wniosków należałoby powtórzyć testy więcej razy, dla większej liczby zbiorów, dla różnych wartości parametrów (batch size, learning rate) i dla większej liczby epok.