# Uczenie nienadzorowane (ang. Unsupervised Learning)

*By kama3 kama3*
*lis 3, 2025*

## K-means

Popularny algorytm K-Means, który działa jako analiza skupień czyli klasteryzacja.
Klasteryzacja to technika polegająca na automatycznym grupowaniu podobnych do siebie obiektów w tzw. klastry, gdzie elementy wewnątrz jednej grupy są bardziej do siebie podobne niż do elementów z innych grup. K-Means grupuje dane w klastry, pozwala identyfikować grupy podobnych stanów (np. różne stopnie uszkodzenia).

Zalety algorytmu K-means
Prostota i szybkość: Algorytm jest stosunkowo prosty w implementacji i szybki obliczeniowo, co czyni go wydajnym w przypadku dużych zbiorów danych.
Skalowalność: Dobrze skaluje się do dużych zbiorów danych.
Zawsze zbieżny: Proces iteracyjny algorytmu zawsze prowadzi do zbieżnego rozwiązania (lokalnego minimum sumy kwadratów odległości).
Łatwość interpretacji: Wynikowe klastry są łatwe do zinterpretowania.
Efektywność obliczeniowa: Dzięki zastosowaniu funkcji odległości do dopasowania instancji do centrów skupień, procesy tworzenia skupień są obliczeniowo efektywne.
Dobra wydajność dla danych dobrze rozdzielonych: Algorytm sprawdza się dobrze, gdy dane są dobrze rozdzielone i mają w przybliżeniu sferyczny (lub kulisty) kształt.

Wady algorytmu K-means:

Wymaga wstępnego podania liczby klastrów (K): Przed uruchomieniem algorytmu należy arbitralnie określić liczbę grup (K), co często jest trudne bez wcześniejszej wiedzy o danych.
Wrażliwość na inicjalizację centroidów: Wynik grupowania może zależeć od początkowego wyboru centroidów (środków klastrów). Algorytm może utknąć w lokalnym optimum, a nie globalnym.
Problem z nieregularnymi kształtami i zmienną gęstością: K-means zakłada, że klastry mają kształt kulisty i podobną wielkość. Ma trudności z grupowaniem danych o nieregularnych kształtach (np. wklęsłych) lub zmiennej gęstości.
Wrażliwość na wartości odstające (outliers): Wartości odstające mogą znacząco wpływać na położenie centroidów, co prowadzi do mniej dokładnych wyników.
Wymaga danych numerycznych: Algorytm działa najlepiej na danych numerycznych i wykorzystuje metryki odległości (zazwyczaj odległość euklidesową). Nie radzi sobie bezpośrednio ze zmiennymi kategorycznymi.
Podatność na "przekleństwo wymiarowości": W przypadku danych o bardzo dużej liczbie wymiarów, działanie algorytmu może być mniej efektywne.

Read more

## PCA / t-SNE / UMAP

Srodowisko użytkowe to przestrzeń wysokowymiarowa, przestrzeń niskowymiarowa, redukcja niepotrzebnych informacji, wizualizacja. Zalety algorytmów: szybko się oblicza (zwykłe mnożenie przez macierz), możliwość

inferencji, daje przestrzeń o tej samej wymiarowości
,dobrze wyjaśnialna, jedyny parametr to tylko transformacja na danych wejściowych- skalowanie /normalizacja, zachowuje globalną strukturę danych.

Głównym atutem tych metod jest ich fundamentalna rola w wizualizacji złożonych danych pomiarowych. Przekształcenie przestrzeni wysokowymiarowej w przestrzeń niskowymiarową (np. 2D lub 3D) pozwala człowiekowi naocznie ocenić strukturę danych, zidentyfikować klastry czy wartości odstające.
Wiele z tych algorytmów, zwłaszcza te liniowe jak PCA (analiza głównych składowych), cechuje się dużą efektywnością obliczeniową, ponieważ operacje często sprowadzają się do szybkiego mnożenia macierzy. Zaletą PCA jest również dobra wyjaśnialność (interpretowalność) wyników oraz możliwość inferencji, a także zachowanie globalnej struktury danych i redukcja niepotrzebnych informacji. Prostota transformacji jest często podkreślana jako zaleta — wymagają one zazwyczaj jedynie odpowiedniego wstępnego skalowania lub normalizacji danych wejściowych.

Read more

## Hierarchiczne grupowanie

Nazwa metod hierarchicznych wzięła się od swego rodzaju hierarchii, którą budują algorytmy implementujące to podejście. Ma ona strukturę drzewa (wizualizowanego na wykresie zwanym dendrogramem), które przedstawia proces działania algorytmu i budowy grup (poprzez podział bądź scalanie). Hierarchiczne grupowanie pozwala tworzyć drzewo zależności między stanami maszyny.

Gdy mierzymy się z problemem grupowania algorytmy hierarchiczne mają kilka zalet:

Intuicyjne – dosyć prosta idea działania. Proste w interpretacji – rozwiązanie przybiera kształt drzewa i może być prezentowane na dendrogramie.

Wady:

Mało wydajne obliczeniowo (sprawdzają się raczej na małych bazach) – duża złożoność obliczeniowa i pamięciowa.
Wrażliwe na dobór metody liczenia odległości.
Wrażliwe na wartości odstające.

Read more

## Isolation Forest / One-Class SVM

Metody detekcji anomalii uczą się „normalnego” zachowania i wykrywają odchylenia.

Zalety:

Działa dobrze w przypadku danych wielowymiarowych.

Szybki i skalowalny, z małym zapotrzebowaniem na pamięć.

Nie wymaga zakładania żadnego konkretnego rozkładu.

Zapewnia intuicyjną możliwość wyjaśnienia (anomalie wymagają mniejszej liczby podziałów w celu wyizolowania).

Wady: Mniej skuteczne, gdy anomalie są bardzo podobne do normalnych danych. Nie uwzględnia złożonych zależności czasowych (np. wykrywania oszustw w czasie).

Węzeł SVM z jedną klasą One-Class SVM korzysta z algorytmu uczenia nienadzorowanego. Węzeł ten można wykorzystać do wykrywania nowości. Wykryje on miękką granicę danego zbioru próbek, a następnie sklasyfikuje nowe punkty jako należące do tego zbioru albo do niego nienależące.

https://www.ibm.com/docs/pl/spss-modeler/18.5.0?topic=nodes-one-class-svm-node

Read more

## Autoenkodery beznadzorowane

Autoenkodery (AE) to rodzaj sztucznych sieci neuronowych wykorzystywanych do uczenia się efektywnych reprezentacji (kodowań) danych w sposób beznadzorowy. Ich głównym celem jest kompresja danych wejściowych do mniejszego wymiaru, a następnie ich odtworzenie z jak największą dokładnością. Wykrywają odchylenia poprzez błędy rekonstrukcji.

Autoenkodery oferują szereg istotnych zalet, wynikających głównie z ich zdolności do uczenia się efektywnych reprezentacji danych w sposób beznadzorowy:

Beznadzorowe uczenie się cech (Feature Learning)

Nieliniowa redukcja wymiarowości:

Elastyczność i adaptowalność

Odporność na szum (Noise Reduction)

Skuteczna detekcja anomalii:

Trudności w korzystaniu z autoenkoderów
Choć autoenkodery to potężne narzędzia, wiążą się z nimi pewne wyzwania:

Nadmierne dopasowanie: Autoenkodery potrafią nauczyć się zapamiętywać dane wejściowe, zamiast generalizować je, zwłaszcza w przypadku ograniczonych zbiorów danych.
Wybór odpowiedniego rozmiaru ukrytej przestrzeni: Wybór odpowiedniego rozmiaru warstwy wąskiego gardła ma duże znaczenie dla uzyskania optymalnej wydajności.
Interpretacja: Czasami trudno jest zinterpretować sygnały utajone o niskim poziomie samoregulacji.
Czas szkolenia: Szkolenie autoenkoderów, szczególnie w przypadku zaawansowanych architektur, może wymagać znacznych zasobów obliczeniowych i czasu.

Read more