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 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:

- rożne 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

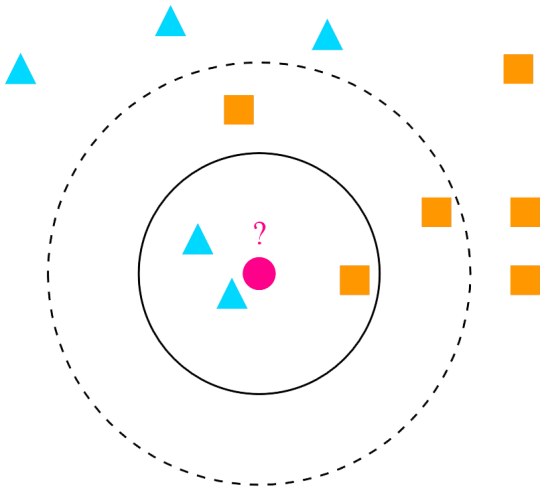
Maszyna wektorów nośnych jest klasyfikatorem liniowym. Algorytm polega na rozdzieleniu obiektów o różnej przynależności klasowej za pomocą hiperpłaszczyzn, które mają być od siebie możliwe jak najbardziej oddalone - taką odległość nazywa się marginesem klasyfikatora, a hiperpłaszczyzny z największym 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 k najbliższych sąsiadów jest klasyfikatorem (ściślej algorytmem regresji regresji nieparametrycznej). Algorytm ten zakłada dany zbiór uczący, w którym znajdują się już sklasyfikowane dane. Schemat składa się z szukania k obiektów najbliższych do obiektu klasyfikowanego. Następnie, przyporządkowuje się nowy obiekt do najczęściej występującej klasy w obrębie jego k-najbliższych sąsiadów.

Rysunek 1 pokazuje działanie algorytmu. W przypadku k=3 (mniejszy okrąg), różowa kropka zostanie zakwalifikowana do niebieskich trójkątów. W przypadku k=5 (większy okrąg) - do pomarańczowych kwadratów.



Rysunek 1:

3.5 Algorytm sztucznych sieci neuronowych

Sztuczna siec neuronowa jest połączeniem wielu elementow nazywanych sztucznymi neuronami, które tworzą conajmniej 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