Metody MDS

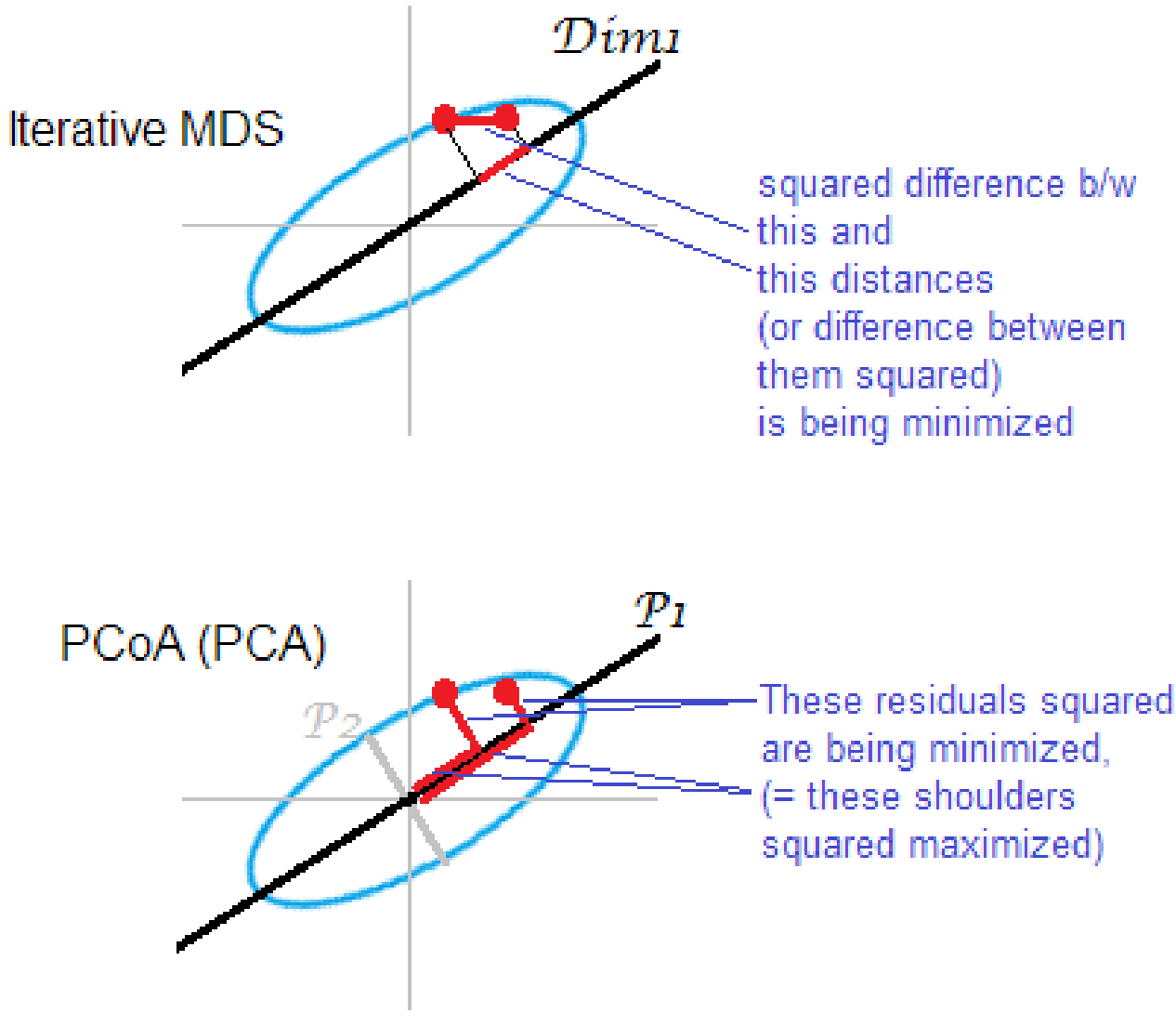
Mikołaj Dłuś
Mateusz Monasterski

github.com/Dwoosh/MDS_tutorial

Po co nam one skoro PCA tak ładnie wizualizowało?

- PCA i jemu podobne algorytmy wybierają „interesującą” **liniową** projekcję danych
- Problem gdy dane zawierają ważne nieliniowe struktury

- PCA – najdokładniejsza reprezentacja danych w przestrzeni nisko-wymiarowej, zachowuje kowariancję danych
- Iteratywny – prawdziwy MDS ma jako główny cel rekonstrukcję/zachowanie odległości między parami obiektów


Iterative MDS
Dim1
squared difference b/w
this and
this distances
(or difference between
them squared)
is being minimized
PCoA (PCA)
P1
P2
These residuals squared
are being minimized,
(= these shoulders
squared maximized)

Manifold Learning

- Podejście do nonlinear dimensionality reduction (NLDR)
- Można go rozumieć jako próbę uogólnienia liniowych algorytmów takich jak PCA, aby były wrażliwe na nieliniową strukturę danych

Multidimensional Scaling

- Są to sposoby wizualizacji poziomu podobieństwa lub odmienności poszczególnych przypadków zbioru danych.
- Celem tej analizy jest wykrycie sensownych ukrytych wymiarów, które pozwalają badaczowi wyjaśnić obserwowane podobieństwa lub odmienności (odległości) między badanymi obiektami.
- Szuka nisko-wymiarowej reprezentacji danych, w których odległości odpowiednio szanują odległości w pierwotnej przestrzeni o wysokim wymiarze.

Multidimensional Scaling

- Obrót układu współrzędnych i odbicie zwierciadlane nie zmieniają odległości pomiędzy punktami, więc wynik skalowania można poddać rotacji lub odbiciu.

- Korzysta z symetrycznej macierzy odległości – różnic miedzy obiektami

- W scikit-learn w przypadku atrybutu dissimilarity=='precomputed‘
  Należy na wejściu podać właśnie taką macierz

Algorytm metryczny MDS

- W metrycznym MDS macierz podobieństwa wejściowego wynika z metryki (a zatem respektuje nierówność trójkąta)

- odległości między wyjściowymi dwoma punktami są następnie ustawiane tak, aby były jak najbliżej danych podobieństwa lub niepodobieństwa.

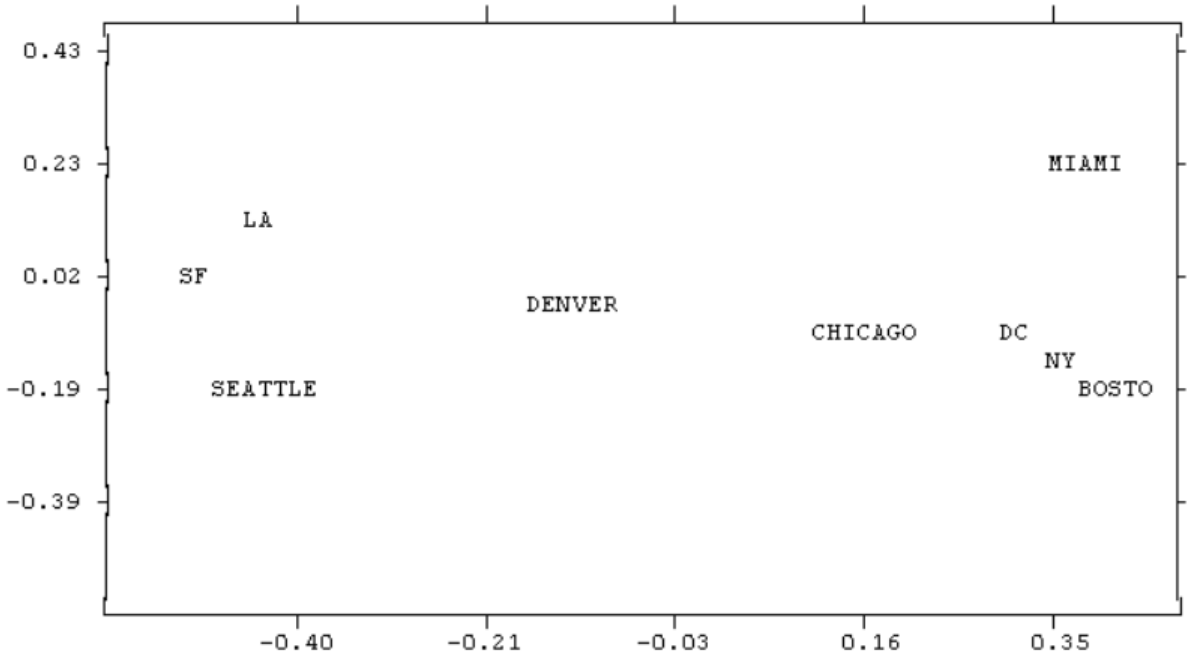
Algorytm niemetryczny MDS

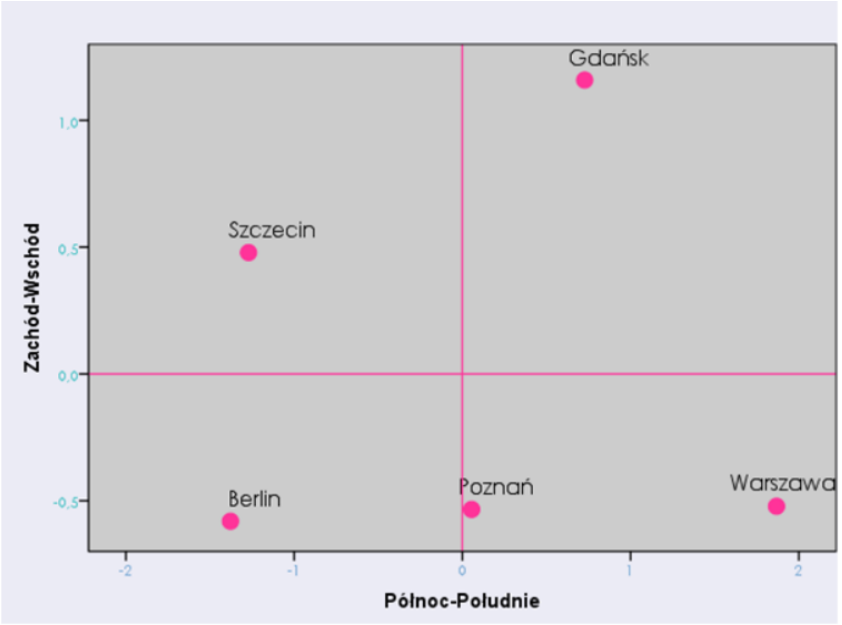
- W wersji niemetrycznej algorytmy będą próbowały zachować kolejność odległości, a zatem poszukują monotonicznego związku między odległościami w przestrzeni osadzonej a podobieństwami / odmiennościami.

Przykład

		1	2	3	4	5	6	7	8	9
		BOST	NY	DC	MIAM	CHIC	SEAT	SF	LA	DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

Korzystając z takiej macierzy odległości między miastami, jest w stanie przedstawić na wykresie 2d rozmieszczenie tych miast względem siebie


0.43
0.23
0.02
-0.19
-0.39
MIAMI
LA
SF
DENVER
CHICAGO
DC
NY
SEATTLE
BOSTO
-0.40
-0.21
-0.03
0.16
0.35


Gdańsk
Szczecin
Berlin
Poznań
Warszawa
Zachód-Wschód
Północ-Południe
1,0
0,5
0,0
-0,5
-2
-1
0
1
2

Multidimensional scaling

Wady:

- Iteracyjny proces
- Wymaga dużej mocy obliczeniowej do obliczania macierzy niepodobieństwa przy każdej iteracji

Isomap – Isometric Mapping

- Jest jednym z pierwszych podejść do Manifold Learning
- Może być postrzegana jako rozszerzenie MDS albo Kernel PCA
- Isomap służy do obliczania quasi-izometrycznego, nisko-wymiarowego embeddingu zestawu punktów danych o dużych wymiarach.

Działanie Isomap składa się z trzech etapów:

- wyszukanie najbliższych sąsiadów
- wyszukanie najkrótszej ścieżki w grafie
- częściowy rozkład wartości własnej

Isomap

Wady:

- działa słabo jeśli manifold zawiera dziury i/lub nie jest gęsty

Locally Linear Embedding – LLE

- Tak jak poprzednio przedstawione metody, wyszukuje nisko-wymiarowej projekcji danych zachowując odległości w lokalnych sąsiedztwach

- Może być traktowany jako seria lokalnych PCA, które są globalnie porównywane w celu znalezienia najlepszego nieliniowego embeddingu.

Działanie LLE składa się z trzech etapów:

- wyszukanie najbliższych sąsiadów
- budowa macierzy wag
- częściowy rozkład wartości własnej

LLE

Zaleta w stosunku do Isomap:

- szybsza optymalizacja jeśli jest zaimplementowany do wykorzystania algorytmów na macierzach rzadkich