Sprawozdanie

Indukcyjne Metody Analizy Danych

Ćwiczenie 4. Algorytm klasyfikacji k-najbliższych sąsiadów

Autor: Paweł Mielniczuk

**Spis treści:**

1. **Opis działania algorytmu KNN**
2. **Wprowadzenie do zbiorów danych**
3. **Opis implementacji**
4. **Analiza wyników zbiorów:**
   1. **Iris**
   2. **Pima diabetes**
   3. **Glass**
   4. **Wine**
   5. **Seeds**
5. **Podsumowanie**
6. **Opis działania algorytmu KNN**

**Knn –** Knn jestmodelem o działaniu instancyjnym/leniwym. Oznacza to, że wszystkie obliczenia zostają dokonane dopiero podczas klasyfikacji. KNN, czyli k nearest neighbors, może zostać użyty zarówno do regresji jak i klasyfikacji. Pseudo algorytm może być opisany jakos:

Kroki wstępne:

* 1. Załadowanie zbioru danych
  2. Przypisanie wartości k

Kroki predykcji

Dla każdej instancji ze zbioru treningowego:

1. Oblicz dystans (np. Euklidesowy) pomiędzy nową instancją, a wszystkimi ze zbioru
2. Posortuj dystanse od najmniejszej do największej
3. Wybierz k pierwszych wyników
4. Wybierz najczęściej występującą klasę z danych odległości (głosowanie)
5. Zwróć wynik głosowania

Jednym z problemów algorytmu KNN jest odpowiednie dobranie parametru k. Jeżeli dany parametr będzie zbyt niski pojawi się overfitting. Natomiast w przypadku zbyt dużego k granice klas stają się bardziej gładkie a błąd predykcji, ponieważ nie mamy zbyt dokładnych granic.

1. **Wprowadzenie do zbiorów danych**

Podczas analizy i implementacji użyte zostały cztery zbiory danych. Zbiory podzielone są na dwie części. Pierwszą z nich są cechy, dokładnie wektor, cech oraz etykiety mówiące o przynależności wektora cech do konkretnej klasy.

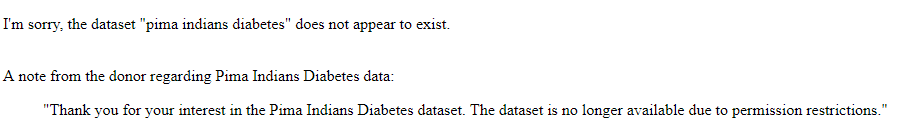
Wszystkie zbiory dostępne są do pobrania ze strony *https://archive.ics.uci.edu/ml/datasets.html*

Zbiory danych zostały ściągnięte i załadowane przy użyciu biblioteki *pandas* lub bezpośrednio załadowane za pomocą biblioteki *scikit-learn*.

Zbiory danych:

* Iris data set
* Wine data set
* Glass identification data set
* Pima diabetes data set
* Seeds dataset

Ciekawostką jest, że w trakcie badania klasyfikatora i tworzenia sprawozdania ostatni ze zbiorów *Pima diabetes* został usunięty ze strony UCI ze przez ograniczenie uprawnień do udostępniania danego zbioru.



Rysunek 1 Wiadomość ze strony https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes mówiąca o braku dalszego dostępu do danego zbioru.

Poniżej zaprezentowano opis zbiorów. Opis ten pomoże w zrozumieniu danych, które będą analizowane. Dobre zrozumienie danych z którymi się pracuje jest niezbędną częścią do poprawnego przeprowadzenia badań.

**Zbiór Iris**

Jest to prawdopodobnie jeden z najbardziej znanych i podstawowych zbiorów danych przy problemach klasyfikacji i rozpoznawania wzorców.

Zbiór składa się ze 150 instancji, podzielonych na 3 równe zbiory po 50 klas każda.

Definicje atrybutów:

* Sepal – zielony płatek u dołu kielicha służący do ochrony kwiatu w trakcie kwitnięcia,
* Petal – płatek kwiatu, służący do przyciągania uwagi ptaków i insektów

Cechy zbioru zawierają cztery informacje:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

Ostatnią, piątą kolumną jest klasa mówiąca o typie irysa. Możliwe są trzy klasy:

1. Iris Setosa
2. Iris Versicolour
3. Iris Virginica

**Zbiór Wine**

Zbiór ten został skonstruowany w wyniku analizy składu chemicznego win stworzonych w tym samym rejonie Włoch lecz przy użyciu trzech różnych odmian uprawnych.

Zbiór składa się ze 178 instancji.

Definicje atrybutów oraz cechy zbioru:

1. Alcohol – alkohol
2. Malic acid – kwas jabłkowy
3. Ash – popiół
4. Alkalinity of ash – alkaliczność popiołu
5. Magnesium – magnez
6. Total phenols – całkowita zawartość fenoli
7. Flavonoids – flawonoidy
8. Nonflavanoid phenols – fenole nieflawonowe
9. Proanthocyanidins – proantocyjanidyny
10. Color intensity, intensywność koloru
11. Hue – odcień
12. OD280/OD315 of diluted wines - OD280 / OD315 rozcieńczonych win
13. Proline – Proline

Pierwszy atrybut w pliku zawierającym dane jest identyfikatorem klasy od 1 do 3.

Rozłożenie instancji klas jest następujące:

* Klasa 1 – 59 instancji,
* Klasa 2 – 71 instancji,
* Klasa 3 – 48 instancji.

**Zbiór Glass identification**

Zbiór powstał poprzez analizę składu chemicznego badanego szkła aby określić typ powstałego szkła oraz jego przeznaczenie.

Zbiór składa się z 214 instancji podzielonych na 6 klas.

Rozłożenie instancji klas jest następujące:

* Klasa 1 – 70 instancji,
* Klasa 2 – 76 instancji,
* Klasa 3 – 17 instancji,
* Klasa 4 - 13,
* Klasa 5 - 9,
* Klasa 6 - 29.

Definicje atrybutów oraz cechy zbioru:

1. Id – numer porządkowy
2. Refractive index – współczynnik załamania światła
3. Sodium – sód
4. Magnesium – magnez
5. Aluminium – glin
6. Silicon – krzem
7. Potassium – potas
8. Calcium – wapń
9. Barium – bar
10. Iron – żelazo

**Zbiór Seeds**

Zbiór reprezentuje atrybuty 3 różnych typów zbóż.

Zbiór składa się z 210 instancji podzielonych na 3 klasy.

Rozłożenie instancji klas jest następujące:

* Klasa 1 (Kama) – 70 instancji,
* Klasa 2 (Rosa) – 70 instancji,
* Klasa 3 (Canadian) – 70 instancji,

Definicje atrybutów oraz cechy zbioru. Wszystkie atrybuty są miarami nasion zboża:

1. Area – pole
2. Perimeter – obwód
3. Compactness – ścisłość
4. Length of kernel – długość nasiona
5. Width of kernel – szerokość nasiona
6. Asymmetry coefficient – współczynnik asymetrii
7. Length of kernel groove - długość rowka nasiona

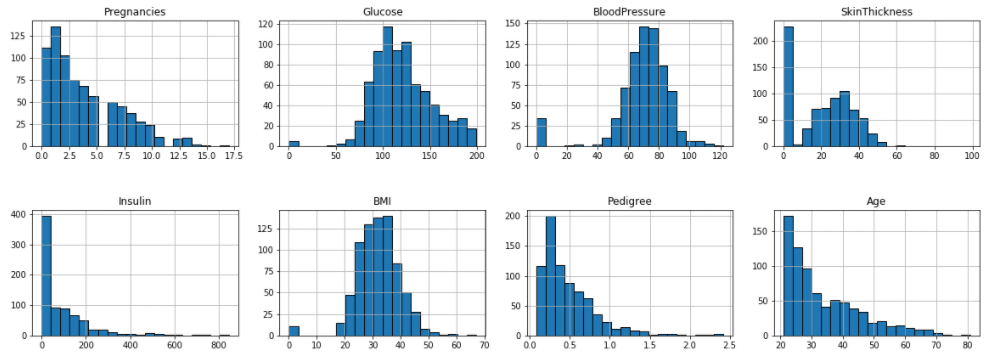
**Zbiór Pima diabetes**

Celem zbioru jest umożliwienie zdiagnozowania czy dany pacjent ma cukrzycę, bazując na diagnostykach zamieszczonych w cechach zbioru. Wszyscy pacjenci przebadani byli kobietami mającymi przynajmniej 21 lat oraz byli pochodzenia indiańskiego plemienia Pima.

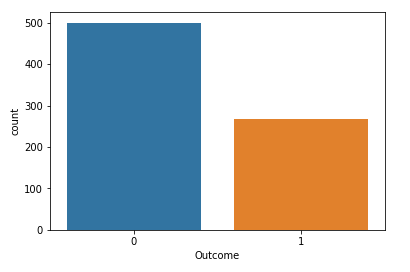
Zbiór składa się z 768 instancji posiadających dwie możliwe klasy 1 – oznaczające że zbadana osoba jest chora na cukrzycę, 0 – oznaczające że dana osoba nie jest chora na cukrzycę.

Definicje atrybutów oraz cechy zbioru:

1. Pregnancies – liczba ciąży
2. Glucose – poziom glukozy
3. Blood ressure – ciśnienie krwi
4. Skin thickness – grubość skóry
5. Insulin – poziom insuliny
6. BMI – body mass index
7. Diabetes pedigree function – funkcja pedigree
8. Age – wiek



Rysunek 2 Histogramy atrybutów danych zbioru Pima diabetes



Rysunek 3 Rozkład klas

1. **Implementacja**

Do implementacji użyty został język programowania Python wraz z bibliotekami takimi jak: pandas, sklearn, numpy, matplotlib

**Parametry badane w knn**

**K (number of votes) –** Podstawowy parametr, który odpowiada za liczbę sąsiednich punktów, które biorą udział w głosowaniu.

**Weights –** Parametr używany podczas głosowania, przyjmuje 2 wartości: ‘uniform’ i ‘distance’. W pierwszej wersji podczas głosowania wszystkie głosy są tak samo ważne i na żadne z nich nie są nakładane wagi. W przypadku ‘distance’ głosy są ważone z wagą równą odwrotności ich odległości do badanego punktu. Oznacza to, że głosy instancji, które są bliżej nowego punktu są ważniejsze podczas klasyfikacji.

**Metric –** Ten parametr mówi o tym w jaki sposób obliczane są odległości do badanego punktu. Zbadane zostaną 3 miary mierzenia odległości:

1. **Euclidean** – pierwiastek z sumy kwadratów różnic między współrzędnymi, linia prosta pomiędzy dwoma punktami,
2. **Manhattan** – suma bezwzględnych różnic współrzędnych punktów,
3. **Chebyshev** – maksimum bezwzględnych różnic między współrzędnymi punktów.

**Parametry analizy k-means i k-medoids**

**Accuracy –** uśredniony błąd zbioru

**F1 Score –** 2 \* precision \* recall / (precision + recall)

**Dobór optymalnego parametru k –** do wyboru optymalnego parametru liczby klastrów została użyta metoda łokcia “elbow method”. Mówi ona o o tym aby wypróbować kilka różnych k i poszukać miejsca na grafie, w którym przestajemy uzyskiwać duże zyski (w tym przypadku zmniejszamy sumę odległości) informacji i większa liczba klastrów przestaje dawać duże zmiany. To miejsce najczęściej jest zakrzywieniem w grafie, które przypomina łokieć.

1. **Analiza wyników zbiorów danych**
   1. **Analiza wyników zbioru Iris dataset**
   2. **Analiza wyników zbioru Pima Diabetes dataset**
   3. **Analiza wyników zbioru Glass dataset**
   4. **Analiza wyników zbioru Wine dataset**
   5. **Analiza wyników zbioru Seeds dataset**
2. **Podsumowanie**