Sprawozdanie z laboratorium: Skalowanie i Wizualizacja Danych

Laboratorium 5: MDS / nieliniowa redukcja wymiarowości

18 maja 2020

Autor: **Adam Czyżewski** 127198 ISWD adam.czyzewski@student.put.poznan.pl

1 Wykorzystane zbiory danych

1. *Cars* (`http://www.cs.put.poznan.pl/ibladek/students/skaiwd/lab5/cars.csv`)
   Opis: Zestawienie marek samochodów opisanych na 4 atrybutach.
   Rozmiar: (15×4)
2. *Swiss Roll* (`sklearn.datasets.make_swiss_roll(1600)`)
   Opis: Wygenerowany zbiór danych.
   Rozmiar: (1600×3)
3. *Digits* (`sklearn.datasets.load_digits()`)
   Opis: Reprezentacja graficzna ręcznie pisanych cyfr z zakres $0 - 9$.
   Rozmiar: (1797×64)

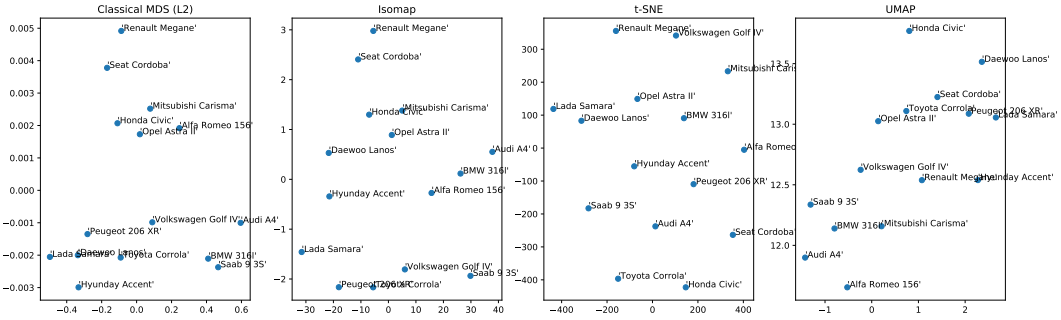
2 Wykorzystane algorytmy

1. *Metric MDS* (`własna implementacja oparata na metodzie PCA`)
2. *Isomap* (`sklearn.manifold.Isomap`)
3. *t-SNE* (`sklearn.manifold.TSNE`)
4. *UMAP* (`https://github.com/lmcinnes/umap`)

3 Wyniki

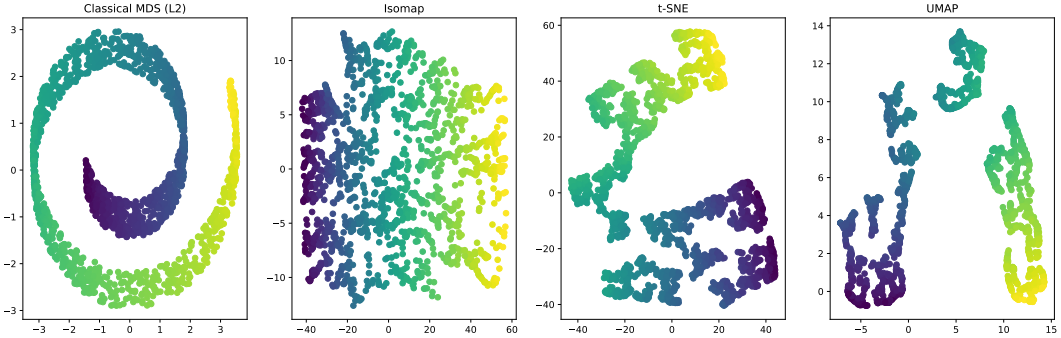
3.1 Domyślne parametry

3.1.1 Cars



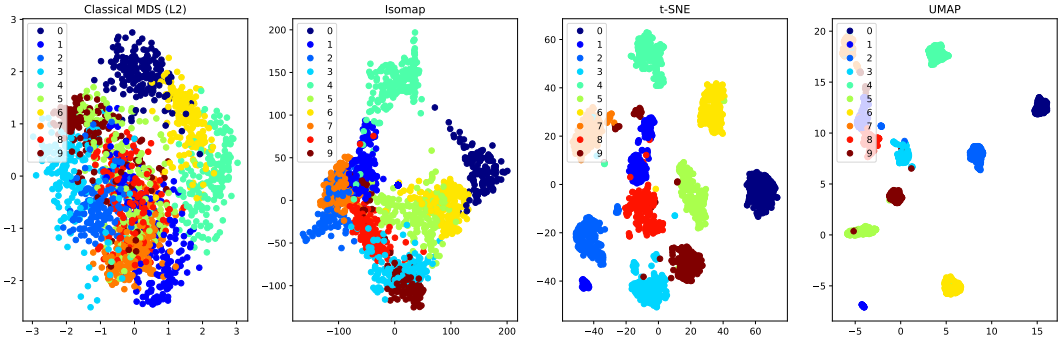
Rysunek 1: Wynik działania wszystkich metod na zbiorze *Cars*

3.1.2 Swiss roll



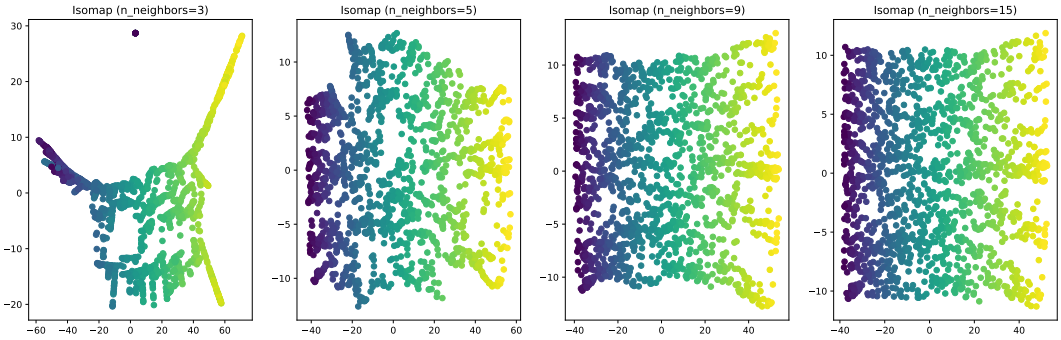
Rysunek 2: Wynik działania wszystkich metod na zbiorze *Swiss roll*

3.1.3 Digits

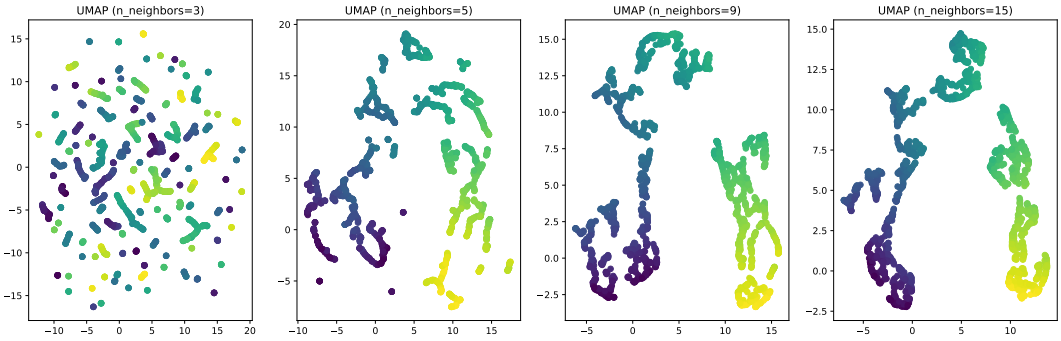


Rysunek 3: Wynik działania wszystkich metod na zbiorze *Digits*

3.2 Tuning parameterów na zbiorze *Swiss roll*



Rysunek 4: Wizualizacja różnych wartości parametru n_neighbors algorytmu *Isomap* na zbiorze *Swiss roll*



Rysunek 5: Wizualizacja różnych wartości parametru n_neighbors algorytmu *UMAP* na zbiorze *Swiss roll*

4 Wnioski

Kolejność prezentowania wybranych algorytmów złożyła się z ich możliwościami (rosnąco). Uzyskane wnioski:

- W przypadku klasycznego MDS opartego na (PCA) wykorzystanie metryki *euklidesowej* sprawia, że metoda ta zachowuje się identycznie jak zastosowanie metody *PCA* na danych wejściowych (minimalizacja liniowych odległości pokrywa się z maksymalizacją liniowych korelacji). Przedstawione wykresy zastosowania tej metody na nieliniowych danych obrazują słabość tej metody. Prawdopodobnie jest możliwa poprawa wyników w przypadku zastosowania nielinowej metodą odległości (np. *fold change*)
- Na podstawie eksperymentów z regulacją ilości sąsiadów w algorytmach *Isomap* i *UMAP* widać różnice koncepcyjne ich działania. *Isomap* jako algorytm grafowy bazuje na odległościach pomiędzy punktami, a więc kiedy liczba jego sąsiadów jest na tyle mała, że nie sięga ona do innych "krawędzi" danych (w przypadku *Swiss roll*) to liczba ta nie wpływa znacząco na wygląd projekcji. *UMAP* jako algorytm opierający się również na metodach grafowych jak i elementach probabilistycznych jest wrażliwy

na zmienną liczbę sąsiadów, a im większa ona jest tym tworzone grupy są coraz bardziej jednolite i oddalone od siebie.

- Jakość algorytmów została prawdopodobnie najlepiej zobrazowana na ostatnim zbiorze (*Digits*). *Classical MDS* (jako *PCA*) nie był w stanie liniowo oddzielić od siebie konkretnie żadnej grupy. *Isomap* potrafił dość dobrze oddzielić od siebie grupy cyfr o specficznej strukturze (np. 4), lecz nadal trudniejsze przykłady (podobne do siebie, np. 1 i 7) są ze sobą nierozróżnialne. Dopiero *t-SNE* jest w stanie oddzielić te grupy (i podobnie jak jego następnik *UMAP*) praktycznie w stanie rozwiązać problem klasyfikacji, co może stanowić bardzo dobry i ciekawy przykład uczenia nienadzorowanego i możliwości tego typów algorytmów.