

Warsztaty z ucznia Maszynowego semestr letni 2018-2019

Mikołaj Słoń

Instrukcja obsługi programu służącego do regresji liniowej:

Wymagania: System Windows 7 lub nowszy

Program jest aplikacją konsolową i wymaga jednego argumentu, nazwy pliku konfiguracyjnego.

W pliku konfiguracyjnym każda linijka to zestaw argumentów dla których ma zostać wykonany test. Każdy argument jest oddzielony spacją. Na końcu linijki nie powinna znajdować się spacja. W przypadku nieprawidłowej liczby argumentów bądź nieprawidłowego formatu argumentów program może zachowywać się w sposób nieprzewidywalny.

1. Nazwa pliku z danymi. Plik musi być w formacie csv z separatorem ','. Pierwsza linia zawiera nazwy argumentów. Ostatni argument to zawsze wartość funkcji dla podanych wcześniej argumentów. W kolejnych rzędach znajdują się liczby reprezentujące wartości argumentów i spodziewaną wartość funkcji. Wszystkie wartości po pierwszej linii muszą być liczbami w przeciwnym przypadku zachowanie programu może być nie zdefiniowane. Dla liczb rzeczywistych należy używać '.' Aby oddzielić część dziesiętną. Przykładowy plik źródłowy znajduje się w projekcie i nazywa się „input.csv”.

Plik z danymi powinien zostać zamknięty przed uruchomieniem programu inaczej działanie programu może być nie zdefiniowane.

2. Drugim argumentem jest wartość z przedziału [0-1] określająca podział danych na testowe i treningowe. Przykładowo przy parametrze 0.3 dane testowe będą stanowiły 30% zbioru danych.
3. Wybierz czy dane mają zostać znormalizowane do wartości z przedziału [0-1]. Wpisz 'y' jeżeli mają zostać znormalizowane.
4. Wybierz implementację regresji liniowej. Należy wybrać '1', '2' lub '3'. W przypadku wybrania innego znaku program zostanie zamknięty i należy uruchomić go ponownie.
 - 1 regresja liniowa sumaryczna
 - 2 regresja liniowa alternatywna implementacja
 - 3 metoda gradientowa (iteracyjna)

5. Nazwa eksperymentu. Na podstawie tej nazwy będą tworzone pliki z wynikami.

6. Opcjonalny argument w przypadku algorytmu numer 3 należy wybrać ilość iteracji. Ilość iteracji musi być większa niż 0 i być liczbą całkowitą. Przy dużym zbiorze danych i dużej liczbie iteracji wykonanie programu może zająć znaczną ilość czasu.

5. Program wytrenuje model regresji i policzy MSE (Mean Squared Error), SSE(Sum of Squared Errors) oraz MAE (Mean absolute Error) dla zbioru treningowego i wypisze ich wartości. Następnie dla każdego zestawu argumentów z zbioru testowego wyznaczy obliczoną przez model spodziewaną wartość. Jako ostatni wypisze MSE, SSE oraz MAE dla zbioru testowego. Program wygeneruje również pliki wyjściowe. NazwaEksperymentuStatistics.txt oraz NazwaEksperymentuComparision.txt. W pierwszym zostaną wypisane MSE, SSE oraz MAE dla zbioru treningowego. W drugim Będzie wartość predykowana modelem oraz wartość prawdziwa y.