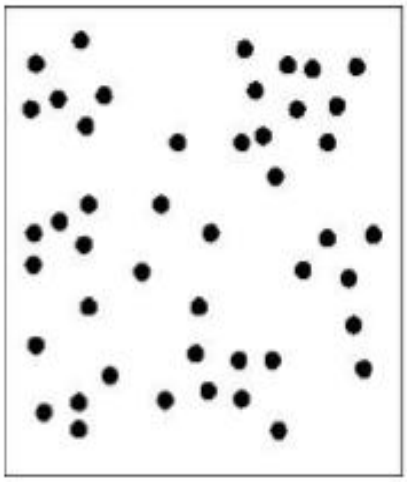
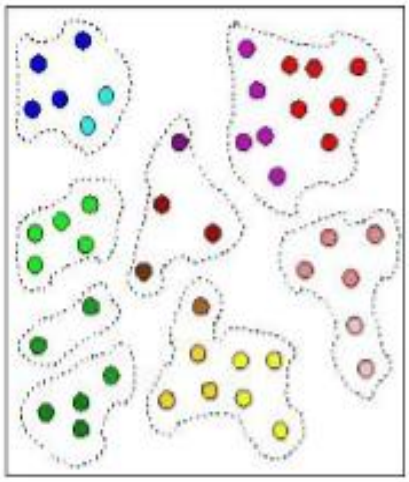
Laboratorium 3+4

Grupowanie

Wstęp

Tematem laboratoriów będzie grupowanie (klasteryzacja). W klasteryzacji nie znamy klasy. Algorytmy grupujące same starają się wyodrębnić klasy we wszystkich rekordach starając się je pogrupować, poszukać „skupisk na wykresie". Rekordy, które są do siebie podobne (leżą blisko siebie) często wpadają do jednego klastra.





Jedną z metod grupowania jest algorytm k-średnich, który omówiony był na wykładzie. Przypomnijmy jego działanie.

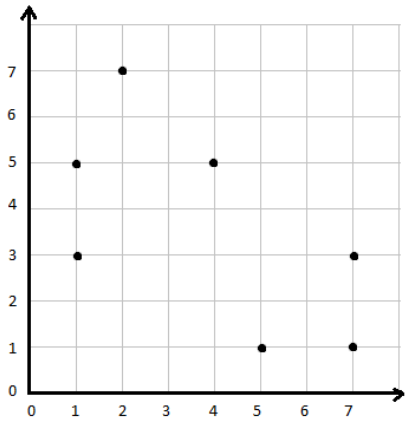
Algorytm k-średnich:

1. Ustal wartość k (liczbę grup/klastrów).
2. Losowo ustal k początkowych środków grup (centroidy).
3. Dla każdego rekordu danych znajdź najbliższy centroid (nowy środek grupy)- w ten sposób wszystkie rekordy zostaną przydzielone do k grup (klastrów).
4. Dla każdej z grup znajdź centroid (średnia z rekordów w klastrze) i uaktualnij położenie środka grupy jako nowa wartość centroidu.
5. Powtarzaj kroki 3-5 dopóki są zmiany lub nie osiągniemy maks. liczby iteracji.

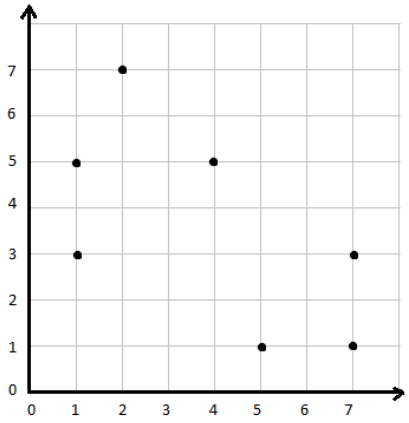
Zadanie 1

Przeprowadź symulację algorytmu grupowania za pomocą k-średnich na małej siedmiorekordowej bazie danych zaznaczonej na wykresie. Początkowe centroidy zostały ustalone w punktach (2,2) i (7,5). W każdej iteracji algorytmu podaj obliczone centroidy i zaznacz na wykresie podział na klastry. Podaj obliczenia na kolejne centroidy, zawrzyj obliczenia na odległości jeśli przynależności nie są oczywiste na podstawie rysunku.

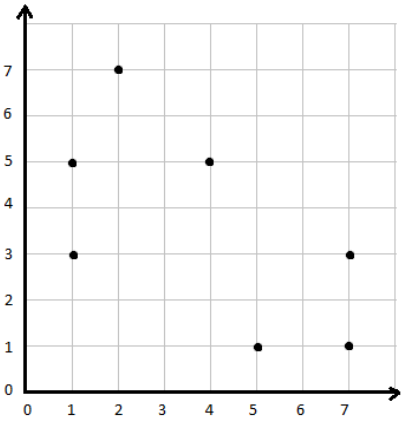
Iteracja 1:
Centroid1 = (2,2)
Centroid2 = (7,5)



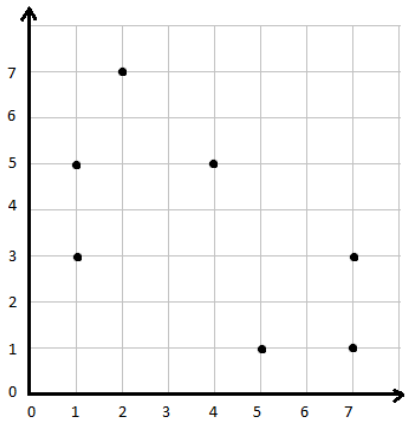
Iteracja 2:
Centroid1 = ...
Centroid2 = ...



Iteracja 3:
Centroid1 = ...
Centroid2 = ...



Iteracja 4:
Centroid1 = ...
Centroid2 = ...



Zadanie 2

Ściągnij bazę danych iris2D.csv (jest to 150 irysów skompresowanych przez PCA do dwóch wymiarów) i przetestuj na niej grupowanie metodą k-średnich, przyjmując k=3. Zwizualizuj klastry na wykresie punktowym. Każdy irys to punkt, a różne klastry mają różne kolory.

Zadanie 3

Podobnie jak w zadaniu drugim, przetestuj jak działa algorytm grupowania opartego na gęstości: DBSCAN. Poeksperymentuj z doborem parametrów algorytmu. Powstałe klastry przedstaw na wykresie punktowym. Wyjaśnij w kilku zdaniach ogólną ideę działania algorytmu.

Zadanie 4

Jak można weryfikować poprawność algorytmu grupującego? Zajrzyj pod link https://en.wikipedia.org/wiki/Cluster_analysis i rozdział Evaluation (External Evaluation).

Dla algorytmów grupujących z zadania 2 i 3 dokonaj oceny obliczając:

a) Czystość (purity) klastrów. Do tego celu trzeba będzie wykorzystać tak zwaną confusion matrix:
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cluster.contingency_matrix.html

b) Zgodność klastrów z prawdziwymi klasami (Rand Index).
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rand_score.html#sklearn.metrics.rand_score

Do obliczeń potrzebny jest oryginalny plik z danymi zawierający etykiety klas. Jest to plik danych, który znajdziemy w materiałach (plik irisORG.csv).

Przesłanie rozwiązania

Rozwiązanie powinno być w formie małego sprawozdania w **pliku pdf / html / doc** i powinno obejmować zadania 2,3,4. Można je zrobić ręcznie lub Jupyterem, lub inną techniką. Nazwa pliku powinna być: Raport2XXXXXX.YYY gdzie XXXX nazwisko, a YYY rozszerzenie pliku. Sprawozdanie powinno zawierać wstawki kodu (komendy), ich wyniki, twoje komentarze i ewentualnie inne elementy np. twoje obliczenia. Nie powinno przekraczać 1 MB pamięci. Termin nadsyłania raportów to 15 maja 2025 roku.

Na wyższą liczbę punktów mają wpływ następujące rzeczy:

• Dobra struktura sprawozdania (wprowadzenie, eksperymenty, objaśnienia, interpretacje).
• Szczegółowość i dokładność eksperymentów, dodawanie komentarzy i własnych interpretacji wyników
• Dobre przygotowanie bazy danych do eksperymentów
• Przejrzystość i estetyka prezentacji
• Nadprogramowe analizy mile widziane…