234053
Numer indeksu
Paweł Galewicz
Imię i nazwisko

234067
Numer indeksu
Bartosz Jurczewski
Imię i nazwisko

234102
Numer indeksu
Zbigniew Nowacki
Imię i nazwisko

234106
Numer indeksu
Karol Podlewski
Imię i nazwisko

234128
Numer indeksu
Piotr Wardęcki
Imię i nazwisko

Kierunek Informatyka Stosowana
Stopień II
Specjalizacja Data Science
Semestr 1

Data oddania 18 marca 2020

Metody uczenia maszynowego

Problem set 1

Spis treści

1 Cel

Zadanie polega na analizie procesu klasyfikacji danych za pomocą wybranych metod:

1. Algorytm drzew decyzyjnych
2. Naiwny klasyfikator Bayesa
3. Maszyna wektorów nośnych
4. Klasyfikator k-najbliższych sąsiadów
5. Algorytm sztucznych sieci neuronowych

Należy zaimplementować każdą metodę, a następnie zweryfikować jej działanie biorąc pod uwagę:

- różne możliwe ustawienia parametrów konfiguracyjnych i ich wpływ na wyniki klasyfikacji
- zbiory danych o różnej charakterystyce (przynajmniej 3 różne zbiory)

Każdą metodę należy przetestować na tych samych zbiorach, a następnie porównać wyniki i wyciągnąć wnioski dotyczące skuteczności poszczególnych metod. Jako kryterium porównawcze wystarczy omówić dokładność klasyfikacji (accuracy), pozostałe kryteria są opcjonalne.

2 Wprowadzenie

2.1 Algorytm drzew decyzyjnych

Opis

2.2 Naiwny klasyfikator Bayesa

Opis

2.3 Maszyna wektorów nośnych

Opis

2.4 Klasyfikator k-najbliższych sąsiadów

Algorytm ten należy do grupy algorytmów analizy skupień (wyszukiwanie i wyodrębnianie grup obiektów podobnych do siebie). Algorytm k-średnich polega na przenoszeniu punktów skupień (centroidów) do środków ciężkości podzbiorów punktów. Przebieg algorytmu jest następujący:

1. Określamy liczbę skupień (k)
2. Wybieramy losowo środki skupień (centroidy)
3. Obliczamy odległości wybranych obiektów od środków skupień za pomocą odległości euklidesowej
4. Przypisujemy obiekty do skupień
5. Ustalamy na nowo środki skupień

Kroki od 3 do 5 są wykonywane, aż zostanie spełniony warunek zatrzymania algorytmu. W tym przypadku będzie to przekroczenie narzuconej wcześniej liczby iteracji, lub doprowadzenie skupień do stanu w którym nie będzie dochodziło już do żadnych przesunięć obiektów.

2.5 Algorytm sztucznych sieci neuronowych

Opis

3 Opis implementacji

Algorytmy zostały zaimplementowane za pomocą języka Python w wersji 3.8.2. Wykorzystano w nim biblioteki NumPy, Sklearn i Pandas. Bazowaliśmy na trzech zestawach danych:

- Fall Detection Data from China - `https://www.kaggle.com/pitasr/falldata`
- Rain in Australia - `https://www.kaggle.com/jsphyg/weather-dataset-rattle-package`
- Suicide Rates Overview 1985 to 2016 - `https://www.kaggle.com/russellyates88/suicide-rates-ove`

4 Badania

Cytuję: "Należy zaimplementować każdą metodę, a następnie zweryfikować jej działanie biorąc pod uwagę:
A. różne możliwe ustawienia parametrów konfiguracyjnych i ich wpływ na wyniki klasyfikacji"
B. zbiory danych o różnej charakterystyce (przynajmniej 3 różne zbiory)