Miniprojekt 4

Udostępnionych jest 10 plików z danymi.

1. Plik `rp.data` to plik znany już z drugiego miniprojektu. Zawiera on dane o 9 cechach oraz przypisaną im klasę (2 lub 4).
2. Plik `dane_18D.txt` zawiera dane o 18 cechach, jednak bez przypisanych klas.
3. Pliki `dane_2D_`n`.txt` ($n = 1, \ldots, 8$) zawiera współrzędne punktów na płaszczyźnie oraz przypisaną im klasę. Liczba klas jest różna w różnych plikach.

Dla każdego z powyższych zbiorów danych porównaj działanie następujących trzech algorytmów klasteryzacji:

- algorytm k-średnich (*k-means*) lub k-średnich++ (*k-means++*),
- grupowanie hierarchiczne (*hierarchical clustering*),
- klasteryzacja spektralna (*spectral clustering*).

Jeśli uznasz to za pomocne, zawsze możesz korzystać z funkcji jądrowych.

W każdym przypadku poza danymi z pliku `rp.data` uzasadnij decyzję o wybranej liczbie klastrów. W przypadku danych z plików `dane_2D_`n`.txt` sprawdź, czy zwracasz liczbę klastrów równą liczbie zadanych klas. Znając liczbę klastrów, sprawdź jakość każdego algorytmu klasteryzacji obliczając błąd klasyfikacji.

Zwizualizuj (np. używając różnych kolorów) klastry po zastosowaniu algorytmów dla danych z plików `dane_2D_`n`.txt`.

Na podstawie uzyskanych wyników wyciągnij wnioski: który algorytm Twoim zdaniem sprawdził się najlepiej w każdym z przypadków?

Do miniprojektu dołączone są dwa artykuły dotyczące klasteryzacji spektralnej oraz algorytmu k-means++:

- David Arthur, Sergei Vassilvitskii
  k-means++: The Advantages of Careful Seeding,
- Andrew Ng, Michael Jordan, Yair Weiss
  On Spectral Clustering: Analysis and an algorithm.