Uczenie nienadzorowane

Sylwester Piątek

Wydział Matematyki, Politechnika Wrocławska

Wrocław, 23 kwietnia 2025

Przypomnienie

- W uczeniu nadzorowanym (ang. *supervised learning*) korzystamy z danych, które mają etykiety, tzn. znamy wartości zmiennych objaśniających (X) oraz objaśnianych (Y).
- Uczenia nienadzorowanego (ang. *unsupervised learning*) używamy, gdy nie mamy etykiet, tzn. znamy tylko wartości zmiennych objaśniających (X).
- Uczenia nadzorowanego używamy, gdy chcemy rozwiązać problem regresji lub klasyfikacji.
- Uczenia nienadzorowanego używamy, gdy chcemy znaleźć wzorce w danych lub zredukować wymiarowość danych.

Przykłady problemów

Jakie problemy można rozwiązywać z wykorzystaniem uczenia nienadzorowanego?

- segmentacja klientów,
- analiza koszyka zakupowego, rekomendacje produktów,
- analiza obrazów, analiza tekstu itp.,
- wykrywanie anomalii, wykrywanie błędów w danych,
- zmniejszenie liczby cech w danych.

Rodzaje zagadnień

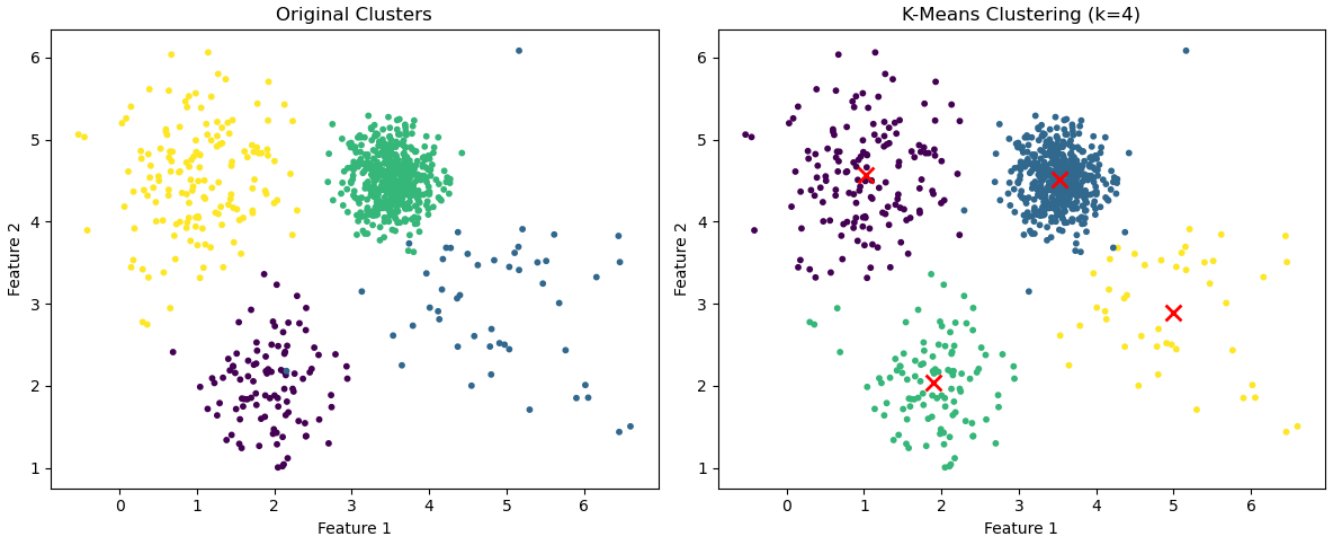
Jakie zagadnienia są przykładami uczenia nienadzorowanego?

- metody asocjacyjne (ang. *association rules*),
- analiza skupień (ang. *cluster analysis*),
- redukcja wymiaru,
- i inne...

Analiza skupień

Analiza skupień (inaczej klastrowanie lub grupowanie) jest to metoda grupowania elementów we względnie jednorodne klasy. Podobieństwo między obserwacjami mierzone jest za pomocą pewnej metryki (np. odległości euklidesowej).

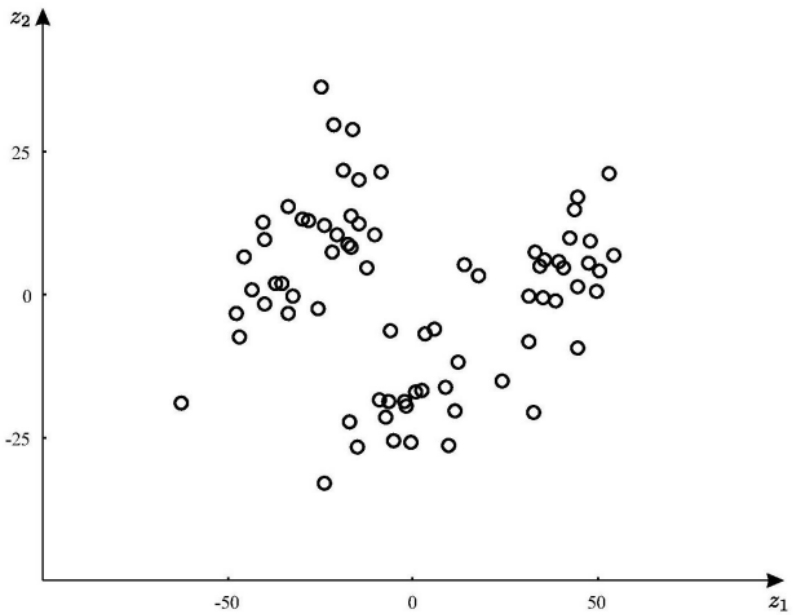
Przykład grupowania danych w skupienia



Rysunek 1: Prawy panel: przykład grupowania danych w skupienia za pomocą algorytmu k-means. Lewy panel: prawdziwy podział punktów na grupy.

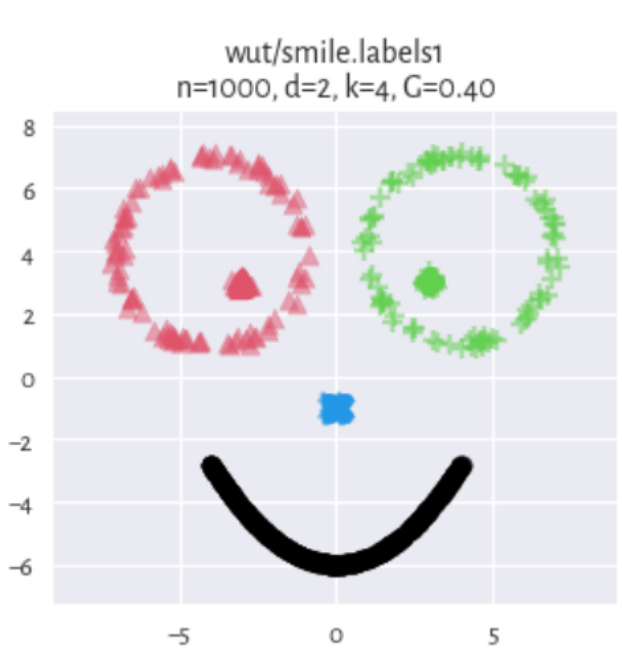
A ile skupień mamy w tym przypadku?

Ile skupień widzimy na poniższym wykresie?

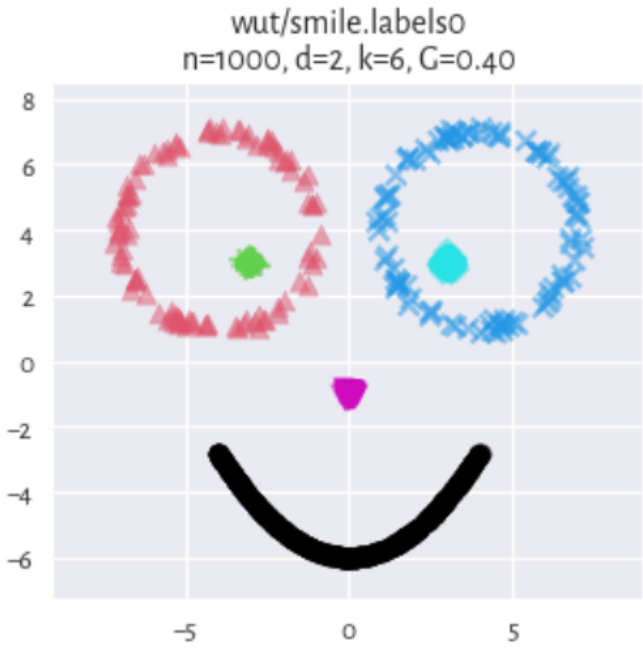


Rysunek 2: Wykres rozrzutu danych opisujących chrząszcze skaczące. Źródło wykresu: [1], źródło danych: [2]

Przykład wygenerowanych danych



(a) Grupowanie w 4 skupienia



(b) Grupowanie w 6 skupień

Rysunek 3: Ilustracja grupowania zbioru danych *smile*. Źródło: `https://clustering-benchmarks.gagolewski.com/weave/data-v1.html#wut/smile`

Jak działa algorytm k-means?

1. Wybieramy liczbę klastrów k oraz losowo wybieramy k punktów jako środki klastrów.
2. Przypisujemy każdy punkt do najbliższego klastra na podstawie odległości (np. euklidesowej) od środka klastra.
3. Obliczamy nowe środki klastrów jako średnie punkty przypisanych do nich punktów.
4. Powtarzamy kroki 2. i 3., aż środki klastrów przestaną się zmieniać lub osiągniemy maksymalną liczbę iteracji.
5. Zwracamy etykiety skupień przypisane do punktów.

Animacja

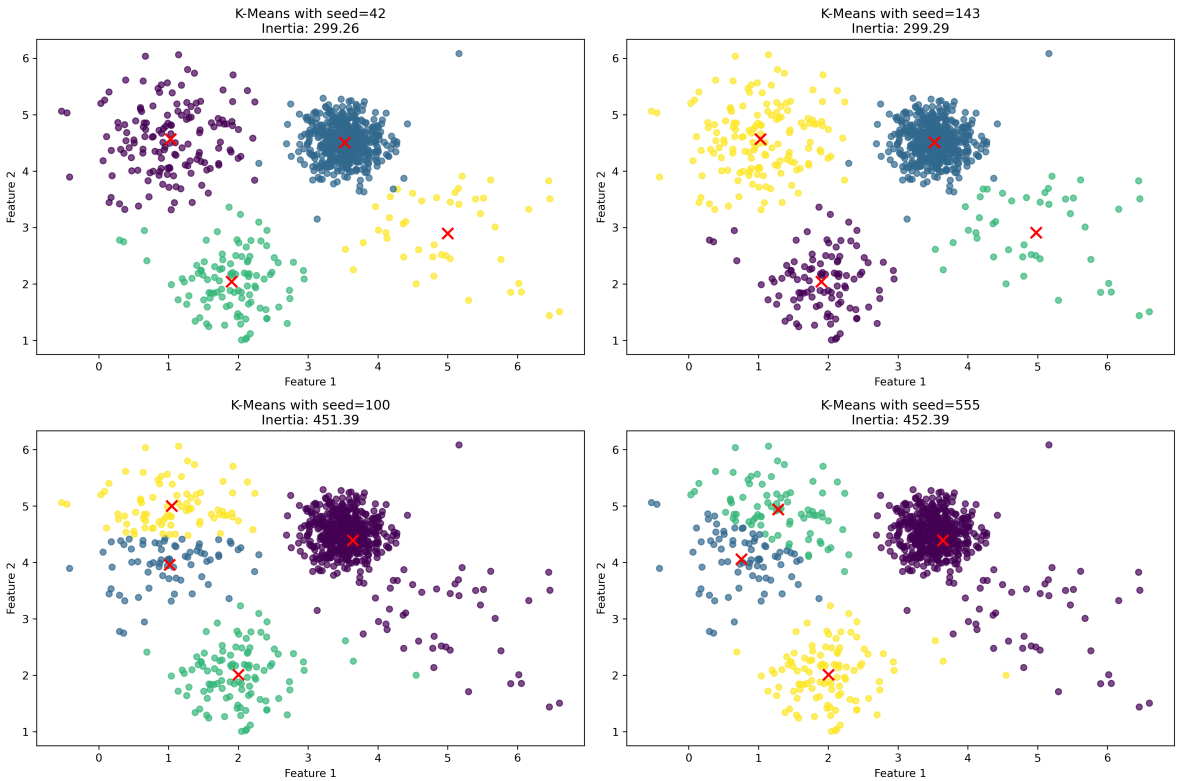
Tu miał być gif :(

Zależność od ziarna

Uwaga 1

Ustawiając różne ziarno losowości, możemy uzyskać różne podziały zbioru danych na skupienia. Algorytm jest więc niedeterministyczny.

Zależność od ziarna — ilustracja

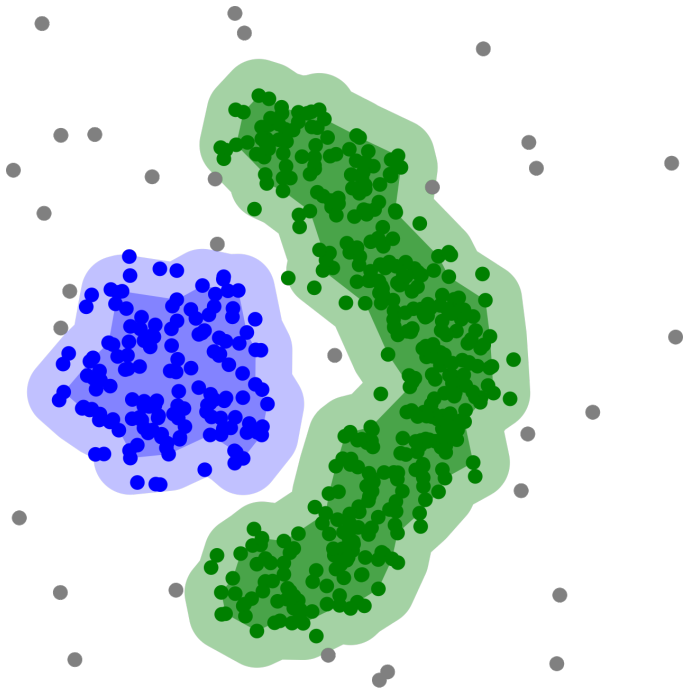


Rysunek 4: Cztery różne podziały zbioru danych na skupienia uzyskane z wykorzystaniem algorytmu k-means z różnym seedem

DBSCAN

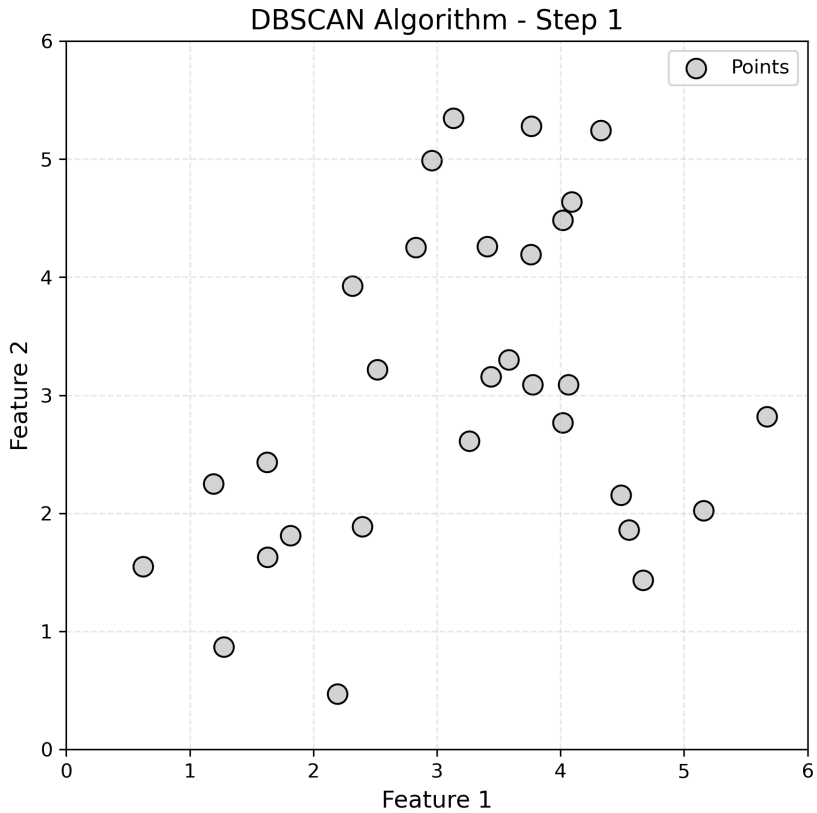
DBSCAN (od ang. *Density-Based Spatial Clustering of Applications with Noise*) — algorytm grupowania danych oparty na gęstości. Nie wymaga wcześniejszej znajomości liczby klastrów, a jego celem jest znalezienie gęsto zaludnionych obszarów w zbiorze danych.
DBSCAN jest odporny na szum i może znaleźć klastry o dowolnym kształcie.
Algorytm przyjmuje dwa parametry: epsilon (ε) i minimalną liczbę punktów (MinPts). Epsilon to maksymalna odległość między punktami, aby mogły być uznawane za sąsiadujące. Minimalna liczba punktów to minimalna liczba punktów, które muszą znajdować się w sąsiedztwie punktu, aby mógł on być uznany za rdzeń klastra.

DBSCAN — przykład



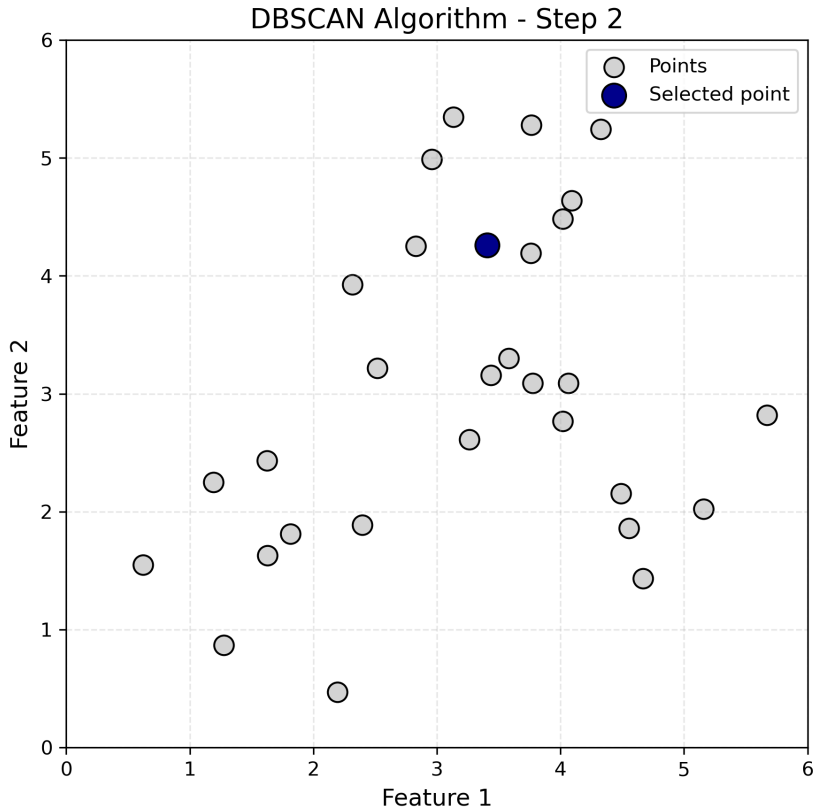
Rysunek 5: Przykład zastosowania algorytmu DBSCAN do grupowania danych. Obszar ciemnoniebieski i ciemnozielony stanowi rdzeń klastra, tzn. zawiera punkty centralne klastra. Źródło: `https://pl.wikipedia.org/wiki/DBSCAN#/media/Plik:DBSCAN-density-data.svg`

DBSCAN — krok 1



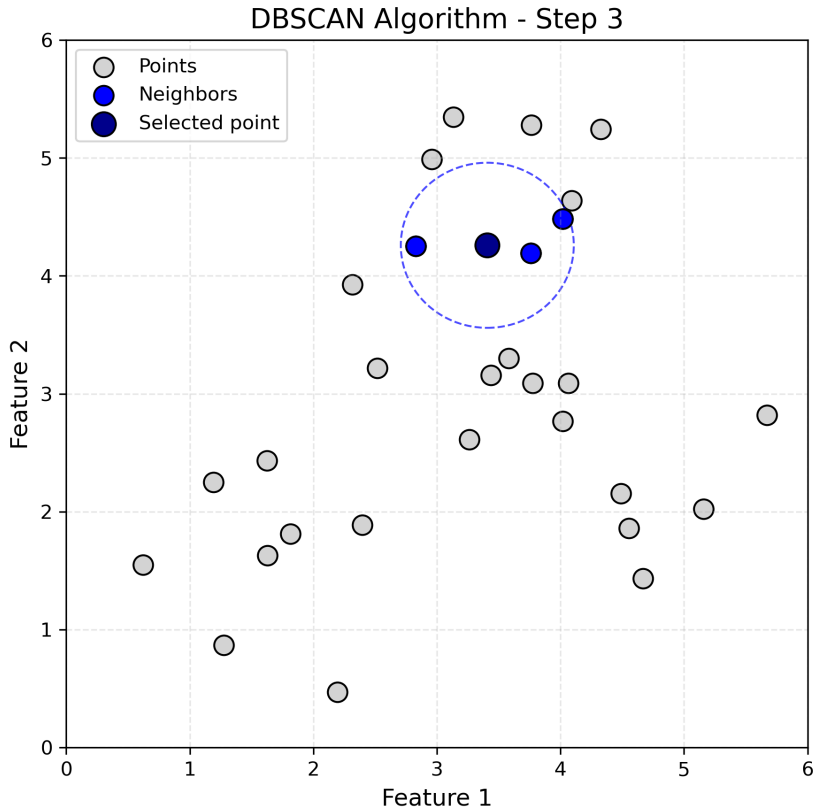
Rysunek 6: Krok 1 — mamy zbiór punktów

DBSCAN — krok 2



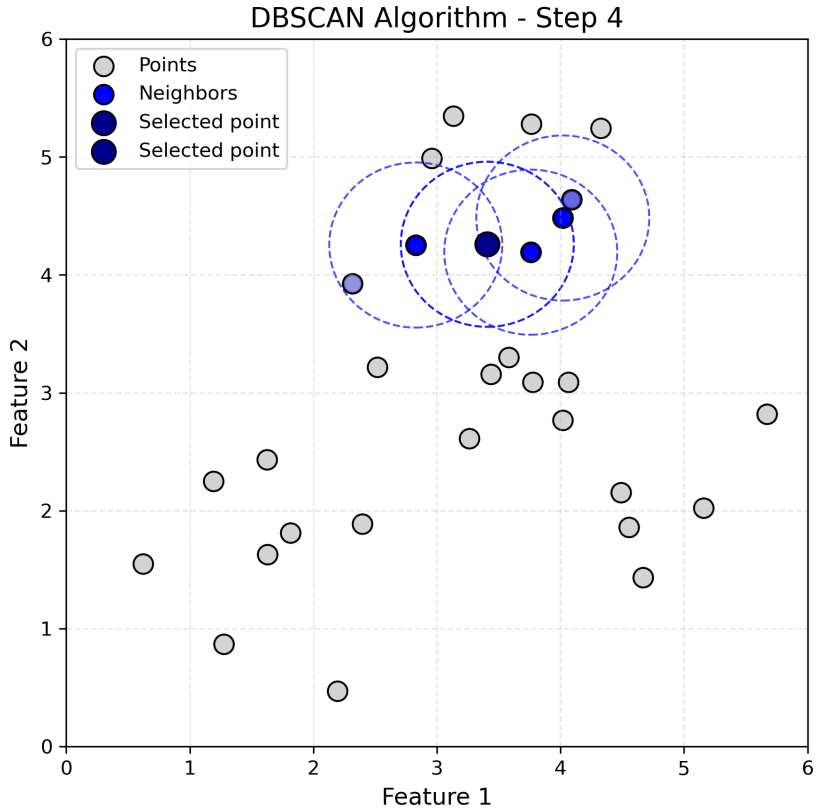
Rysunek 7: Krok 2 — wybieramy losowo jeden punkt - wokół niego będziemy tworzyć skupienie

DBSCAN — krok 3



Rysunek 8: Krok 3 — rysujemy kółko wokół punktu centralnego. Liczba punktów w odległości $< \varepsilon$ jest $\geq MinPts = 3$

DBSCAN — krok 4



Rysunek 9: Krok 4 — rysujemy kółka wokół punktów centralnych. Nowy punkt z lewej strony jest punktem granicznym, a z prawej centralnym.

Inne metody grupowania danych

- Metody hierarchiczne - aglomeracyjne i deglomeracyjne,
- Modyfikacje k-means,
- Metody gęstościowe (np. DBSCAN),
- Metody rozmytej analizy skupień,
- I inne...

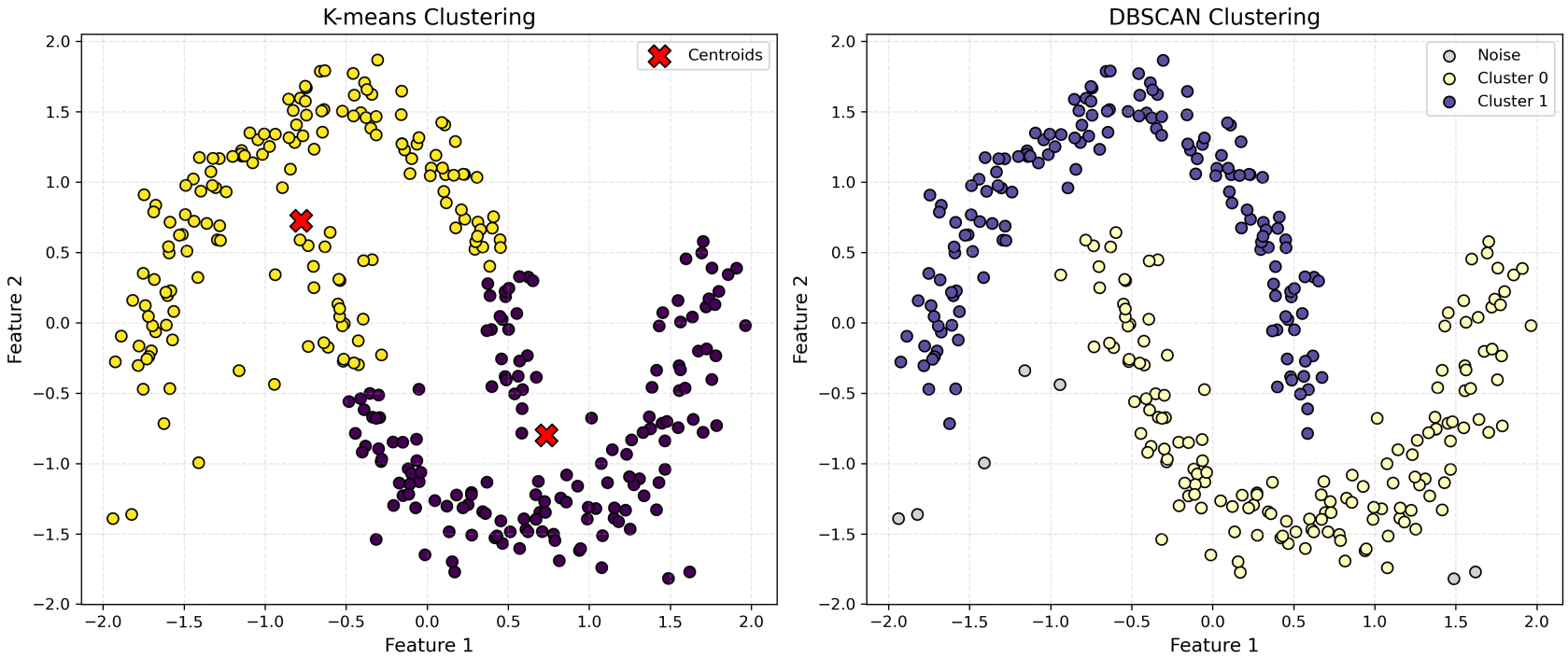
Kiedy używać k-means?

- Gdy skupienia mają podobną liczność i kształt zbliżony do d-wymiarowej kuli,
- Gdy nie mamy zbyt wielu obserwacji odstających (ang. *outliers*),
- Gdy znamy liczbę skupień,
- Gdy mamy do dyspozycji mało pamięci lub chcemy urównoleglić (sparalelizować) procedurę,
- Jest szybszy na dużych zbiorach danych oraz przy dużym wymiarze danych.

Zalety i wady DBSCAN

- Dobrze radzi sobie z „nieregularnymi" kształtami skupień,
- Oznacza obserwacje odstające jako szum,
- Radzi sobie z nierównomiernie rozłożonymi danymi,
- Większe zużycie pamięci, bo przechowuje graf sąsiedztwa,
- Jest zależny od dwóch parametrów: ε i MinPts,
- Jest mniej zależny od punktu startowego.

Porównanie grupowań



Rysunek 10: Porównanie grupowań dokonanych przez k-means oraz DBSCAN

Jak mierzyć jakość klastrów?

Mamy do wyboru miary zewnętrzne i wewnętrzne. Miary zewnętrzne porównują wyniki algorytmu klasteryzacji z rzeczywistymi etykietami klastra (np. określonymi na podstawie wiedzy eksperckiej), podczas gdy miary wewnętrzne oceniają jakość klastrów na podstawie ich struktury i gęstości.

Macierz pomyłek — przypomnienie

Miary zewnętrzne są oparte na macierzy pomyłek (ang. *Confusion matrix*). Przypomnijmy sobie czym jest macierz pomyłek.

Uwaga 2

Miary zewnętrzne służące do oceny jakości grupowania danych w klastry nie mają nic wspólnego z miarami zewnętrznymi, o których się uczymy na teorii miary.

Macierz pomyłek — przypomnienie

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Rysunek 11: Przykład klasyfikacji. Źrodło:
`https://en.wikipedia.org/wiki/Confusion_matrix`

Macierz pomyłek — przypomnienie

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Rysunek 12: Macierz pomyłek — definicja. Źrodło: `https://en.wikipedia.org/wiki/Confusion_matrix`

Macierz pomyłek — przypomnienie

		Predicted condition	
	Total 8 + 4 = 12	**Cancer** 7	**Non-cancer** 5
Actual condition	**Cancer** 8	6	2
	Non-cancer 4	1	3

Rysunek 13: Macierz pomyłek dla powyższego przykładu klasyfikacji.
Źrodło: `https://en.wikipedia.org/wiki/Confusion_matrix`

Przykłady miar zewnętrznych

- C: Prawdziwy podział na skupienia
- K: Predykcja podziału na skupienia
- n: Liczba punktów w danych
- yy: Liczba punktów będących w tym samym klastrze w C oraz K
- nn: Liczba punktów będących w różnym samym klastrze w C oraz K
- yn: Liczba punktów będących w tym samym klastrze w C ale w innym w K
- ny: Liczba punktów będących w innym klastrze w C ale w tym samym w K

Przykłady miar zewnętrznych

Indeks Randa (ang. *Rand Index*) jest zdefiniowany jako:

$$RI = \frac{yy + nn}{yy + yn + ny + nn} = \frac{yy + nn}{\binom{n}{2}} \tag{1}$$

Indeks Fowlkesa–Mallowsa (ang. *Fowlkes–Mallows index*) to:

$$FMI = \sqrt{\frac{yy}{yy + ny} \times \frac{yy}{yy + yn}} = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}}, \tag{2}$$

czyli średnia geometryczna z precyzji i czułości.

Inne miary zewnętrzne

- bazujące na teorii informacji (np. mutual information),
- bazujące na liczbie elementów w klastrach (np. purity),
- modyfikacje indeksu Randa,
- i wiele innych...

Przykłady miar wewnętrznych

- Silhouette Score — ocenia jakość klastrów na podstawie odległości między punktami w klastrze a punktami w innych klastrach. Wartości od -1 do 1, im bliżej 1 tym lepiej.
- Dunn Index — ocenia jakość klastrów na podstawie odległości między klastrami. Konkretnie jest to stosunek minimum z odległości pomiędzy klastrami do maksimum z średnicy klastra. Im większa wartość Dunn Index, tym lepsza jakość klastrów.
- Caliński–Harabasz index — stosunek wariancji pomiędzy skupieniami do wariancji wewnątrz skupień. Im większy tym lepiej.

Czym jest redukcja wymiarowości?

Technika analizy danych, która polega na zmniejszeniu liczby cech w zbiorze danych. Celem jest uproszczenie danych, aby ułatwić analizę i wizualizację. Stosowana w eksploracyjnej analizie danych, aby zrozumieć strukturę danych i znaleźć wzorce.

Zastosowania redukcji wymiaru

Mogą być używane zarówno w uczeniu nadzorowanym, jak i nienadzorowanym

- analiza obrazów,
- analiza tekstu,
- analiza genów,
- i wiele innych...

Zastosowania redukcji wymiaru

Redukcja wymiaru może się przydać w następujących zagadnieniach:

- Wizualizacja danych,
- Przyspieszenie algorytmów uczenia maszynowego,
- Ułatwienie interpretacji modeli,
- Usuwanie multiliniowości z danych,
- Wykrywanie anomalii.

Przykładowe metody redukcji wymiaru

- Analiza głównych składowych (ang. *PCA — Principal component analysis*) - przekształcenie danych do nowego układu współrzędnych; nowe osie są kombinacjami liniowymi oryginalnych cech.
- t-SNE (ang. *t-distributed Stochastic Neighbor Embedding*) — przekształca dane do niższej wymiarowości, zachowując lokalną strukturę danych.
- UMAP (Uniform Manifold Approximation and Projection) — przekształca dane do niższej wymiarowości, zachowując globalną i lokalną strukturę danych.

Na czym polega PCA?

1. Centrowanie danych: $X_c = X - 1\bar{x}^T$ gdzie $\bar{x}$ to wektor średnich.
2. Liczymy macierz kowariancji: $S = \frac{1}{n-1} X_c^T X_c$.
3. Przeprowadzamy dekompozycję macierzy S, tzn.: $S = V \Lambda V^T$ gdzie V to macierz wektorów własnych a Λ jest macierzą diagonalną z wartościami własnymi $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Na czym polega PCA?

k-wymiarowa PCA projekcja (rzut) ($k < d$) to:

$$Z = X_c V_k, \tag{3}$$

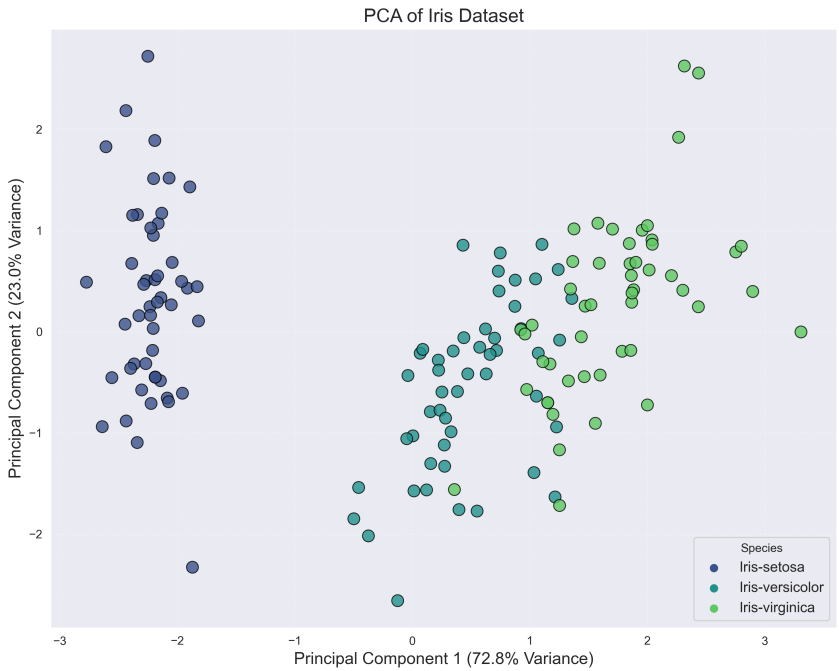
gdzie V_k zawiera k pierwszych kolumn V.
Proporcja wariancji wyjaśnianej przez i-tą główną składową:

$$\text{PVE}_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{4}$$

Skumulowana wariancja wyjaśniana przez pierwszych k głównych składowych:

$$\text{CVE}(k) = \sum_{i=1}^{k} \text{PVE}_i \tag{5}$$

Przykład PCA dla danych *iris*



Rysunek 14: Wykres rozrzutu na podstawie dwóch pierwszy głównych składowych dla zbioru danych *iris*. Te dwie składowe wyjaśniają ok. 96 procent wariancji.

Wykres interaktywny

Przykład interaktywnego wykresu na podstawie PCA na danych MPG

Koniec

Dziękuję za uwagę

Bibliografia


Krzyśko, M., Wołyński, W., Górecki, T., and Skorzybut, M.
Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości.
01 2008.


Lubischew, A. A.
On the use of discriminant functions in taxonomy.
Biometrics 18, 4 (1962), 455–477.