

<u>234053</u>
<i>Numer indeksu</i>
<u>Paweł Galewicz</u>
<i>Imię i nazwisko</i>

<u>234067</u>
<i>Numer indeksu</i>
<u>Bartosz Jurczewski</u>
<i>Imię i nazwisko</i>

<u>234102</u>
<i>Numer indeksu</i>
<u>Zbigniew Nowacki</u>
<i>Imię i nazwisko</i>

<u>234106</u>
<i>Numer indeksu</i>
<u>Karol Podlewski</u>
<i>Imię i nazwisko</i>

<u>234128</u>
<i>Numer indeksu</i>
<u>Piotr Wardecki</u>
<i>Imię i nazwisko</i>

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	3
2	Opis implementacji	3
3	Klasyfikatory	3
3.1	Algorytm drzew decyzyjnych	3
3.2	Naiwny klasyfikator Bayesa	3
3.3	Maszyna wektorów nośnych	3
3.4	Klasyfikator k-najbliższych sąsiadów	4
3.5	Algorytm sztucznych sieci neuronowych	4
4	Badania	4
4.1	Fall Detection Data from China	5
4.2	Rain in Australia	5
4.3	Suicide Rates Overview 1985 to 2016	5
5	Wnioski	5

1 Cel

Zadanie polegało 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

Należało 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 należało przetestować na tych samych zbiorach, a następnie porównać wyniki i wyciągnąć wnioski dotyczące skuteczności poszczególnych metod. Jako kryterium porównawcze wykorzystaliśmy **dokładność klasyfikacji (accuracy) oraz ...**.

2 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](#)
- [Rain in Australia](#)
- [Suicide Rates Overview 1985 to 2016](#)

3 Klasyfikatory

3.1 Algorytm drzew decyzyjnych

Opis

3.2 Naiwny klasyfikator Bayesa

Opis

3.3 Maszyna wektorów nośnych

Maszyna wektorów nośnych jest klasyfikatorem liniowym. Algorytm polega na rozdzieleniu obiektów o różnej przynależności klasowej za pomocą hiperpłaszczyzn, które mają być od siebie możliwe jak najbardziej oddalone - taką odległość nazywa się marginesem klasyfikatora, a hiperpłaszczyznę z największym marginesem wektorami nośnymi.

Algorytm bardzo dobrze sobie radzi z danymi liniowo separowanymi, ale nie zawsze będzie istniała hiperpłaszczyzna rozdzielająca, która zapewni poprawną klasyfikację wszystkich elementów zbioru. W takich przypadkach maszyna wektorów nośnych za pomocą funkcji jądrowych transformuje przestrzeń do postaci liniowo separowanej.

3.4 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.

3.5 Algorytm sztucznych sieci neuronowych

Sztuczna sieć neuronowa jest połączeniem wielu elementów nazywanych sztucznymi neuronami, które tworzą co najmniej trzy warstwy: wejściową, ukrytą oraz wyjściową. Neurony przetwarzają informacje dzięki nadaniu im parametrów które nazywane są wagami. Podstawą tworzenia sieci neuronowej jest modyfikowanie współczynnika wagowego połączeń w celu uzyskania poprawnych wyników.

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)

4.1 Fall Detection Data from China

5% zbioru treningowego

10% zbioru treningowego

25% zbioru treningowego

50% zbioru treningowego

4.2 Rain in Australia

5% zbioru treningowego

10% zbioru treningowego

25% zbioru treningowego

50% zbioru treningowego

4.3 Suicide Rates Overview 1985 to 2016

5% zbioru treningowego

10% zbioru treningowego

25% zbioru treningowego

50% zbioru treningowego

5 Wnioski