POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

Wykład 6 - Redukcja wymiaru. Metody nieliniowe

DR INŻ. AGATA MIGALSKA

01 CEL I MOTYWACJA

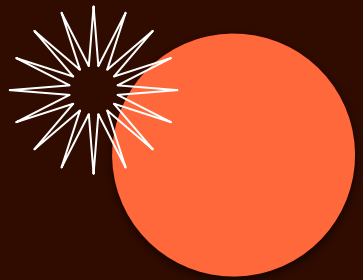
02 KERNEL PCA

03 LLE

04 MDS

05 ISOMAP

06 T-SNE



01

CEL I MOTYWACJA

KLĄTWA WYMIAROWOŚCI

Klątwa wymiarowości objawia się pogarszającymi się i/lub niestabilnymi wynikami modeli uczenia maszynowego wraz ze wzrostem liczby zmiennych.

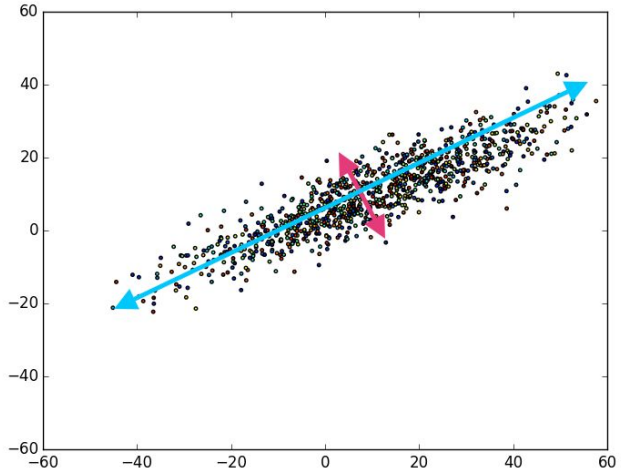
Klątwa wymiarowości występuje, gdy wymiarowość danych wzrasta, a dostępne dane stają się rzadkie.
Aby rozwiązać ten problem, można:

1. zwiększenie próbek danych lub
2. zmniejszenie wymiaru danych.

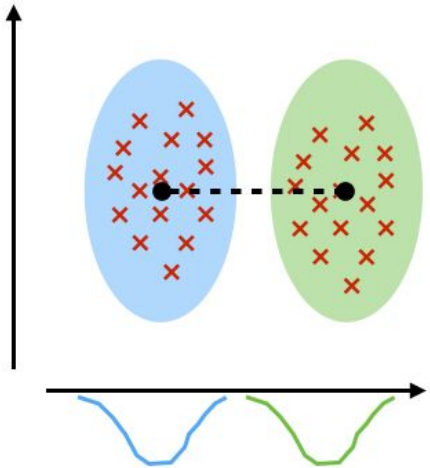
Zwiększenie próbek danych może nie zawsze być możliwe, dlatego zmniejszenie wymiaru danych może być kluczowym wyborem.

LINIOWE METODY REDUKCJI WYMIAROWOŚCI

Analiza głównych składowych (PCA)
- osie składowych, które maksymalizują wariancję

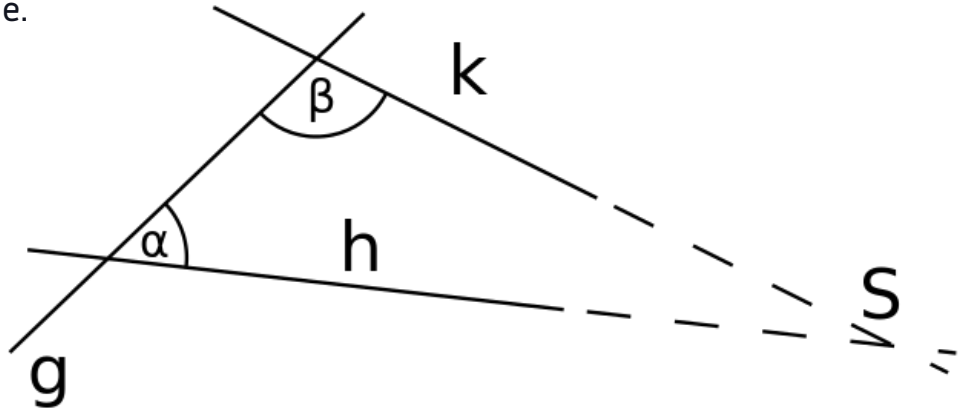


Liniowa analiza dyskryminacyjna (LDA)
- osie składowych, które maksymalizują separację klas



PIĄTY AKSJOMAT EUKLIDESA

Jeżeli prosta przecina dwie proste, tworząc dwa kąty wewnętrzne po tej samej stronie o sumie mniejszej niż dwa kąty proste, to te dwie proste przecinają się po tej stronie, po której znajdują się owe kąty wewnętrzne.



Zdania równoważne:

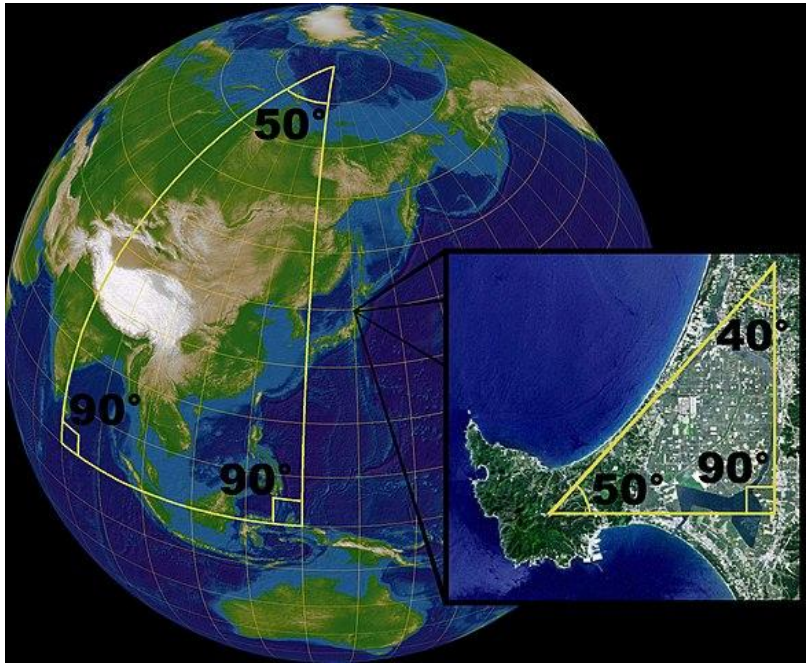
- Na każdym trójkącie można opisać okrąg.
- Suma kątów wewnętrznych trójkąta jest równa dwóm kątom prostym.
- Przez każdy punkt przechodzi tylko jedna prosta równoległa do danej prostej.

ROZMAITOŚĆ

Rozmaitość n-wymiarowa – zbiór punktów, wyposażony w geometrię, która ma lokalnie własności geometrii euklidesowej.

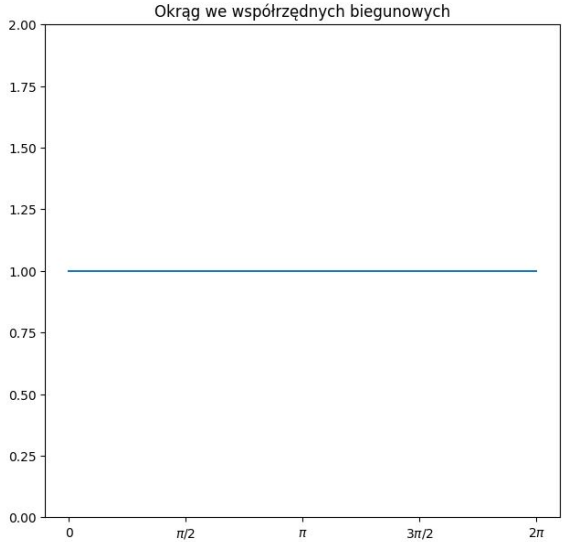
Wyposażenie w geometrię oznacza, że podane zostały wzory na obliczanie odległości między punktami, kątów między prostymi, pól powierzchni itp., przy czym lokalnie odległości między punktami są dane wzorami takimi jak w przestrzeni euklidesowej.

Lokalność - każdy punkt rozmaitości ma otoczenie, które jest homeomorficzne z przestrzenią euklidesową R^n.



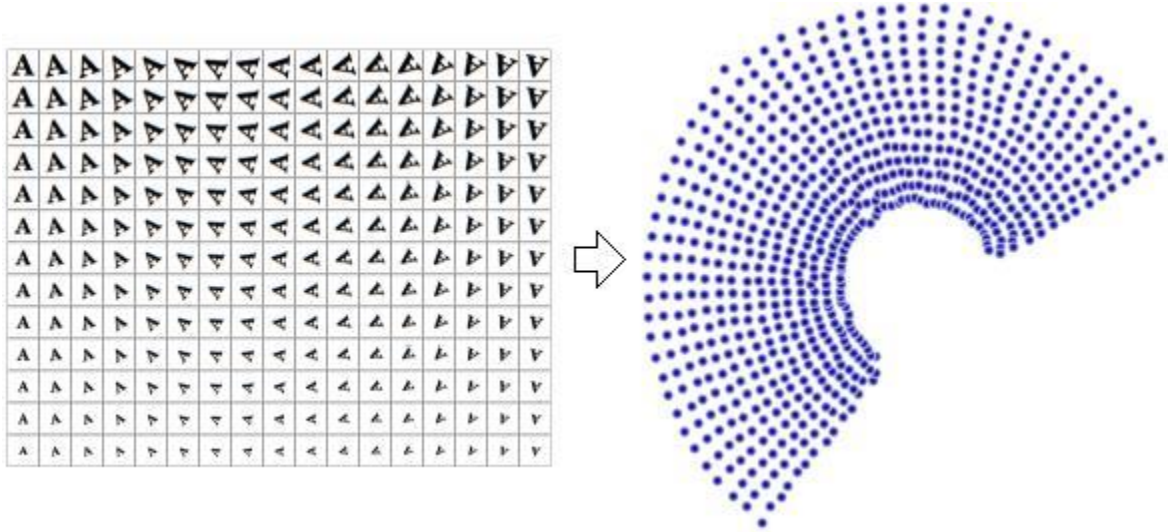
“UCZENIE SIĘ ROZMAITOŚCI” (ANG. MANIFOLD LEARNING)

Algorytmy dla tego zadania opierają się na założeniu, że wymiarowość wielu zbiorów danych jest tylko sztucznie zawyżona.

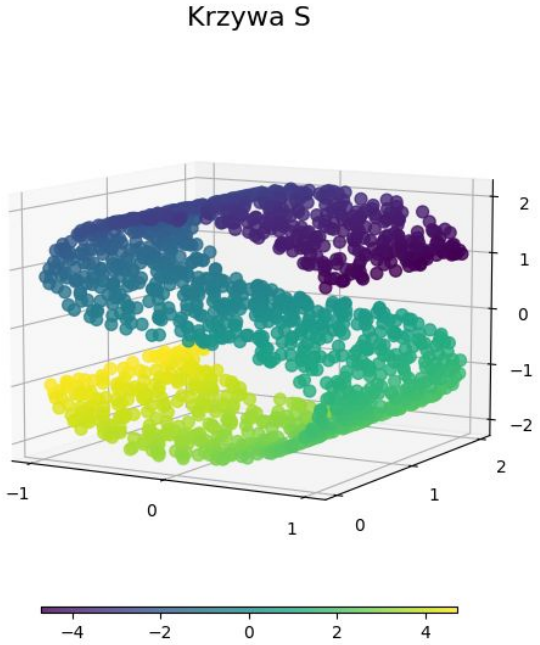


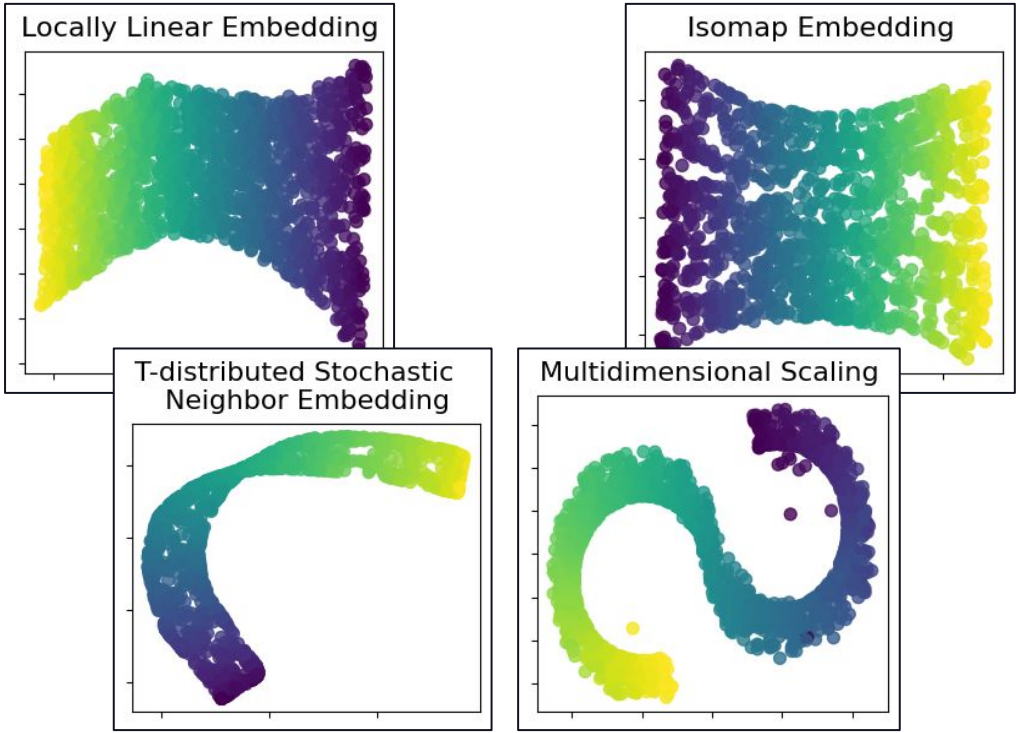
“UCZENIE SIĘ ROZMAITOŚCI” (ANG. MANIFOLD LEARNING)

Algorytmy dla tego zadania opierają się na założeniu, że wymiarowość wielu zbiorów danych jest tylko sztucznie zawyżona.

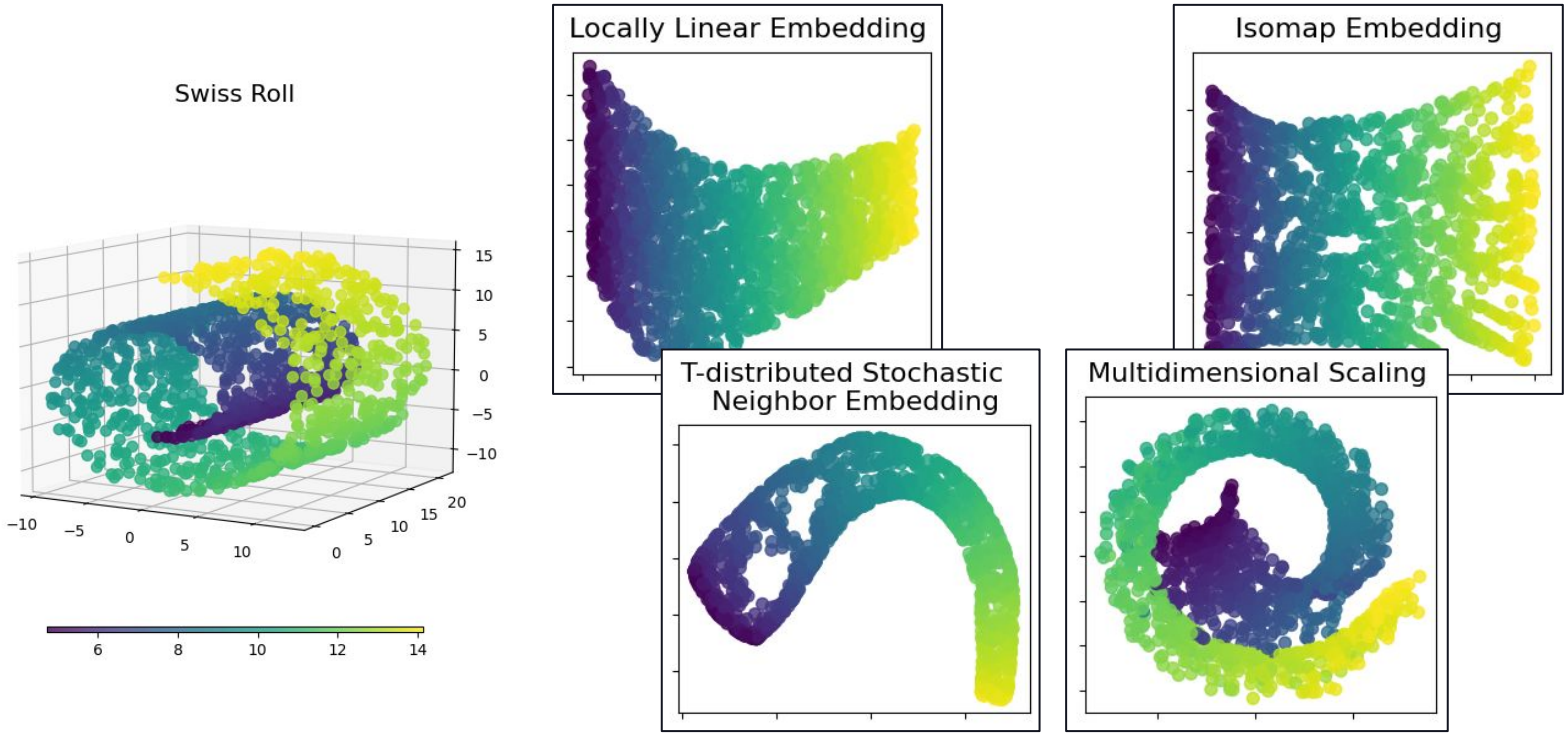


UCZENIE SIĘ ROZMAITOŚCI





UCZENIE SIĘ ROZMAITOŚCI



METODY LINIOWE I NIELINIOWE REDUKCJI

wychwytują globalną strukturę danych

METODY LINIOWE

- Analiza składowych głównych (PCA)
- Liniowa analiza dyskryminacyjna (LDA)
- Classical Multidimensional Scaling (cMDS)

skala struktury danych zależy od liczby sąsiadów

METODY NIELINIOWE

- Kernel PCA
- Quadratic / Generalized Linear Analysis
- Nonmetric Multidimensional Scaling
- Locally Linear Embedding
- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Isomap
- UMAP
- ...

sprzyjają zachowaniu lokalnej struktury

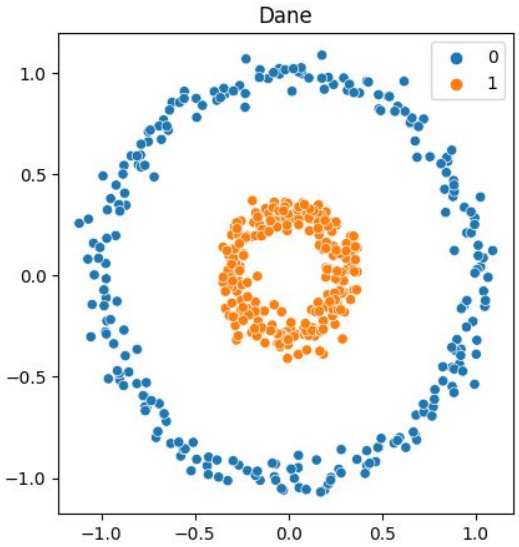
02

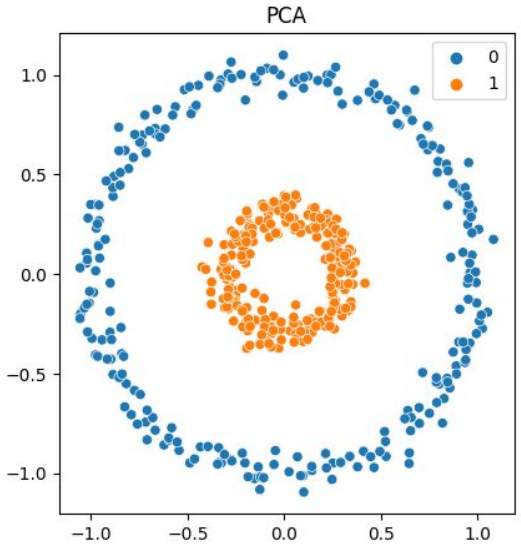
KERNEL PCA

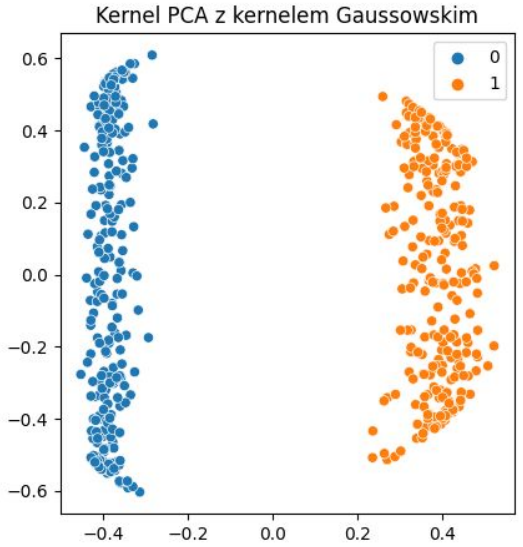
PCA

- Można wykazać, że głównymi składowymi są wektory własne macierzy kowariancji danych.
- Główne składowe są obliczane poprzez rozkład własny macierzy (EVD) lub rozkład wg wartości osobliwych (SVD).
- Pierwszym głównym składnikiem zbioru k zmiennych jest zmienna pochodna utworzona jako liniowa kombinacja pierwotnych zmiennych, która wyjaśnia największą wariancję. Drugi główny składnik wyjaśnia największą wariancję tego, co pozostaje po usunięciu efektu pierwszego składnika. Po k iteracjach cała wariancja zostanie wyjaśniona.

PCA vs KERNEL PCA NA DANYCH NIELINIOWYCH



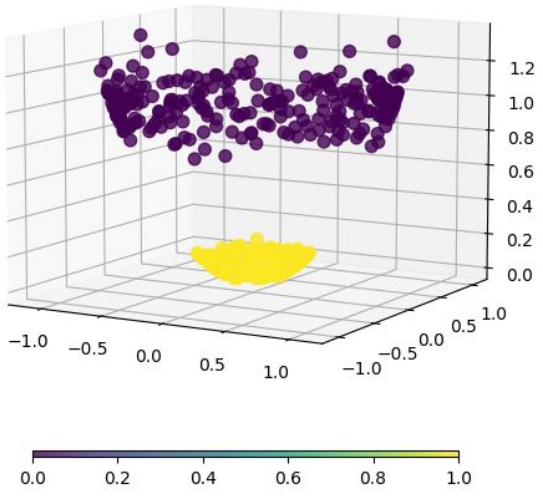




REPREZENTACJA STRUKTURY NIELINIOWEJ

- Rzutowanie do wyższego wymiaru może uprościć dane, których nie da się oddzielić liniowo.
- Zauważmy, że chociaż N punktów nie może być ogólnie rozdzielonych liniowo w d<N wymiarach, prawie zawsze można je liniowo rozdzielić w d>=N wymiarach.
- W Kernel PCA „wybiera się” nietrywialną, arbitralną funkcję Φ, która nigdy nie jest obliczana jawnie, co pozwala na użycie bardzo wysokowymiarowych Φ.

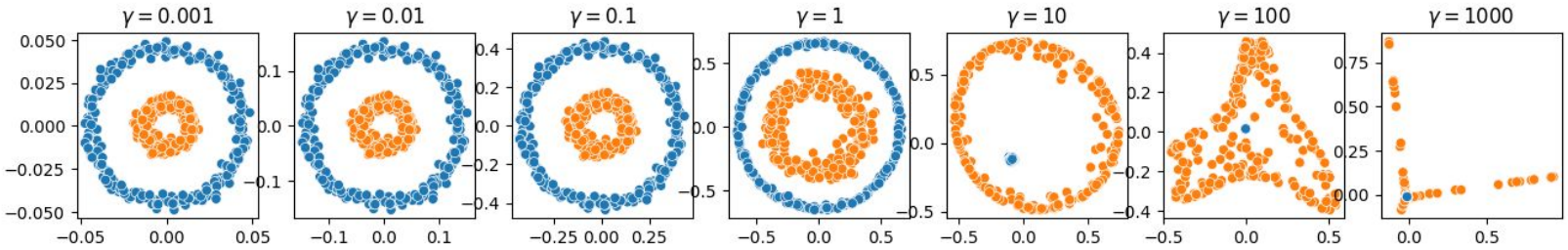
$$(x_1, x_2) \Rightarrow (x_1, x_2, x_1^2 + x_2^2)$$



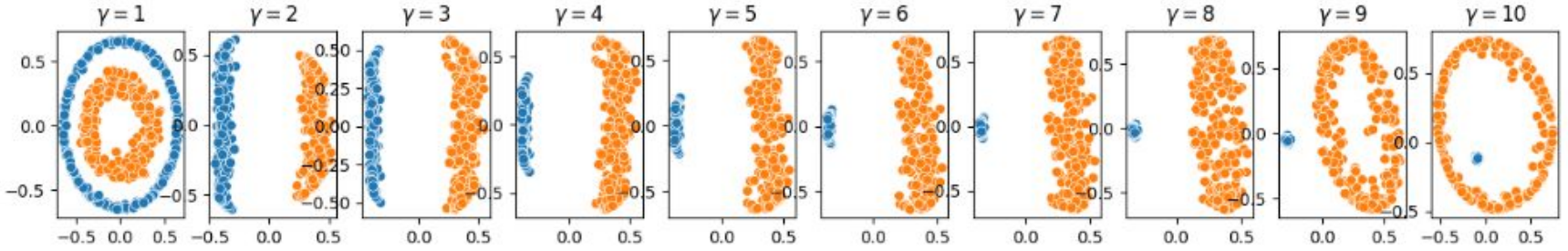
KERNEL PCA

- Obserwacja: PCA polega na znalezieniu wektorów własnych macierzy kowariancji XX^T
- Kernel PCA polega na znalezieniu wektorów własnych macierzy $K = \Phi(X)\Phi(X)^T$
- W rezultacie, możemy (bez obliczania wartości $\mathbf{\Phi}$(X)) zamodelować dowolną nieliniową transformację $\mathbf{\Phi}$(X), o ile jesteśmy w stanie wydajnie obliczać iloczyn skalarny $\phi(x_i)^T\phi(x_j)$. Zabieg ten nazywany jest “kernel trick”.

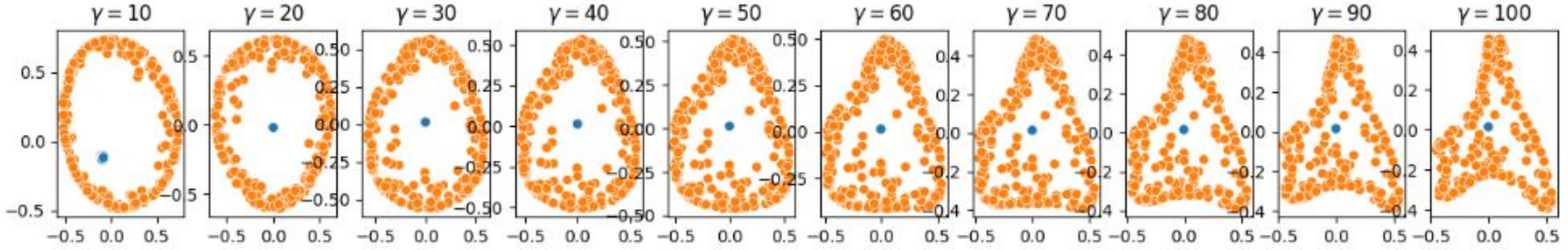
WPSÓŁCZYNNIK JĄDRA GAMMA (KERNEL RBF)



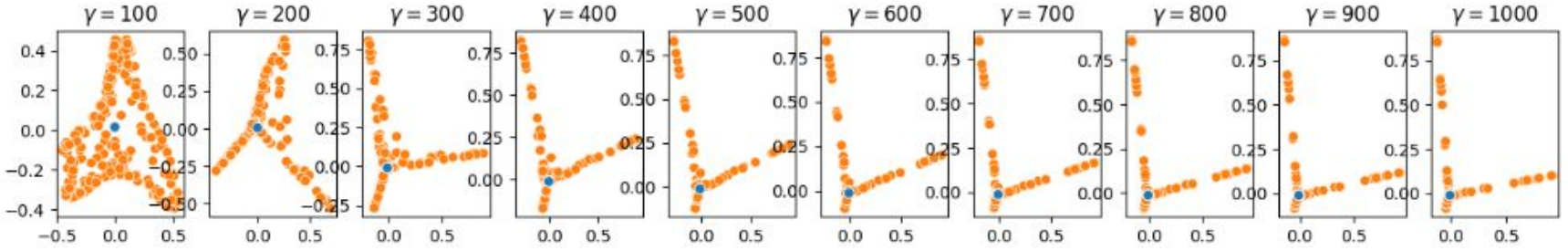
[11]:
```
plot_kernel_pca(circles2, list(range(1, 11)));
```

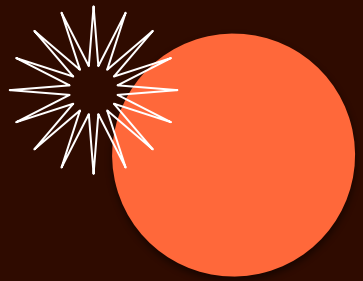


[12]:
```
plot_kernel_pca(circles2, list(range(10, 101, 10)));
```



[13]:
```
plot_kernel_pca(circles2, list(range(100, 1001, 100)));
```

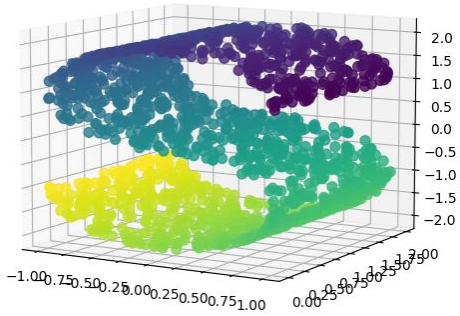


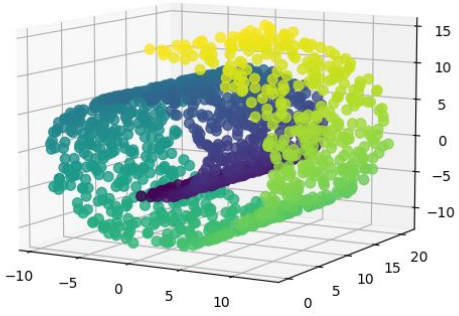


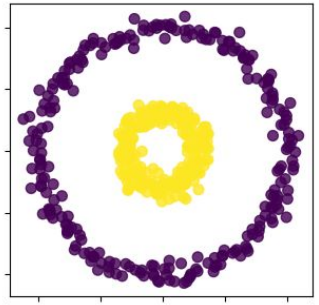
03

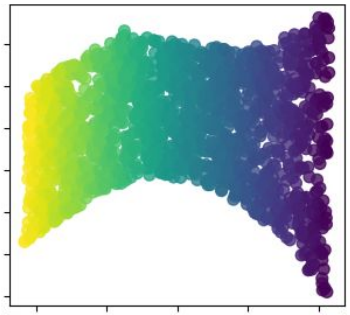
LOCALLY LINEAR EMBEDDING

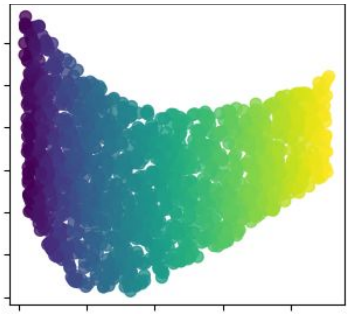
LOCALLY LINEAR EMBEDDING

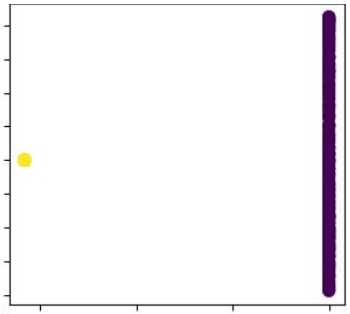




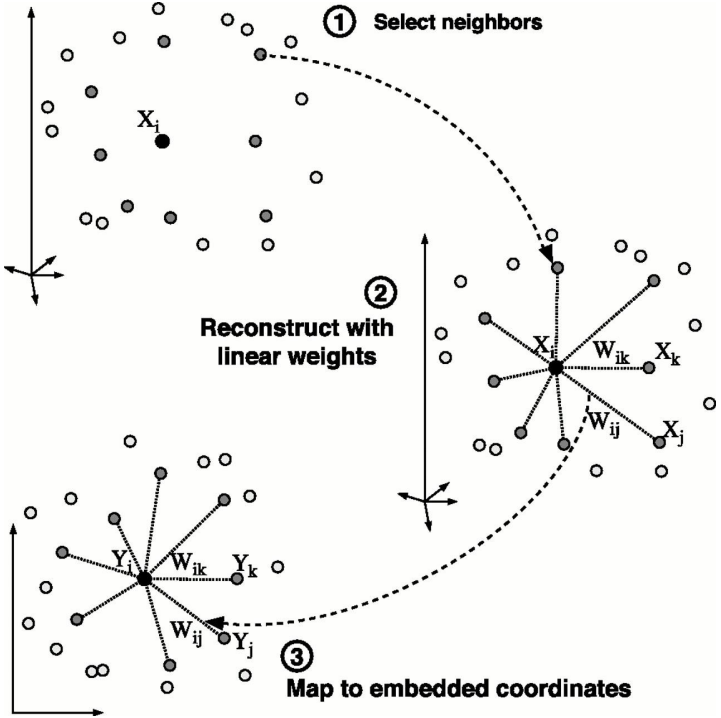








LOCALLY LINEAR EMBEDDING



LLE to nieliniowa metoda redukcji wymiarów, która osadza wysokowymiarowe punkty danych w przestrzeni o niższych wymiarach, zakładając, że każdy punkt i jego najbliżsi sąsiedzi znajdują się w rozmaitości liniowej.

LLE zakłada, że rozmaitość jest wypukła.

LOCALLY LINEAR EMBEDDING

Rozważamy przestrzeń danych o wysokiej wymiarowości D.

Niech $\widetilde{X_i}$, i=1, ..., N to N wektorów w tej D-wymiarowej przestrzeni.

Znajdź k najbliższych sąsiadów (wg odległości euklidesowej) dla każdego wektora $\widetilde{X_i}$.
Niech N_i oznacza zbiór indeksów sąsiadów i-tego punktu.

Znajdź optymalne, lokalnie wypukłe kombinacje k najbliższych sąsiadów do reprezentacji każdego oryginalnego wektora. To jest równoważne znalezieniu minimum

$$\varepsilon(W) = \sum_i \left| \widetilde{X_i} - \sum_{j \in N_i} W_{ij} \widetilde{X_j} \right|^2 ,$$

gdzie $\sum_j W_{ij} = 1.$

Powyższy problem można rozwiązać metodą najmniejszych kwadratów.

LOCALLY LINEAR EMBEDDING

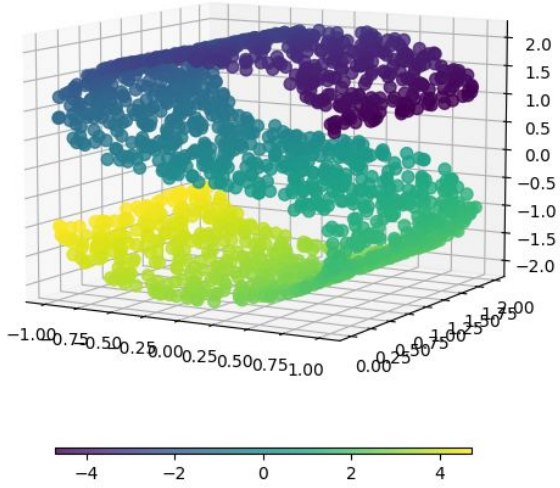
Następnie LLE rozważa przestrzeń projekcyjną. Przestrzeń rzutu ma wymiar znacznie mniejszy niż D. Niech $\widetilde{Y}_i$ będzie rzutem $\widetilde{X}_i$ w przestrzeni projekcyjnej. Rzuty $\widetilde{Y}_i$ są dobrane tak, aby zminimalizować następującą funkcję celu:

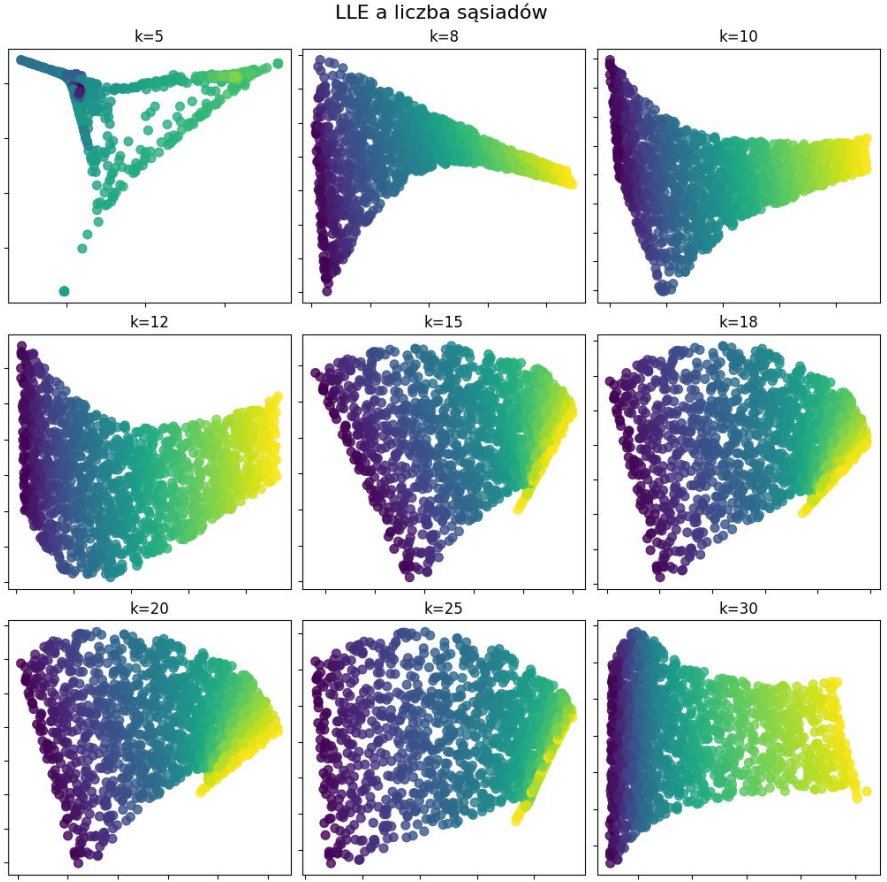
$$\Phi(Y) = \sum_i \left| \widetilde{Y}_i - \sum_{j \in N_i} W_{ij} \widetilde{Y}_j \right|^2.$$

Jest to równoważne ze znalezieniem reprezentacji o niższych wymiarach, tak że zachowane są lokalne reprezentacje wypukłe.

LICZBA SĄSIADÓW

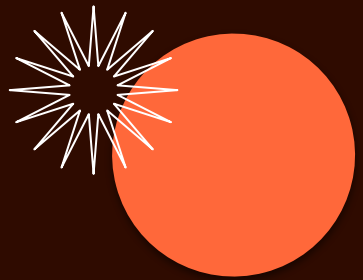
Krzywa S





LICZBA SĄSIADÓW

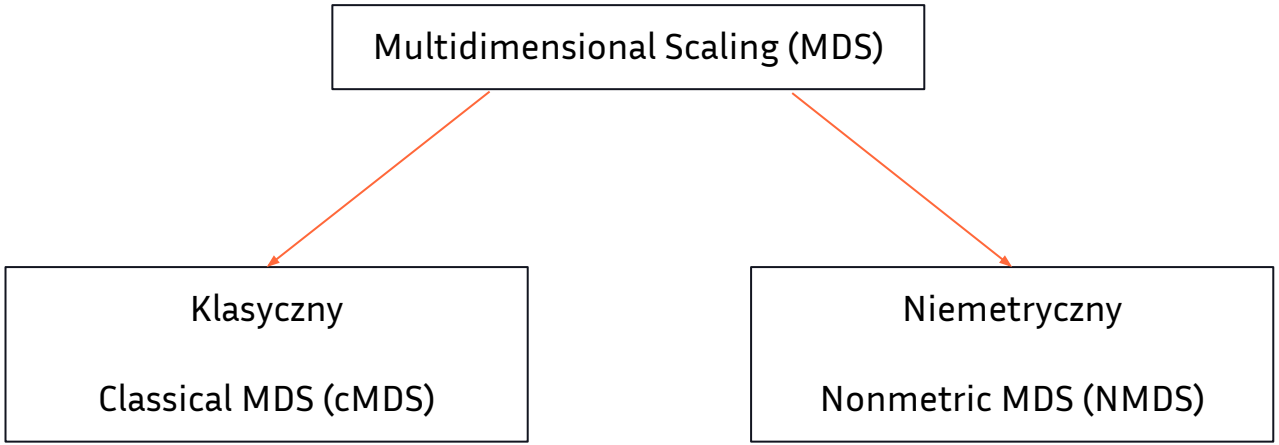
MAŁE K	K=[8,30]	DUŻE K
metoda przechwytuje bardziej lokalną strukturę danych, kosztem pominięcia niektórych informacji globalnych	WIARYGODNE ODWZOROWANIE	poprawa w uchwyceniu struktury globalnej, ale możliwym kosztem utraty informacji lokalnych



04

MULTI-DIMENSIONAL SCALING

cMDS (PCoA) vs NMDS



również nazywany Principle Coordinate Analysis (PCoA)

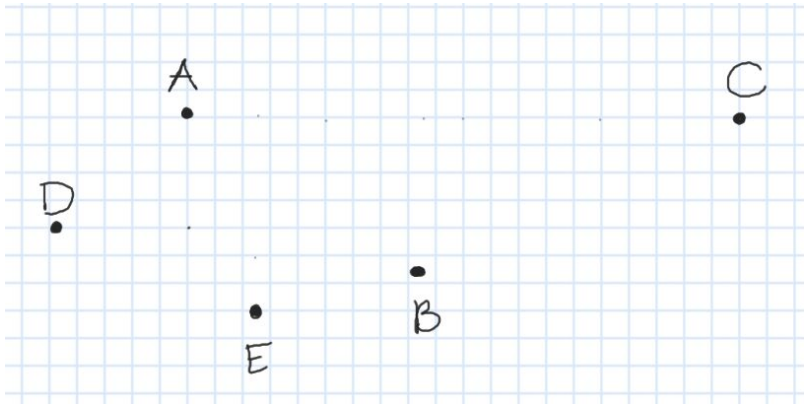
używa porządku (rang) zamiast odległości

KLASYCZNY MDS

Macierz odległości pomiędzy miastami

	A	B	C	D	E
A	0	10	20	5	7
B	10	0	13	12	6
C	20	13	0	24	19
D	5	12	24	0	7
E	7	6	19	7	0

Rekonstrukcja lokalizacji miast



KLASYCZNY MDS

Rozważane są punkty X_i w przestrzeni metrycznej Ω , $X_i \in \Omega$.

Dla $1 \leq l \neq m \leq N$, niech d(l , m) oznacza odległość między X_l i X_m.

Chcemy znaleźć $X_i' \in R^k,\ i = 1, \ldots, N,$ gdzie k jest ustaloną liczbą całkowitą, tak aby rozwiązać następujący problem optymalizacji:

$$\min_{X_i' \in R^k} \sum_{l \neq m} \left(d(l,m) - d'(l,m) \right)^2$$

gdzie d'(l,m) - odległość pomiędzy X'_l i X'_m w R^k.

KLASYCZNY MDS

MDS to doskonała metoda utrzymania globalnej struktury danych.
Podobnie jak PCA, jest również podatna na zaniedbywanie utrzymywania małych odległości i jest mniej przydatny w uchwyceniu struktury danych nieliniowych.

Oczekuje się, że PCA i MDS będą działać lepiej na danych z odpowiednio dużą frakcją dużych odległości między punktami danych.

METRYCZNY MDS

Zamiast metryki euklidesowej, możemy użyć:

- metryki miejskiej / manhattańskiej (L_1)
- maksimum (L_∞)
- metryki Minkowskiego (L_m)

NIEMETRYCZNY MDS

- Zamiast macierzy odległości konstruujemy macierz podobieństwa obiektów.
- Podobieństwo nie musi spełniać reguły trójkąta. Np. ludzie mogą zgodzić się, że Korea Północna jest podobna do Korei Południowej i że Korea Północna jest również podobna do Iranu, ale uważają, że Korea Południowa i Iran bardzo się różnią.
- Niemetryczny MDS zakłada, że ranga podobieństwa ma znaczenie. Optymalizowana funkcja "stresu" sprawdza, czy porządek rang wszystkich odległości w parach w rozwiązaniu jest taki sam, jak porządek rang podobieństwa:

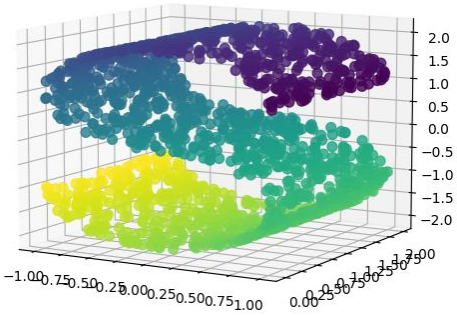
$$Stress = \sqrt{\frac{\sum (f(x) - d)^2}{\sum d^2}}$$

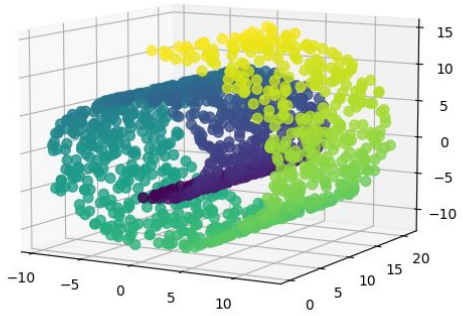
gdzie x - wektor podobieństwa,
f(x) - monotoniczna transformacja x,
d - odległości pomiędzy punktami.

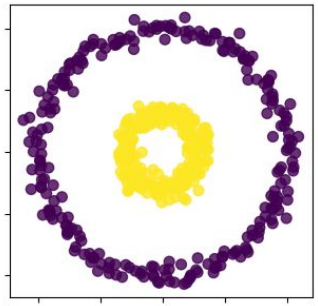
05

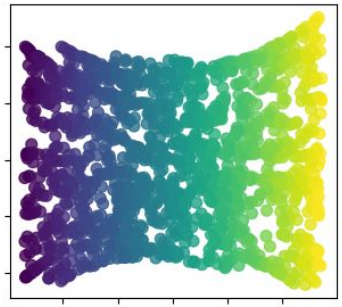
ISOMAP

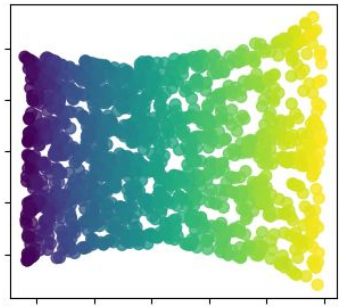
ISOMAP

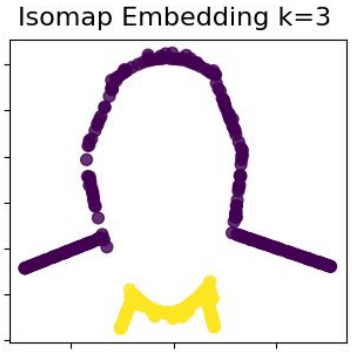

2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

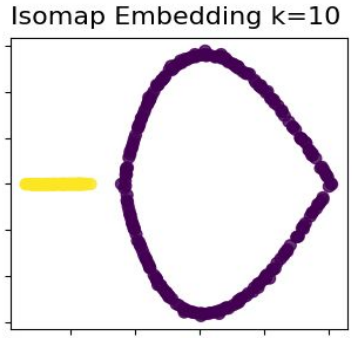

15
10
5
0
−5
−10
−10
−5
0
5
10
0
5
10
15
20



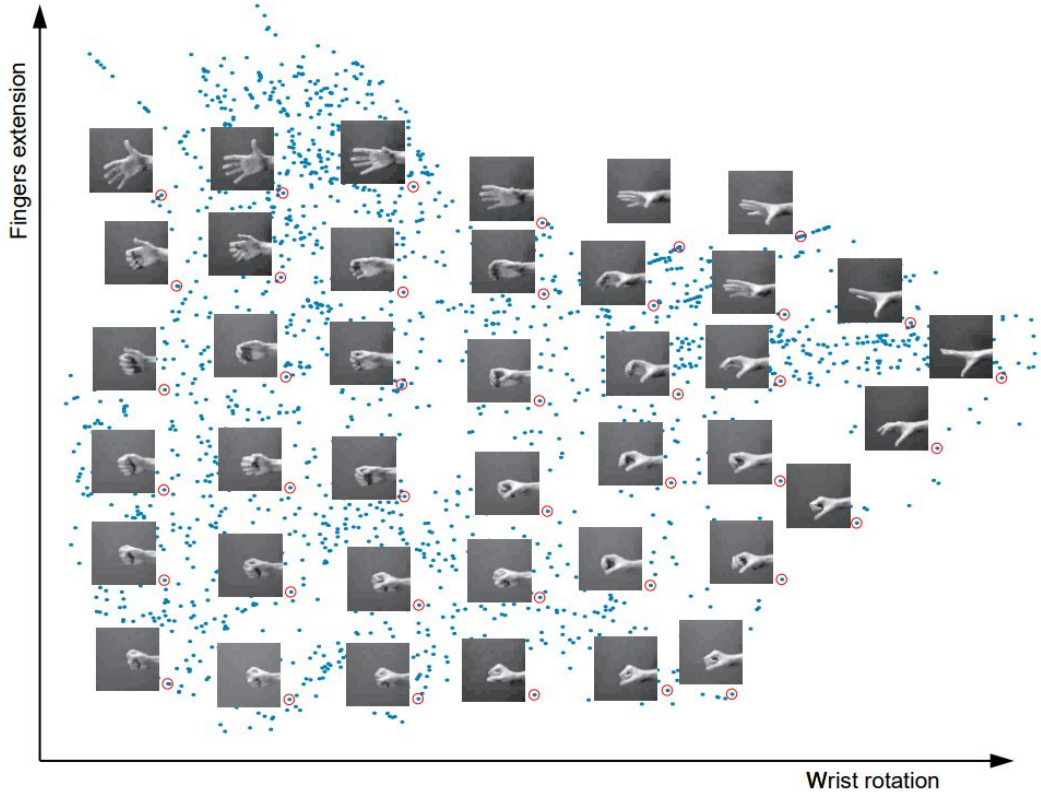





Isomap Embedding k=3


Isomap Embedding k=10

ISOMAP



Isomap (K=6) zastosowany do N=2000 obrazów (64 na 64 piksele) dłoni w różnych konfiguracjach.

Źródło: Tenenbaum, J. B., Silva, V. D., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319-2323.

ALGORYTM

Rozważmy N punktów, X_i , i=1, 2, ..., N, w przestrzeni danych.
Dla każdego punktu danych X_i rozważ jego sąsiadów.
Istnieją dwie możliwości:
1. k-najbliżsi sąsiedzi każdego punktu X_i ; lub
2. ε -sąsiedztwo, które obejmuje wszystkie punkty znajdujące się nie dalej niż ε -odległość od X_i.

Niech N_i oznacza zbiór indeksów punktów sąsiadujących z X_i.

Konstruujemy graf, w którym każdy X_i jest wierzchołkiem, a dwa wierzchołki są połączone wtedy i tylko wtedy, gdy $i \in N_j$ lub $j \in N_i$.

Zdefiniuj odległość między dwoma punktami, X_i i X_j, jako sumę długości łuku najkrótszej ścieżki łączącej X_i i X_j.

Projekcja niskowymiarowa jest następnie generowana przez wywołanie metrycznego MDS.

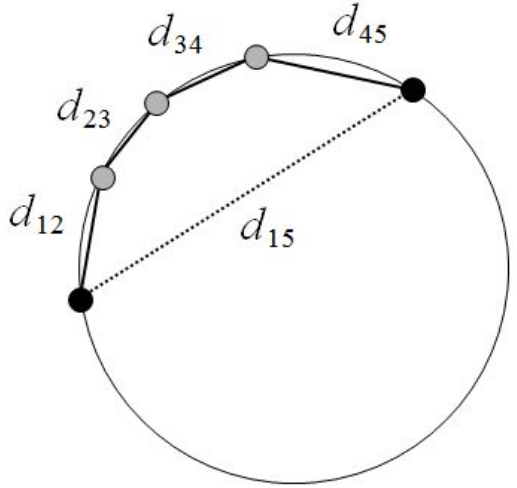
ODLEGŁOŚĆ GEODEZYJNA

Długość najkrótszej ścieżki w grafie
Najkrótszą ścieżkę można obliczyć za pomocą programowania dynamicznego (np. Dijkstra, 1959).

Odległość geodezyjna
Odległość geodezyjna między dwoma punktami na rozmaitości to długość najkrótszej krzywej znajdującej się na rozmaitości i łączącej oba punkty.

Bernstein, de Silva, Langford i Tenenbaum (2000) pokazali, że odległość graficzna jest w pewnym sensie dobrym substytutem odległości geodezyjnej.

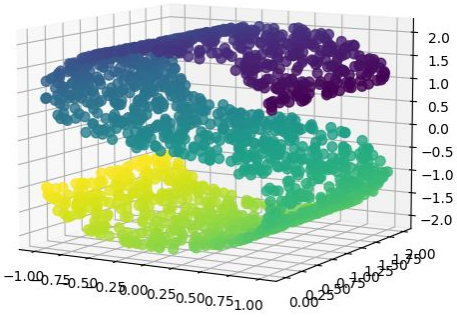
Odległość graficzna jest obliczalna na podstawie danych, podczas gdy odległość geodezyjna nie jest obliczalna.

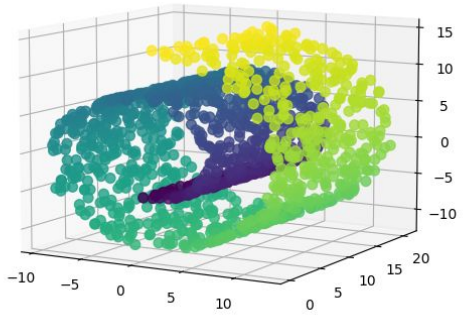


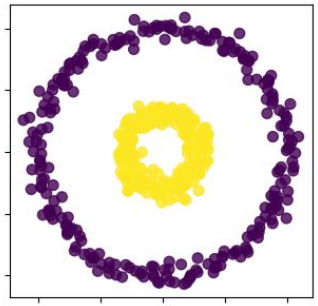
06

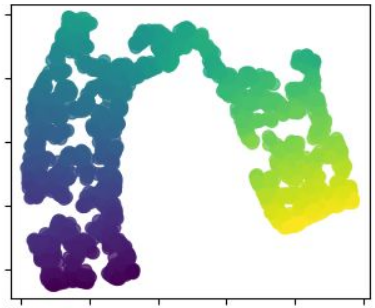
T-SNE

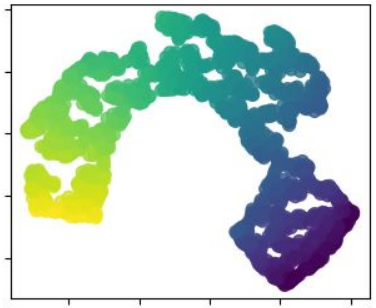
T-SNE

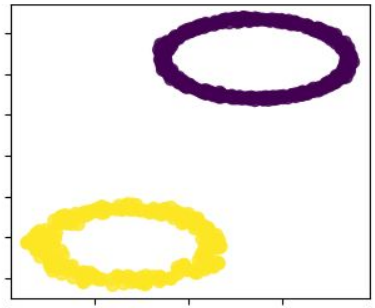

2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


15
10
5
0
−5
−10
−10
−5
0
5
10
0
5
10
15
20









T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

t-SNE jest nieliniową techniką redukcji wymiarowości, która bazuje na podobieństwie pomiędzy punktami.
Podobieństwo punktu danych x□ do punktu danych x_i jest prawdopodobieństwem warunkowym $p_{j|i}$, że x_i wybierze x□ jako swojego sąsiada. Podobieństwo punktu jest proporcjonalne do gęstości prawdopodobieństwa wg rozkładu Gaussa wyśrodkowanym w x_i i z wariancją σ_i^2

$$p_{i||j} = \frac{\exp\left(\frac{-dis(x_i,x_j)^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(\frac{-dis(x_o,x_k)^2}{2\sigma_i^2}\right)}.$$

Podobieństwo punktów przestrzeni wysokowymiarowej jest obliczane jako $p_{ij} = \frac{p_{i|j|} + p_{j|i|}}{2N}$,
a w przestrzeni niskowymiarowej jako

$$q_{ij} = \frac{\left(1 + dis(x_i,x_j)^2\right)^{-1}}{\sum_{k\neq l}\left(1 + dis(x_i,x_j)^2\right)^{-1}}.$$

T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

- Kolejną częścią t-SNE jest stworzenie przestrzeni niskowymiarowej o takiej samej liczbie punktów jak w oryginalnej przestrzeni.
- Punkty rozkładane są losowo w przestrzeni.
- Celem algorytmu jest znalezienie podobnego rozkładu prawdopodobieństwa w przestrzeni niskowymiarowej. Zamiast rozkładu Gaussa, używany jest rozkład t-Studenta z jednym stopniem swobody. (Rozkład t-Studenta ma dłuższe ogony, dzięki czemu punkty nie “tłoczą się”).
- Podobieństwo punktów przestrzeni niskowymiarowej jest obliczane jako $q_{ij} = \frac{\left(1 + dis(x_i, x_j)^2\right)^{-1}}{\sum_{k \neq l} \left(1 + dis(x_i,\ x_j)^2\right)^{-1}}.$
- Aby zoptymalizować ten rozkład, t-SNE wykorzystuje dywergencję Kullbacka-Leiblera między prawdopodobieństwami warunkowymi p_{j|i} i q_{j|i}: $D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$

07

PODSUMOWANIE

JAK WYBRAĆ METODĘ REDUKCJI?

1. Sprawdź typ posiadanych danych - różne metody mają zastosowanie do danych ciągłych, kategorycznych, liczebności lub odległości.
2. Weź pod uwagę swoją intuicję i wiedzę domenową na temat zbieranych pomiarów. Uwzględnienie natury i rozdzielczości danych jest ważne, ponieważ metody redukcji wymiarowości mogą koncentrować się na odzyskiwaniu globalnych lub lokalnych struktur danych.
3. Jeśli obserwacje w Twoich danych mają przypisane etykiety klas, a Twoim celem jest uzyskanie reprezentacji, która najlepiej dzieli je na znane kategorie, możesz rozważyć użycie nadzorowanych technik redukcji wymiarowości.

Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLoS computational biology, 15(6), e1006907.
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907

THANKS!

DZIĘKUJĘ ZA UWAGĘ