

## MINIPROJEKT 1

Każdy wiersz pliku `dane.data` zawiera osiem liczb rzeczywistych: pierwsze siedem odpowiada wartościom pewnych cech (danym wejściowym), natomiast ostatnia liczba odpowiada danej objaśnianej (danej wyjściowej).

Zaimplementuj (w ulubionym języku programowania) algorytm **regresji liniowej** i wyznacz funkcję jak najlepiej opisującą dane wyjściowe w zależności od siedmiu cech. Skorzystaj zarówno z rozwiązania analitycznego, jak i metod gradientowych.

Do ewaluacji otrzymanych modeli wykorzystaj **kwadratową funkcję straty**. Dodatkowo możesz sprawdzić, jak zmieniają się wyniki przy wyborze innych funkcji strat (np. rozważając bezwzględną funkcję straty lub funkcję Hubera).

Sprawdź, jak zastosowanie **funkcji bazowych** wpływa na wynik. Przetestuj funkcje bazowe takie jak wielomiany (np.  $x_1$ ,  $x_1^2$ ,  $x_1 \cdot x_4$ , ...), funkcje gaussowskie (czyli funkcje postaci  $x \mapsto \exp(-(x - \bar{x})^2/s^2)$  z pewnym parametrem  $s$ ) i ewentualnie inne ciekawe funkcje, które mogą poprawić wynik.

Zastosuj **regularyzacje**  $\ell_1$  i  $\ell_2$  oraz regularyzację z siecią elastyczną i w każdym przypadku wyznacz możliwie najlepszą wartość parametru/parametrów regularyzacji.

Nie zapomnij o tym, aby na początku **przeskalować dane**.

Podziel (w sposób losowy) dane na **zbiór treningowy, walidacyjny i testowy**. Hyperparametry modelu (czyli np. parametry regularyzacji, stopień wielomianów, parametr funkcji gaussowskiej) wyznacz w oparciu o dane ze zbioru walidacyjnego, natomiast ocenę modelu oprzyj na danych ze zbioru testowego.

Dla wybranych modeli stwórz **wykresy** zawierające krzywe uczenia, czyli przedstawiające funkcję błędu zależną od rozmiaru zbioru treningowego. Zaznacz punkty odpowiadające błędom obliczonym na całym zbiorze testowym po zastosowaniu algorytmu na następujących frakcjach zbioru treningowego: 0.01, 0.02, 0.03, 0.125, 0.625, 1. Aby uwiarygodnić wyniki, uśrednij kilka przebiegów algorytmu na losowych wyborach obserwacji do zbioru treningowego, walidacyjnego i testowego.

Napisz **raport** opisujący rozważane modele oraz otrzymane wyniki. W opisie uwzględnij wszystkie interesujące aspekty, takie jak podział danych, sposób ich skalowania, wykorzystaną funkcję błędu, wstępną analizę danych, wykorzystane funkcje bazowe, metody regularyzacji, szybkość uczenia itd. Nie musisz się ograniczać tylko do najlepszych wyników – opisywanie porażek również jest cenne.