Podstawy uczenia maszynowego

11.04.2024

Laboratorium 5

Klasteryzacja

Łukasz Stępień, Kacper Fus

1. Cel zadania

Celem zadania było zapoznanie się z trzema popularnymi algorytmami klasteryzacji: k-means, DBSCAN i klasteryzacją spektralną oraz zbadanie ich skuteczności na różnych zbiorach danych.

1. Implementacja

Wykorzystano syntetycznie generowane zbiory „Blobs”, „Circles”, „Moons” oraz „Ellipses”. Następnie zbadano jak algorytmy k-means, DBSCAN oraz klasteryzacja spektralna radzą sobie z tymi zbiorami, ze względu na różne parametry. Następnie wykorzystano zbiór banknotes, który zawiera informacje o cechach banknotów oraz ich statusie jako oryginalne lub sfałszowane. Zadanie polegało na zastosowaniu algorytmów k-means i DBSCAN do pogrupowania banknotów w klastry.

2.1 Wizualizacja zborów

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, krąg, diagram

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, tekst, diagram, Wykres

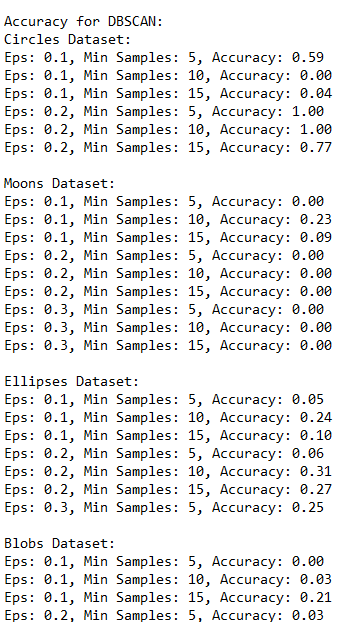
Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, linia, tekst, diagram

Opis wygenerowany automatycznie

2.2. Wyniki testów

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie



Obraz zawierający tekst, Czcionka, czarne i białe, dokument

Opis wygenerowany automatycznie

* 1. Obraz zawierający zrzut ekranu, krąg, tekst, Wielobarwność

     Opis wygenerowany automatycznieWpływ parametrów na kształt klastrów

Obraz zawierający zrzut ekranu, krąg, tekst, Wielobarwność

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, krąg, tekst, diagram

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, krąg, Wielobarwność, tekst

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, krąg, tekst, diagram

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, tekst, krąg, diagram

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, Wielobarwność

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, Wielobarwność

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, Wielobarwność

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, Wielobarwność

Opis wygenerowany automatycznie

* 1. Wnioski do Zadania 1
* K-means
  + Najlepiej sobie poradził ze zbiorem danych Circles oraz Moons. Sugeruje to, że lepiej on radzi sobie ze zbiorami eliptycznymi bądź okrągłymi ,które na dodatek są oddzielone od siebie .
  + Na zaprezentowanych wykresach, widać, że parametr zwiększą liczbę klastrów, na które dzielimy zbiór danych. Każdy ze zbiorów danych, jest równomierny oraz zauważamy granicę między nimi.
* DBSCA
  + Najlepiej poradził sobie ponownie z Circles, prawdopodobnie dlatego, że reprezentują one zbiór o różnej gęstości,
  + Wpływ tych dwóch parametrów na klastry w DBSCAN polega na decydowaniu, jakie punkty będą łączone w klastry na podstawie odległości i gęstości. Ich umiejętne dobranie sprawi, że równomierne kształty bez kolizyjnych szumów.
* Klasteryzacja spektralna
  + Poradziłą sobie najlepiej z danymi Blobs co może świadczyć o tym, że Klasteryzacja spektralna jest skuteczna w identyfikowaniu klastrów w zbiorach danych, które nie są liniowo separowalne.
  + Zwiększanie liczby klastrów, zwiększa ilość podzbiorów na jakie dane zostały podzielone, a gamma sprawią, że zbiory stają się mniej równomierne.
  1. Wyniki doświadczeń do Zadania 2

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

* 1. Wnioski do Zadania 2
* DBSCAN nie przypisał wszystkich banknotów do klastrów. To sugeruje, że niektóre banknoty mogą zostać uznane za punkty odstające lub nieprzypisane do żadnego klastra.
* Homogeniczność mierzy: wartość 0.0 oznacza, że klastry nie są homogeniczne, co sugeruje, że punkty z różnych klas mogą być przypisane do tych samych klastrów.
* Zupełność mierzy: wartość 1.0 oznacza, że wszystkie punkty oryginalnie należące do jednej klasy zostały przypisane do jednego klastra, ale jakość tych klastrów jest nadal niska.
* V-miara: w tym przypadku, V-miara wynosząca 0.0 sugeruje, że klasteryzacja nie jest zgodna z rzeczywistym podziałem na banknoty oryginalne i sfałszowane.
* K-means, ma wysoką skuteczność w przypisywaniu klastrów, co świadczy, że struktura zbioru była dobrze odzwierciedlona przez algorytm.
* K-means:
  + K-means dobrze radzi sobie z danymi, które mają w miarę jednorodne i wypukłe klastry o podobnej wielkości.
  + Głównymi parametrami k-means są liczba klastrów (k). Wybór odpowiedniej liczby klastrów może być trudny, a niepoprawne ustawienie może prowadzić do złego podziału danych. Wybór punktów startowych może również wpływać na wynik, ale zwykle algorytm jest odporny na różne punkty startowe.
* DBSCAN
  + DBSCAN jest bardziej elastyczny i może radzić sobie z danymi o różnych kształtach klastrów oraz gęstości.
  + Głównymi parametrami DBSCAN są promień epsilon (eps) określający odległość, w której punkty są uznawane za sąsiadów, oraz minimalna liczba punktów (MinPts), które muszą znajdować się w otoczeniu punktu, aby został uznany za centralny. Wybór odpowiednich wartości eps i MinPts może być kluczowy. Zbyt małe eps mogą powodować, że wiele punktów zostanie uznanych za szum, podczas gdy zbyt duże eps mogą prowadzić do połączenia klastrów.
* Klasteryzacja spektralna:
* Klasteryzacja spektralna może radzić sobie z danymi, które mają nieliniowe kształty klastrów oraz z danymi wysokowymiarowymi.
  + Klasteryzacja spektralna opiera się na wyznaczeniu macierzy podobieństwa lub odległości między punktami, a następnie wyznaczeniu kierunków (wektorów własnych) w tej przestrzeni. Głównym parametrem jest liczba kierunków (k), która określa liczbę klastrów. Wybór odpowiedniej liczby kierunków może być trudny, a niepoprawny wybór może prowadzić do złego podziału danych.