Własności danych o dużej liczbie wymiarów

Mateusz Kobos

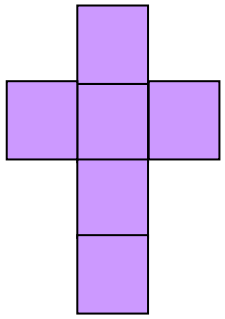
08.03.2006

Spis treści

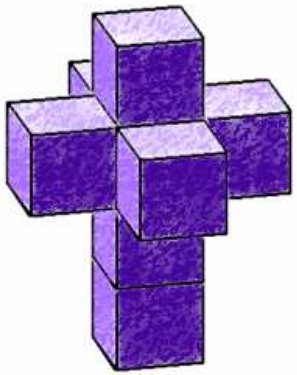
Nieintuicyjność wymiarów > 3

- Dane można reprezentować jako n-wymiarowe punkty (wektory)
- Przestrzeń n-D dla n>3 trudno sobie wyobrazić
- Charles Howard Hinton (1853-1907)
  - Zajmował się popularyzacja przestrzeni 4-D
  - ukuł nazwę Tesserakt dla siatki hiperześcianu 4-D

Siatka dla sześcianu 3-D:



Siatka dla sześcianu 4-D:



Przykłady HDD (High-Dimensional Data)

- Dane:
  - medyczne (np. EEG)
  - finansowe (wskaźniki giełdowe)
  - multimedialne (obrazki, filmy)
  - zakupy dokonywane za pomocą kart kredytowych
  - szeregi czasowe

Spis treści

Nieintuicyjność HDD

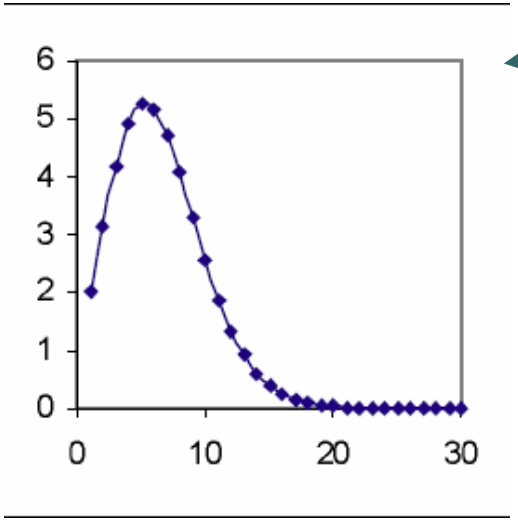
Objętość kuli

- Wzór na objętość kuli:

$$V_r(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d$$

- Gdzie:
  - r – promień kuli
  - d – liczba wymiarów

Wykresy dla kul, gdzie r=1:

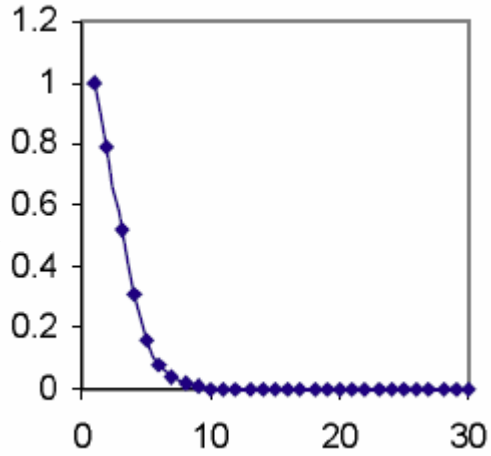


X: liczba wymiarów(d)

Y: objętość kuli

X: liczba wymiarów (d)

Y: objętość kuli/objętość hiperześcianu o boku 2r



Rysunek z [11]

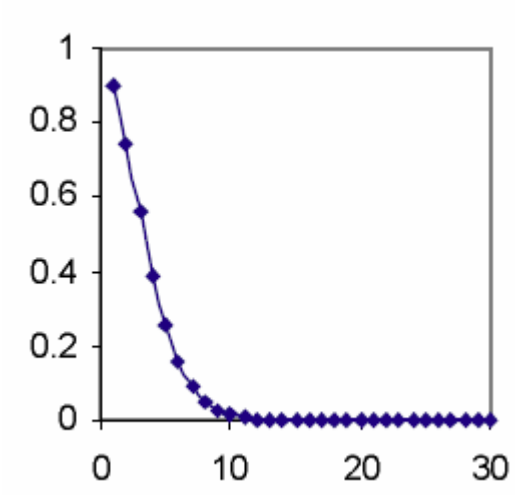
Rysunek z [11]

Nieintuicyjność HDD

Rozkład normalny

- Znany fakt: w 1-D rozkładzie normalnym N(0,1)

90% próbek znajduje się w przedziale [-1,65; 1,65]



Rysunek z [11]

X: liczba wymiarów (d)

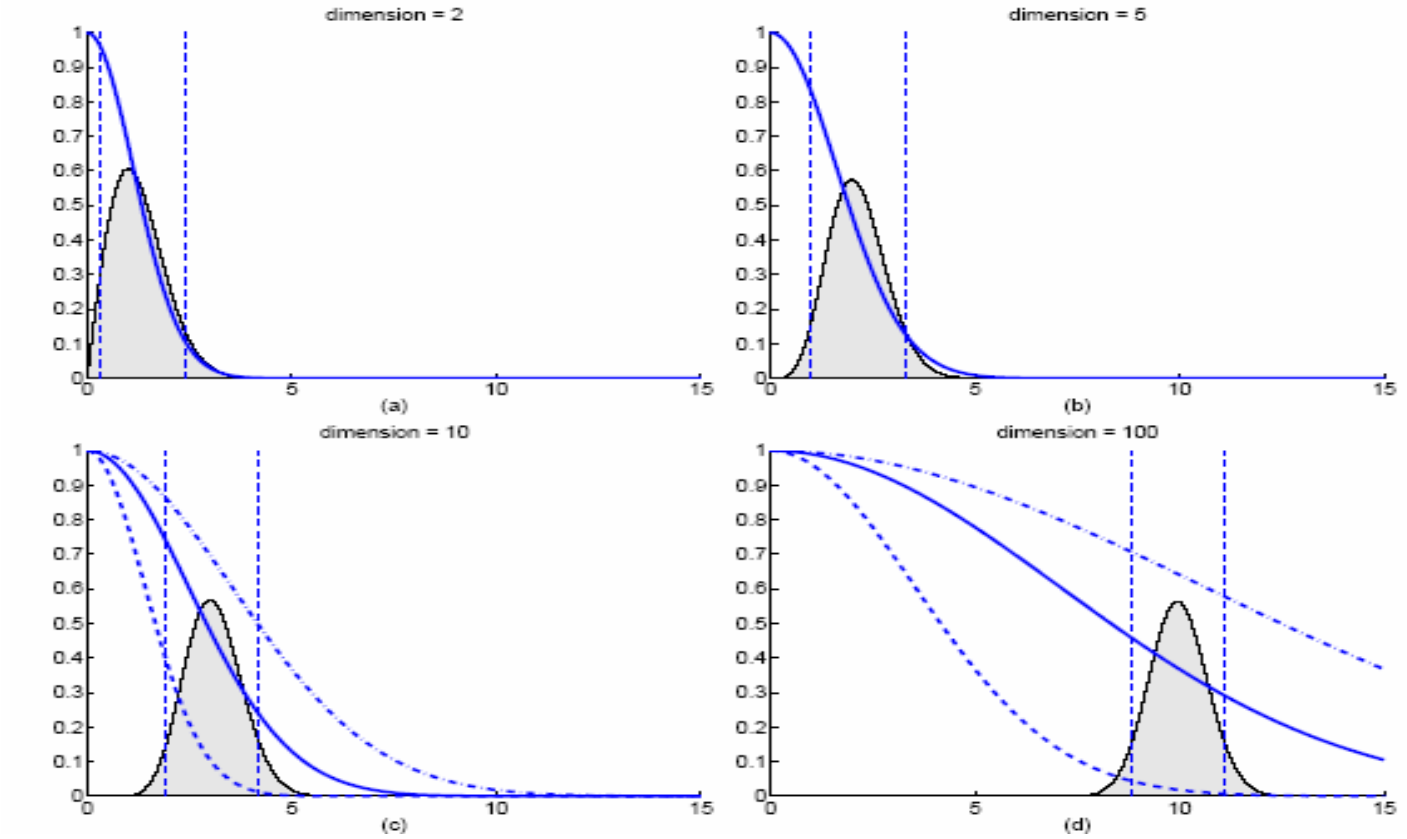
Y: część próbek znajdującej się w kuli o promieniu 1,65

- Większość objętości rozkładu znajduje się w „ogonie” zamiast blisko centrum

Nieintuicyjność HDD

Jądra Gaussowskie

$$K(x, y) = \exp(-\frac{d(x,y)^2}{\sigma^2})$$



Rysunek z [6]

-Punkty mają rozkład normalny z centrum w punkcie C

- X (szary dzwon): rozkład odległości punktów od punktu C

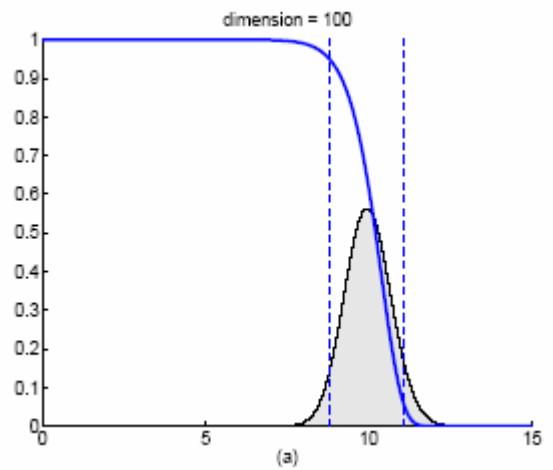
- niebieska linia: wartości jądra Gaussowskiego (z centrum w punkcie C) dla danej odległości od punktu C

- dla HD jądro traci swoje własności odróżniania punktów dalekich od bliskich

Nieintuicyjność HDD

Jądra p-Gaussowskie (cd.) – rozwiązanie problemu (Francois i inni [6])

$$K_p(x, y) = \exp(-\frac{d(x,y)^p}{\sigma^p})$$



Rysunek z [6]

- Lepsze rozróżnianie między punktami bliskimi i dalekimi

Nieintuicyjność HDD

Punkty dalekie i bliskie są blisko siebie

- W pracy [2] (Beyer i inni) pokazano, że dla łagodnych ograniczeń na rozkład danych (sztuczne dane o różnych rozkładach i dane prawdziwe), przy rosnącej wymiarowości odległość do najbliższego punktu zbliża się do odległości do najdalszego (dla kwerendy NN). W tym przypadku znajdowanie k-NN przestaje mieć większy sens
- ten efekt może być widoczny już dla 10-15 wymiarów
- sugerują wykorzystywanie nowego algorytmu zamiast k-NN: epsilon-Radius NN

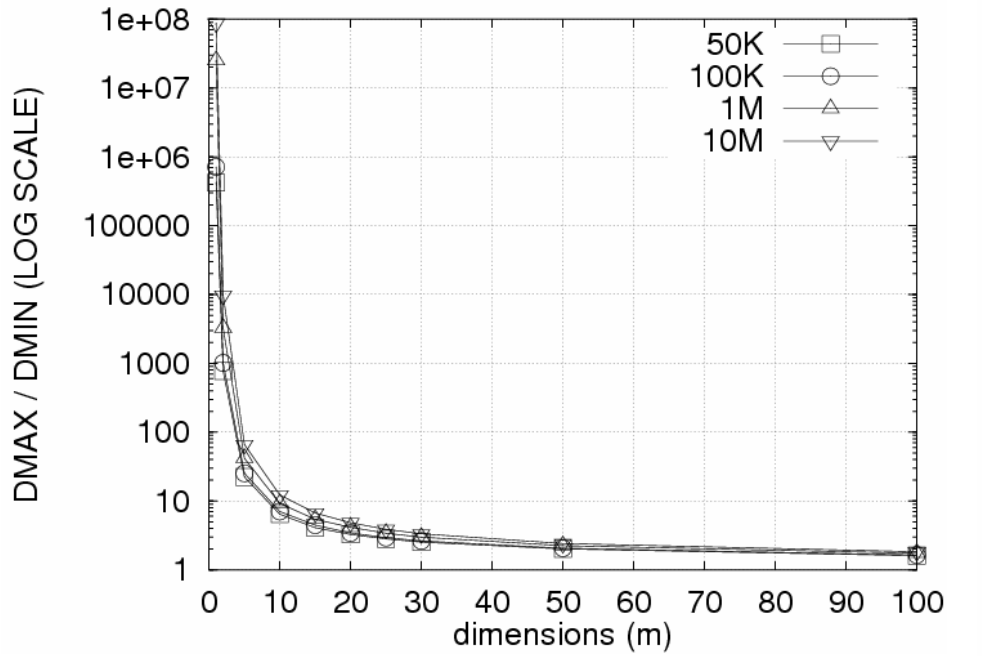
Nieintuicyjność HDD

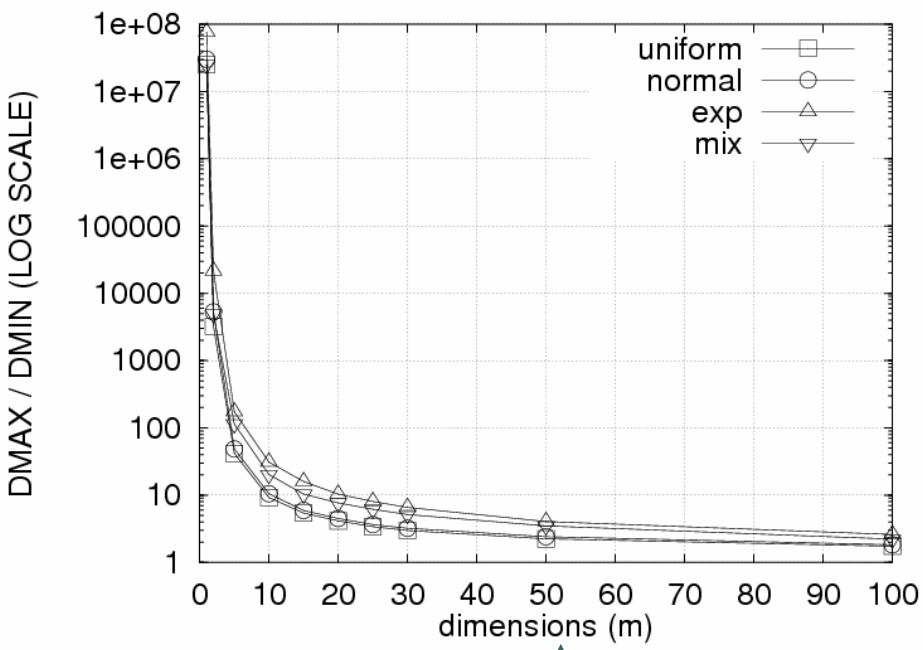
Punkty dalekie i bliskie są blisko siebie (cd.)

DMAX – odległość do najdalszego punktu

DMIN – odległość do najbliższego punktu

Wszędzie wariancja=1





Rysunki z [2]

dla rozkładu jednostajnego

dla miliona punktów

Punkty dalekie i bliskie są blisko siebie – inne podejście[1]

- Zajmijmy się metryką L_k

$$L_k(x, y) = \sum_{i=1}^{d} (\left\| x^i - y^i \right\|^k)^{1/k}$$

- Autorzy twierdzą, że im wyższe *k*, to metryka gorzej się sprawdza w wysokich wymiarach i przedstawiają dowód teoretyczny i wyniki praktyczne potwierdzające tezę
- Więc najlepsza jest metryka dla *k=1* (Manhattan), albo wręcz ułamkowa<1

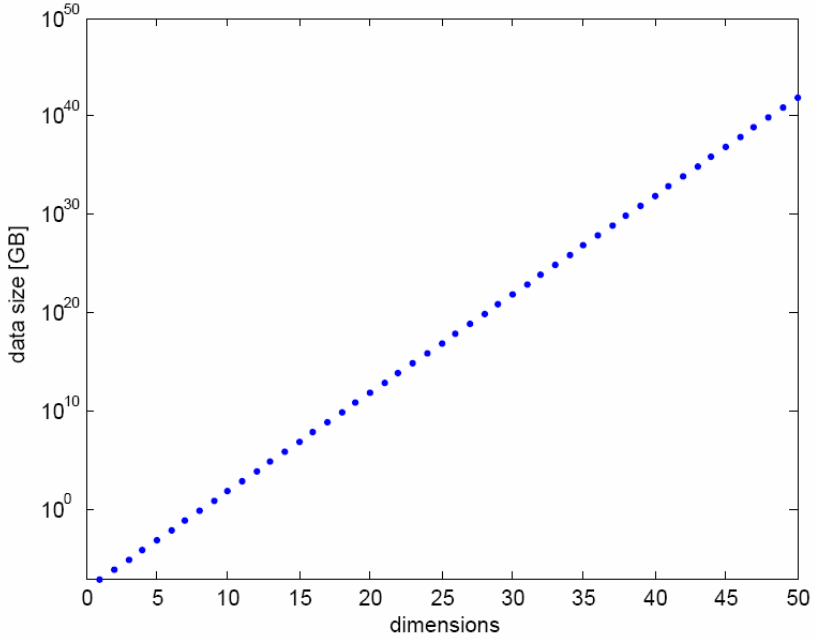
Spis treści

„Klątwa” wymiarowości

- Termin wymyślony prawdopodobnie przez matematyka Richard’a Bellman’a (1920-1984)
- Ogólnie: pojawianie się zjawisk dla HDD, które mają negatywny wpływ na zachowanie i wydajność algorytmów uczących się [12]
- Przykład: aproksymowanie funkcji (tworzenie modelu) na podstawie pomiarów (przykładów) przez algorytm uczący:
  - Zakładamy, że dobra aproksymacja 1-D: 10 punktów
  - Więc dobra aproksymacja dla d-wymiarową: 10^d punktów (duża liczba nawet dla małych d)

Zjawisko pustej przestrzeni

- By wypełnić w miarę gęsto przestrzeń d-wymiarową potrzeba np.10^d punktów
- Warunek niemożliwy do spełnienia dla prawdziwych HDD
- Wniosek: Rzeczywiste HDD w sposób rzadki wypełniają przestrzeń danych – przestrzeń jest „pusta”



Założenia wykresu:

-Każdy punkt to wektor d-wymiarowy

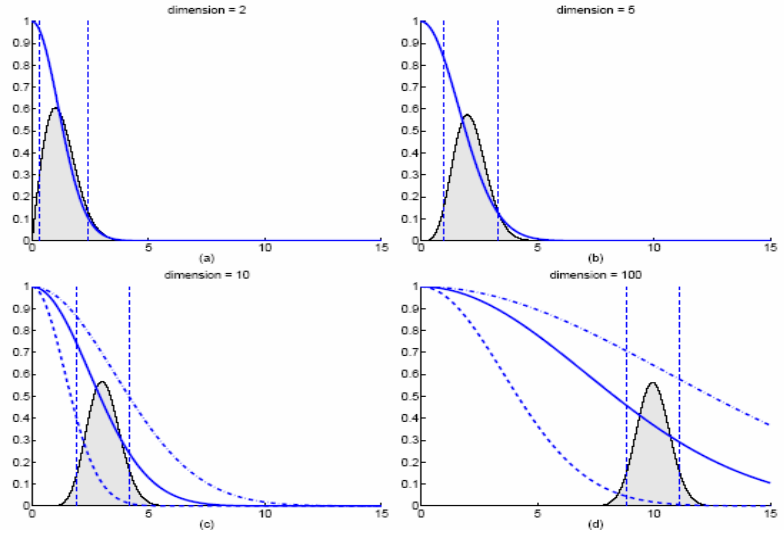
-Każdy z d elementów wektora to 32 bity pamięci

- do aproksymacji funkcji potrzebujemy 10d punktów

-1GB=10^9 B

„Błogosławieństwa” wymiarowości

- Zjawisko „koncentracji miary” – przy łagodnych ograniczeniach na rozkład danych wariancja każdej miary (odległość, norma) jest stała podczas gdy średnia rośnie (dla rosnącej liczby wymiarów d)
  - Konsekwencja: w HD punkty wydają się być znormalizowane



- Samopodobieństwo (własności fraktalne) – część z danych rzeczywistych ma wysoki wymiar nominalny, własności samopodobieństwa i niski wymiar fraktalny [9].
  - np. wydajność przeszukiwania bazy danych (algorytm k-NN) zależy od wymiaru fraktalnego danych (a nie wymiaru nominalnego).

„Reguły kciuka”, n=x*d

- „Reguły kciuka” dotyczące potrzebnej liczby danych (n) dla określonej liczby wymiarów (d):
  - n=5*d, n=10*d (wg. [10])
  - n>10*d (dla zagadnienia feature selection) (wg. [7])

Nie jest tak źle jak się wydaje

- W rzeczywistych danych przy mierzeniu paru zmiennych danego zjawiska (wymiary) część z tych zmiennych jest uzależniona od innych z tych zmiennych
- Może to być wytłumaczeniem, dlaczego ANN potrafią dobrze się uczyć nawet na danych, dla których d~n

Spis treści

Wewnętrzna wymiarowość (Intrinsic Dimensionality (ID))

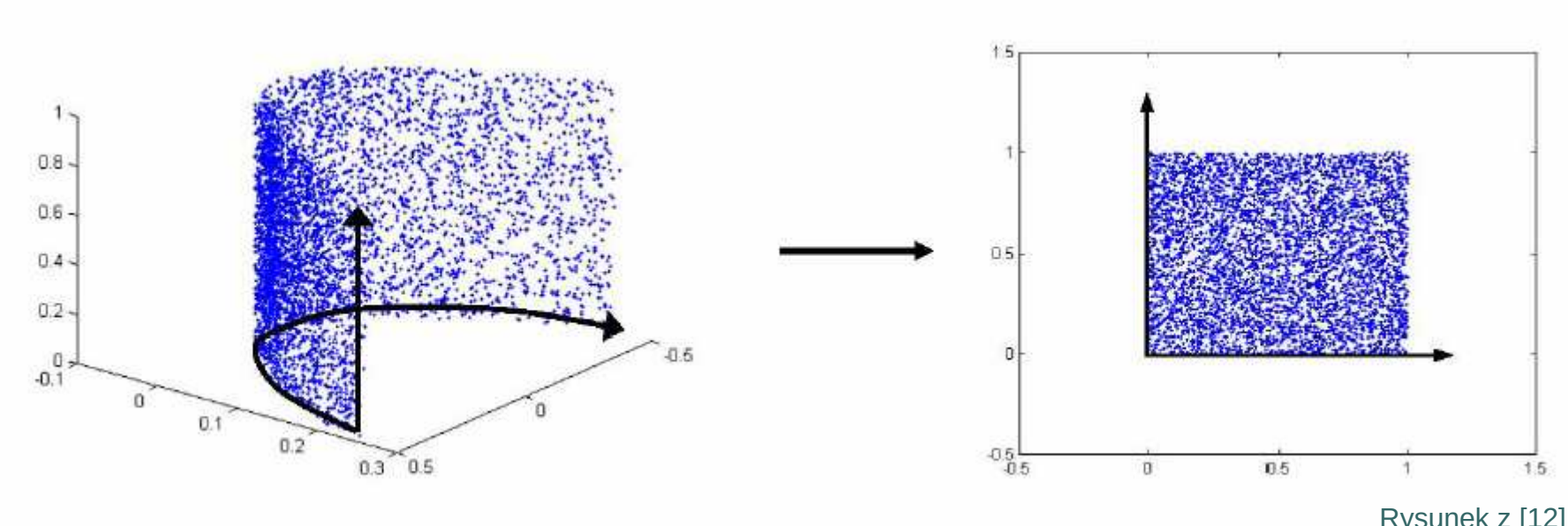
- Dane są reprezentowane jako punkty (wektory) d-wymiarowe ≠> dane „naprawdę” są d-wymiarowe
- ID - minimalna liczba niezależnych zmiennych (stopnie swobody) potrzebnych do reprezentacji danych bez starty informacji, a bardziej ogólnie:
- Zbiór $\Omega \subset R^d$ ma ID = m, gdy jego elementy leżą całkowicie w m-wymiarowej podprzestrzeni R^d (m<d) [3]

Rozmaitość topologiczna (manifold)

- Rozszerzenie na wyższe wymiary pojęcia „powierzchni”
- Definicja [4]: przestrzeń topologiczna *M*, w której dla każdego punktu istnieje otoczenie otwarte, homeomorficzne z przestrzenią kartezjańską R^n dla pewnego $n \in N$
- Liczbę n taką samą dla wszystkich punktów $x \in M$ nazywa się wymiarem rozmaitości (my będziemy go utożsamiać z ID)

Rozmaitości topologiczne - przykłady

- Kula jest rozmaitością topologiczną o wymiarze 2
- Rozkład „podkowa” („horseshoe”)
  - wymiar nominalny=3
  - ID=2

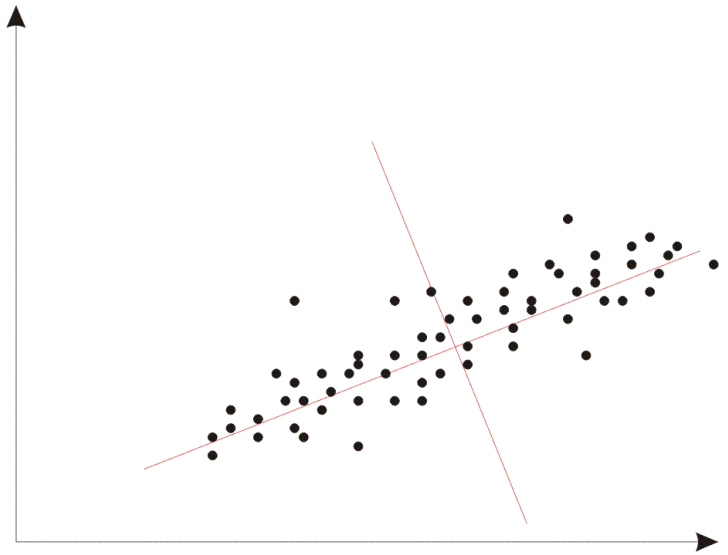


Rysunek z [12]

Spis treści

Analiza Głównych Składowych
(Principal Component Analysis (PCA))

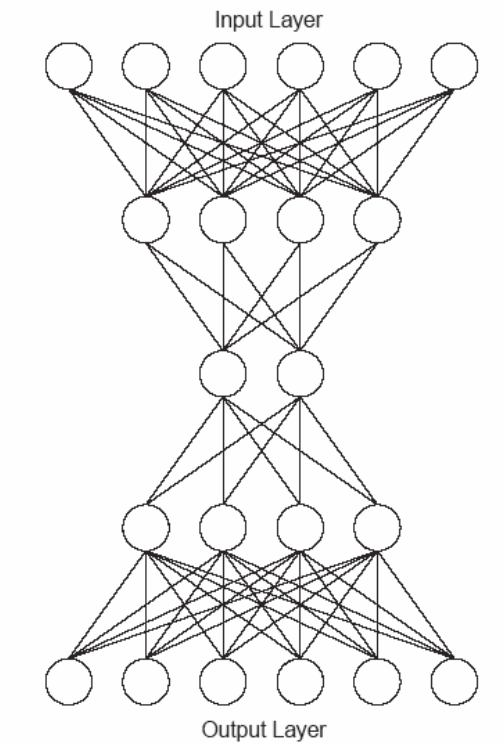
- Klasyczna metoda statystyczna do redukcji wymiarowości
- Za pomocą transformacji liniowych tworzy nowy układ współrzędnych tż:
  - Rzut na 1. oś ma największą wariancję z możliwych rzutów
  - Rzut na 2. oś ma drugą największą wariancję z pozostałych rzutów itd.
- Można „obciąć” końcowe („nieważne” – o małej wariancji) wymiary
- Wady: przeszacowywanie ID, liniowość
- Zalety: obliczeniowo proste, nadaje się do np. wstępnego rzutowania danych



Spis treści

Nieliniowa PCA

- realizowana za pomocą 5-warstwowego MLP (w kształcie klepsydry) - liczba neuronów w środkowej warstwie, to oszacowanie ID
- na wejście i wyjście dajemy taki sam wektor



- uczymy za pomocą backpropagation (najczęściej metoda sprzężonych gradientów)
- Wady:
  - rzuty na krzywe i powierzchnie są suboptymalne,
  - nie może modelować krzywych i powierzchni, które się przecinają
  - ograniczony do prostych problemów ze względu na problemy z doborem parametrów warstw (wg. [12])

Rysunek z [3]

Spis treści

Metody skalowania Wielowymiarowego (MDS)

- Metody MDS (Multi-Dimensional Scaling) – implementują rzutowanie nieliniowe
- Główna idea algorytmów: starają się tak rozmieścić punkty w przestrzeni o niższej wymiarowości, by odległości między parami punktów były jak najlepiej zachowane
- Są to algorytmy iteracyjne

MDS (cd.)

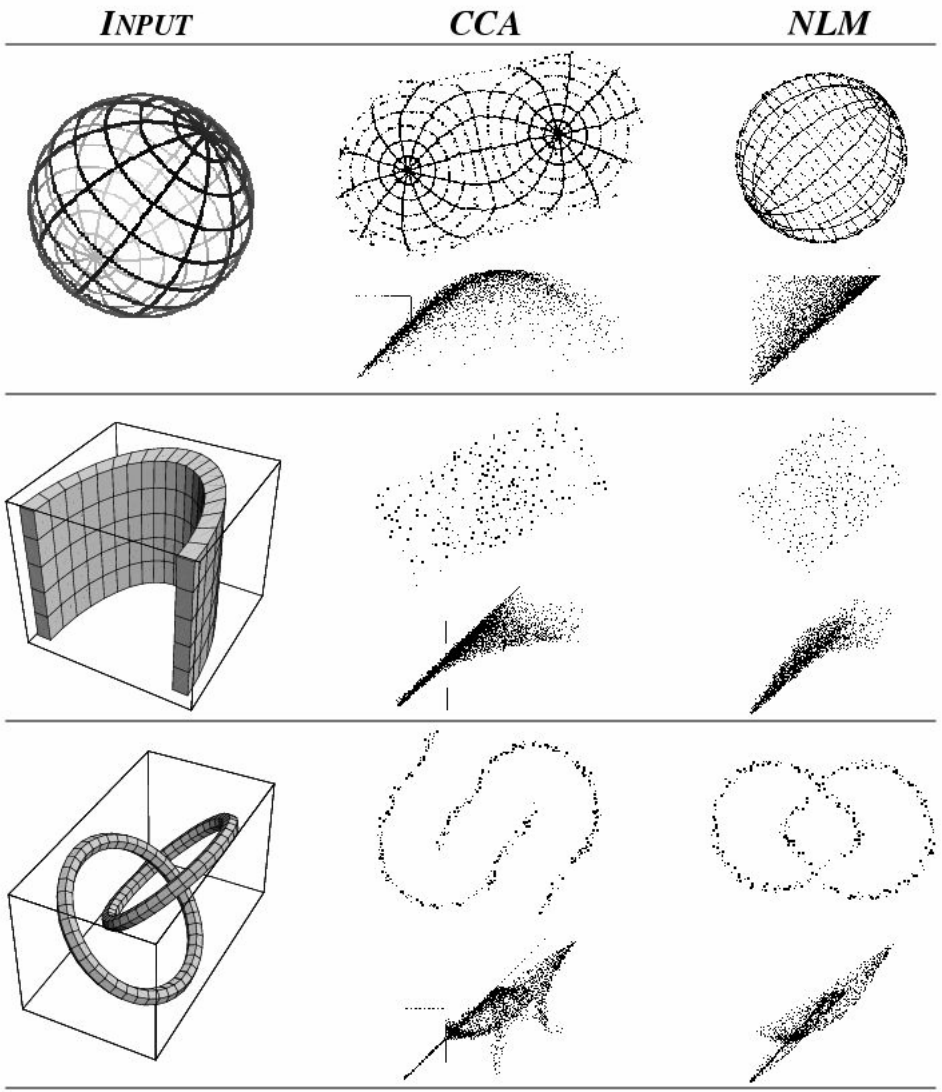
- MDSCAL - klasyczny algorytm tego typu
- Sammon’s mapping – podobny algorytm
- Dla obu:
  - jakość rzutowania określa wzór na naprężenie (stress), który zależy od odległości między danymi.
  - Gdy odległość między danymi zrzutowanymi przypomina odległość wejściową, to naprężenie małe
  - Rozmieszczamy losowo punkty, a następnie nimi poruszamy zgodnie z malejącym gradientem
- By określić wymiarowość:
  - rzutujemy na różną liczbę wymiarów
  - tworzymy wykres naprężenia (końcowego) od wymiaru.
  - Szukamy na wykresie „kolanka” – w tym miejscu określamy ID
- Wady:
  - „kolanko” może nie istnieć
  - Duża złożoność obliczeniowa

MDS (cd.)

- Curvilinear Component Analysis (CCA)
  - Oparty na SOM ANN
  - 2 etapy działania:
    - Kwantowanie wektorowe zbioru danych -> wyłonienie reprezentantów (przy pomocy SOM)
    - Nieliniowe rzutowanie reprezentantów na przestrzeń o niższej liczbie wymiarów (technika analogiczna do innych algorytmów MDS)
  - Zalety:
    - Kwantowanie ułatwia zbieżność (eliminacja szumu), zmniejsza nakład obliczeniowy
  - Wady:
    - Problem z doborem parametrów, problemy z silnie zwiniętymi rozmaitościami

MDS (cd.) - CCA - przykłady

- Rzutowanie wykonane za pomocą CCA
- W kolumnie NLM – Sammon's mapping



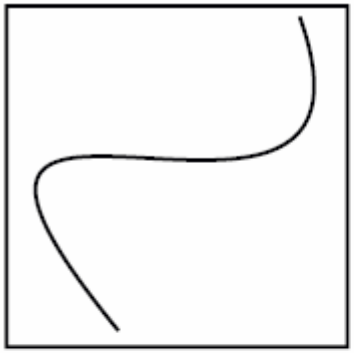
Rysunek z [5]

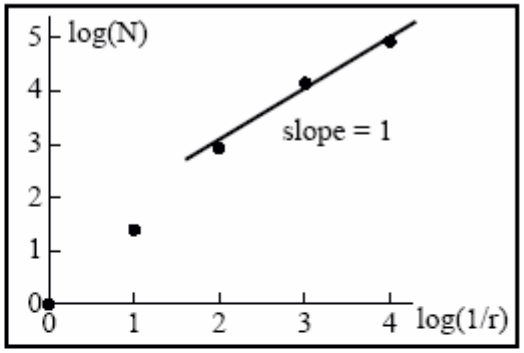
Spis treści

Box Dimension

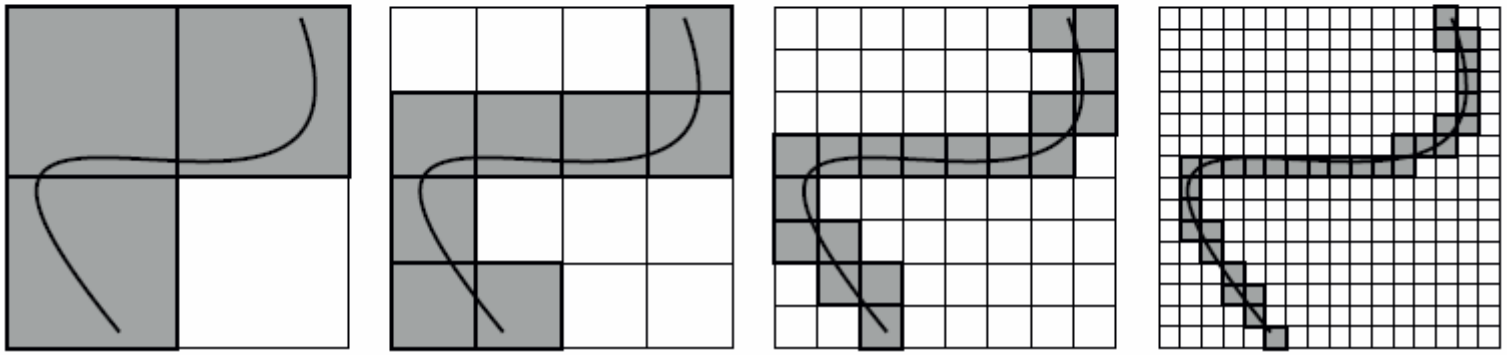
- Dzielimy przestrzeń na hipersześcianiki o boku r
- N(r) z tych sześcianików zawiera co najmniej 1 punkt
- N(r) obliczamy dla kolejnych wartości r
- Szukany wymiar to:

$$D_0 = \lim_{r \to 0} - \frac{\log N(r)}{\log r}$$





- Wady:
  - Duża złożoność obliczeniowa dla dużej liczby punktów i wymiarów



Rysunek z [11]

Correlation dimension

- Dzielimy przestrzeń na hipersześcianiki o boku r
- Obliczamy (dla kolejnych r): $S(r) = \sum_i (p_i)^2$
  - Gdzie p_i – procent punktów, które wpadły do i-tego hipersześcianiku
- Szukany wymiar to:

$$D_2 = \lim_{r \to 0} \frac{\log S(r)}{\log r}$$

Zalety i wady metod fraktalnych:

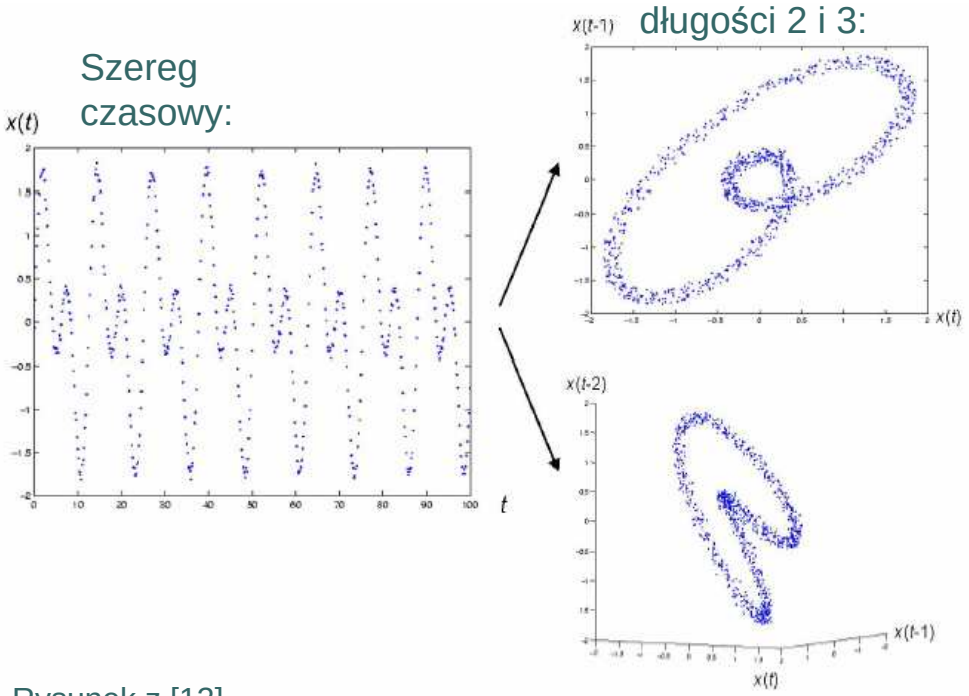
- Zalety [9]:
  - Pokazano, że wydajność przeszukiwania przestrzeni (k-NN) zależy od wymiaru fraktalnego a nie nominalnego
  - Pokazano tam parę rzeczywistych zbiorów danych, które:
    - mają dużą (nominalną) liczbę wymiarów
    - mają własność samopodobieństwa
    - mają niski wymiar fraktalny
  - podają uzasadnienie teoretyczne i praktyczne, że tego typu zbiory danych są uodpornione na klątwę wymiarowości
- Wady [3]:
  - zostało udowodnione, że by dokładnie określić wymiarowość d-wymiarowego zbioru, potrzeba conajmniej $10^{d/2}$ punktów
    - (można temu próbować zaradzić zwiększając liczbę punktów metodą surrogate data)

Wykorzystanie ID

- feature extractors - na przykład ustalamy ile jest neuronów w warstwie ukrytej w MLP ANN
- do charakteryzacji ludzkich twarzy, wymiar fraktalny obrazka może służyć do jego klasyfikacji
- w szeregach czasowych - do ustalania rzędu modelu (na podstawie tw. Teken'a)

Twierdzenie Teken'a

- Jeśli: q=ID regresorów (regressors)



- Wtedy:
  - liczba regresorów wystarczająca do przewidzenia szeregu jest w przedziale [q, 2q-1]
  - Regresory z przestrzeni 2q-1- wymiarowej mogą być zrzutowane bez utraty informacji do przestrzeni q-wymiarowej

Rysunek z [12]

Podsumowanie metod szacowania ID

- na topie są są metody fraktalne i zastosowanie ANN, rzutowanie nieliniowe
- nadal nie ma algorytmów, które by szacowały ID dla dużego wymiaru i małej liczby danych

Spis treści

Success story [11]

- Predykcja wskazań indeksu giełdowego BEL20
- Wejście: 42 wskaźniki finansowe
- Przebieg eksperymentu:
  - Zastosowanie PCA – przy zachowaniu 95% wariancji redukcja do 25 wymiarów
  - Szacowanie ID za pomocą metody Grosberg'a-Procaccia (ID=9)
  - Użycie CCA do zrzutowania danych do przestrzeni 9-D
  - Użycie sieci radialnych (RBFN) do przewidywania zwrotu z następnego dnia
- Wynik: 57% poprawnego przewidywania czy indeks wzrośnie czy zmaleje

Literatura

- [1] C.C. Aggarwal , A. Hinneburg , D.A. Keim, *On the Surprising Behavior of Distance Metrics in High Dimensional Spaces*, Proceedings of the 8th International Conference on Database Theory, str.420-434, January 04-06, 2001
- [2] K.S. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, *When is "Nearest Neighbor" meaningful?*, Proceeding of the 7th International Conference on Database Theory, str. 217 – 235, 1999
- [3] F. Camastra, *Data dimensionality estimation methods: a survey*, Pattern Recognition, volume: 36, issue: 12, str 2945 – 2954, 2003
- [4] K. Cegiełka, E. Stachowski, K. Szymański, *Encyklopedia dla wszystkich. Matematyka,* WNT, 2000
- [5] P. Demartines, J. Herault, *Curvilinear Component Analysis: a Self-Organizing Neural Network for Nonlinear Mapping*, IEEE Transactions on Neural Networks, vol. 8, no. 1, str. 148-154, 1997
- [6] D. Francois, V. Wertz, M. Verleysen , *About the locality of kernels in high-dimensional spaces*, International Symposium on Applied Stochastic Models and Data Analysis ASMDA 2005, 2005
- [7] A.K. Jain, R.P.W. Duin, Jianchang Mao, *Statistical Pattern Recognition: A Review*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, str. 4-37, Jan., 2000
- [8] M. Kaku, *Hiperprzestrzeń – wszechświaty równoległe, pętle czasowe i dziesiąty wymiar*, Prószyński i S-ka,1996

Literatura

- [9] F. Korn, Bernd-Uwe Panel, C. Faloutsos, *On the 'Dimensionality Curse' and the 'Self-Similarity Blessing*', IEEE Trans. Knowl. Data Eng. 13(1), str. 96-111, 2001
- [10] C. T. Leondes (ed.), *Image processing and Pattern Recognition*, z serii Neural Network Systems Techniques and Applications, Harcourt Publishers Ltd, 1998
- [11] M. Verleysen, *Learning high-dimensional data*, LFTNC'2001 - NATO Advanced Research Workshop on Limitations and Future Trends in Neural Computing, 2001
- [12] M. Verleysen, D. François**,** *The Curse of Dimensionality in Data Mining and Time Series Prediction*, 8th International Workshop on Artificial Neural Networks, IWANN 2005, str. 758-770, 2005

Dziękuję za uwagę!