Praca domowa 1

Zaawansowane Metody Uczenia Maszynowego

Termin oddania: 28.03.2024

1 Cel

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

2 Dane

W tej pracy domowej wykorzystaj 3 różne zbiory danych dla problemu klasyfikacji binarnej. Można skorzystać z repozytoriów danych takich jak: https://archive.ics.uci.edu/, https://www.openml.org/ lub innych źródeł. Wybierz **dwa małe** zbiory danych zawierające co najwyżej 10 zmiennych i **jeden duży** zbiór danych zawierających więcej niż 10 zmiennych. W przypadku wszystkich zbiorów danych liczba obserwacji powinna być większa niż liczba zmiennych. Zbiory danych należy przygotować pod model regresji logistycznej, pamiętaj o uzupełnieniu brakujących wartości oraz usunięciu zmiennych współliniowych.

- Niestandardowe, interesujące zbiory danych zostaną docenione - zestawy danych są interesujące, gdy ponad 50% z nich różni się od używanych na poprzednich przedmiotach (np. *Wstęp do Uczenia Maszynowego*).
- Możesz zamienić wieloklasowe zbiory danych na binarne zbiory danych poprzez łączenie klas.

3 Implementacja algorytmów optymalizacji

Zaimplementuj algorytmy optymalizacji do estymacji parametrów w regresji logistycznej:

1. Gradient Descent
2. Stochastic Gradient Descent (w wersji standardowej, aktualizacja gradientu dla pojedynczej obserwacji)
3. Stochastic Gradient Descent (w wersji mini batch, aktualizacja gradientu dla podzbioru obserwacji np. wielkość batch = 20)

Używanie implementacji dostępnych w Internecie jest niedozwolone.

4 Analiza

Podczas analizy rozważ dodatkową metodę optymalizacji algorytmu – Iterative Reweighted Least Squares (IWLS), możesz użyć gotowej implementacji.

Reguła stopu Zaproponuj regułę zatrzymania dla powyższych algorytmów. Pamiętaj, aby użyć tej samej reguły we wszystkich algorytmach.

Analiza zbieżności Sprawdź jak wartość funkcji log-wiarogodności zależy od liczby iteracji dla 4 powyższych algorytmów. Analizę zbieżności należy przeprowadzić na danych treningowych.

Analiza jakości modeli W celu zbadania jakości modeli posłużymy się miarą zrównoważonej dokładności (*balanced accuracy*). Modele powinny być trenowane na zbiorze uczącym. Miara powinna być obliczana na danych testowych. Należy uśrednić wyniki z co najmniej 5 podziałów trening-test. Jeśli dany algorytm nie osiągnie zbieżności w ciągu 500 iteracji, należy użyć rozwiązania z ostatniej iteracji.

5 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- folder `Kody` zawierający wszystkie potrzebne kody do reprodukcji wyników zadania domowego, w tym implementację algorytmów optymalizacji,
- plik `NUMERINDEKSU_raport.pdf` opisujący wyniki (maksymalnie 5 stron, w tym maksymalnie 3 strony tekstu oraz maksymalnie 2 strony wykresów).

6 Ocena

Łączna liczba punktów do zdobycia jest równa 10, w tym:

Kod (4 punkty)

- jakość kodu (porządek, czytelność) - 1 punkt,
- poprawność algorytmów - 1 punkt
- reprodukowalność wyników - 2 punkty

Raport (6 punktów)

- przetestowanie metod - 3 punkty,
- raport - 3 punkty.

7 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_PD1`. Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 28.03.2024 do godziny 23:59. Tytuł wiadomości: *[ZMUM][PD1] Nazwisko Imię*.