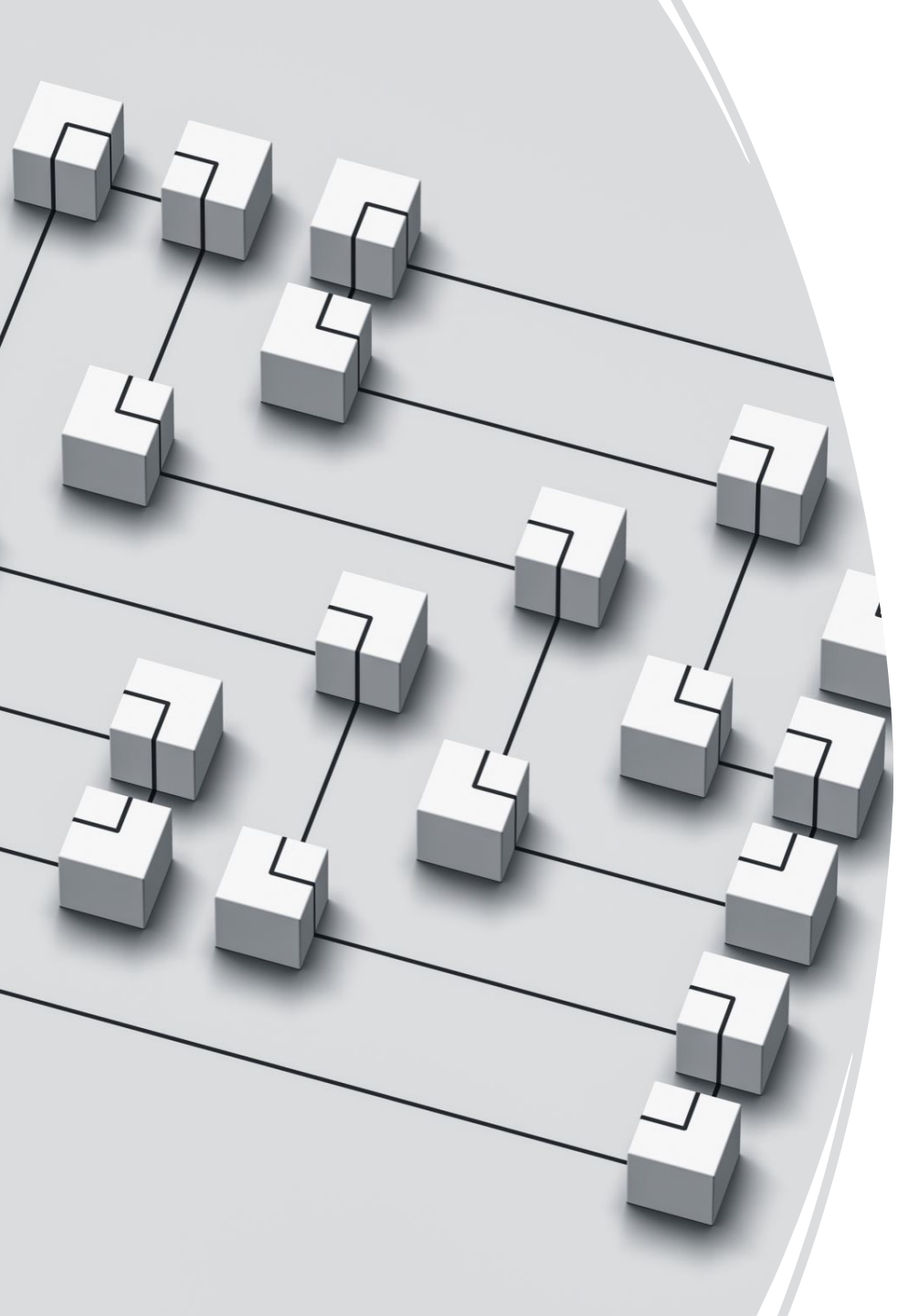

Supervised visualization of high-dimensional data using a centroid-based approach

Karolina Klisz

Bogusz Laszczyk

Justyna Mastek

Zarys tematu

W ramach projektu zostaną obliczone centroidy za pomocą różnych algorytmów klasteryzacji, a następnie wygenerowane wektory zawierające odległości do tych centroid. Dokonana zostanie wizualizacja wyznaczonych wektorów przy użyciu UMAP, t-SNE i IVHD dla różnych parametrów. Stosując dostępne metryki jakości embeddingu, porównane zostaną wizualizacje oryginalnych danych oraz danych po przekształceniach.

Etapy projektu

Klasteryzacja danych za pomocą 3 różnych algorytmów. Obliczenie centroid.
Wygenerowanie nowego zbioru danych jako odległości oryginalnych punktów od centroid.
Wizualizacja otrzymanych wektorów przy użyciu UMAP, t-SNE i IVHD.
Porównanie jakości embeddingów.

Wybrane algorytmy klasteryzacji

- K-means
- Agglomerative clustering
- DBSCAN

K-means

- Algorytm oparty na grupowaniu danych za pomocą centroid.
- Na początku wybierana jest liczba klastrów k oraz losowane są startowe centroidy. Następnie obliczana jest odległość wszystkich obserwacji od wszystkich centroid. Punkty zostają przypisane do najbliżej leżącej centroidy. Dla każdego tak uzyskanego klastra obliczana jest nowa centroida jako środek ciężkości, po czym powtarzane są powyższe kroki aż do momentu, gdy położenie centroidy pozostaje takie samo.
- W algorytmie można stosować różne metryki pomiaru odległości.

Agglomerative clustering

- Algorytm należący do rodziny hierarchicznych algorytmów klasteryzacji.
- W początkowej fazie działania algorytmu każda obserwacja stanowi osobny klaster. W każdym kolejnym kroku podobne do siebie obserwacje są łączone, aż do otrzymania jednego klastra, obejmującego cały zbiór danych.
- Algorytm ten nie zwraca centroid bezpośrednio, ale można je obliczyć jako środek ciężkości otrzymanych klastrów.

DBSCAN

- Algorytm klasteryzacji oparty na gęstości.
- Dla każdej obserwacji znajdowani są jej sąsiedzi w odległości bliższej niż założona wartość epsilon. Każda obserwacja, która ma co najmniej min_samples sąsiadów w odległości mniejszej niż epsilon, nazywana jest punktem centralnym. Wszystkie obserwacje spełniające kryteria opisane powyżej są ze sobą łączone w jedną grupę. Obserwacje, które nie są punktami centralnymi, a znajdują się w odległości epsilon, zostają przyłączone do istniejących grup. Obserwacje, które należą do grup, lecz w ich zasięgu epsilon nie znajduje się żadna nowa obserwacja, nazywane są obserwacjami granicznymi danej grupy. Wszystkie obserwacje, które nie zostały przyłączone do żadnej z grup, stają się obserwacjami odstającymi.
- Algorytm ten nie zwraca centroid bezpośrednio, ale można je obliczyć jako środek ciężkości otrzymanych klastrów.

Problemy

Wybór optymalnego algorytmu klasteryzacji

Odpowiednie dobranie parametrów (np. liczby klastrów)

Dobór metody wizualizacji zbioru danych