Wyzwania motywujące głębokie uczenie maszynowe – przekleństwo wymiarowości

Weronika Sieińska

16 października 2018

Przekleństwo wymiarowości

Jak można zdefiniować przekleństwo wymiarowości?

Zjawisko polegające na tym, że wraz ze wzrostem wymiarowości danych obserwujemy również spadek ich gęstości.

Przekleństwo wymiarowości na przykładzie klasyfikacji

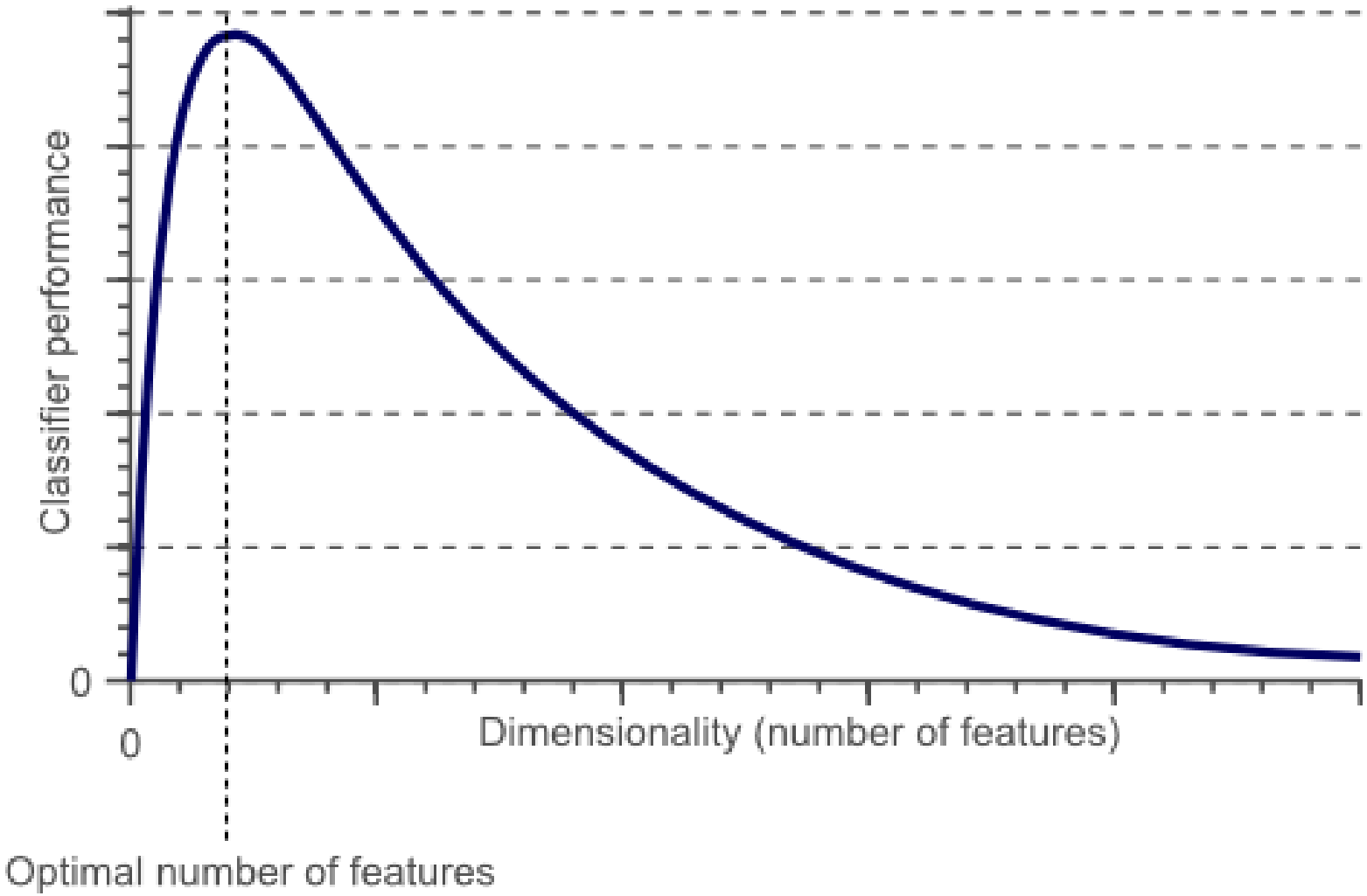
Przykład

- zbiór uczący zawierający 10 obrazków
- każdy obrazek przedstawia psa lub kota
- klasyfikator
- cechy (wymiary)
  - średnia wartość koloru czerwonego
  - średnia wartość koloru zielongo
  - średnia wartość koloru niebieskiego
- możliwe działanie klasyfikatora:

if $0.5 \times red + 0.3 \times green + 0.2 \times blue > 0.6$ **then**
    return *cat*
else
    return *dog*
end if

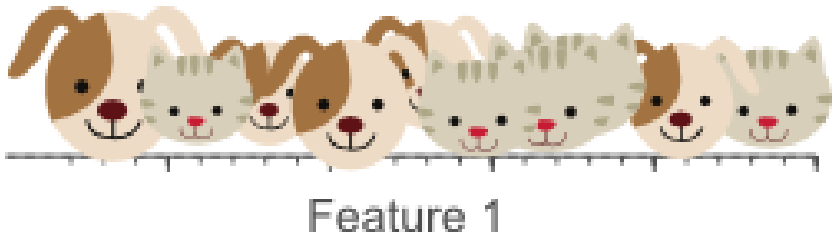
Przekleństwo wymiarowości na przykładzie klasyfikacji

- 3 cechy mogą nie wystarczyć do uzyskania satysfakcjonującego podziału na klasy
- Może dodać nowe cechy?
  - tekstura
  - średnia intensywność krawędzi
  - cechy na podstawie histogramu obrazu (opisu statystycznego wartości obrazu (jasność / intensywność))
  - itp.
- Lepiej nie! Od pewnego momentu zwiększanie liczby wymiarów danych może prowadzić do obniżenia wydajności klasyfikatora.

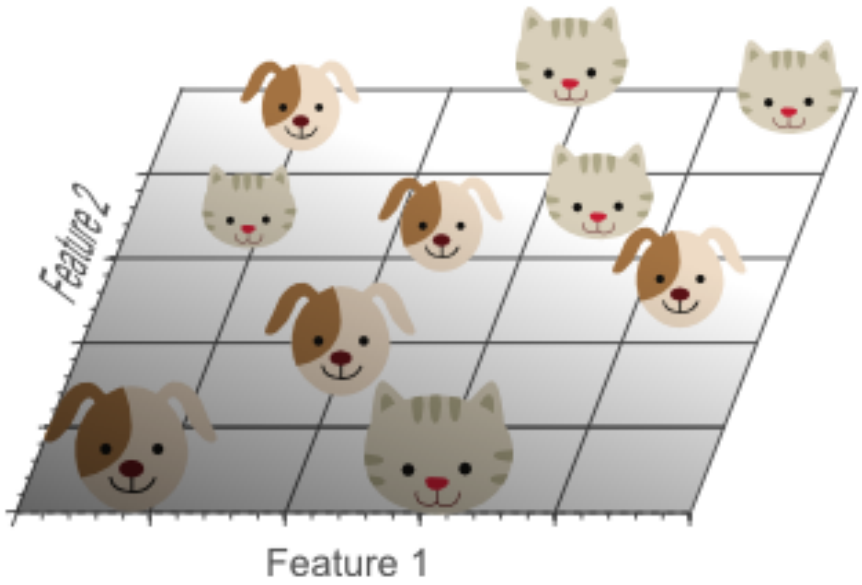


Rysunek 1: źródło:
`http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`

- Wróćmy do mniejszej liczby cech np. jednej:
  - średnia wartość koloru czerwonego
- 1D
- Na rysunku widać, że nie otrzymaliśmy podziału przykładów na klasy.

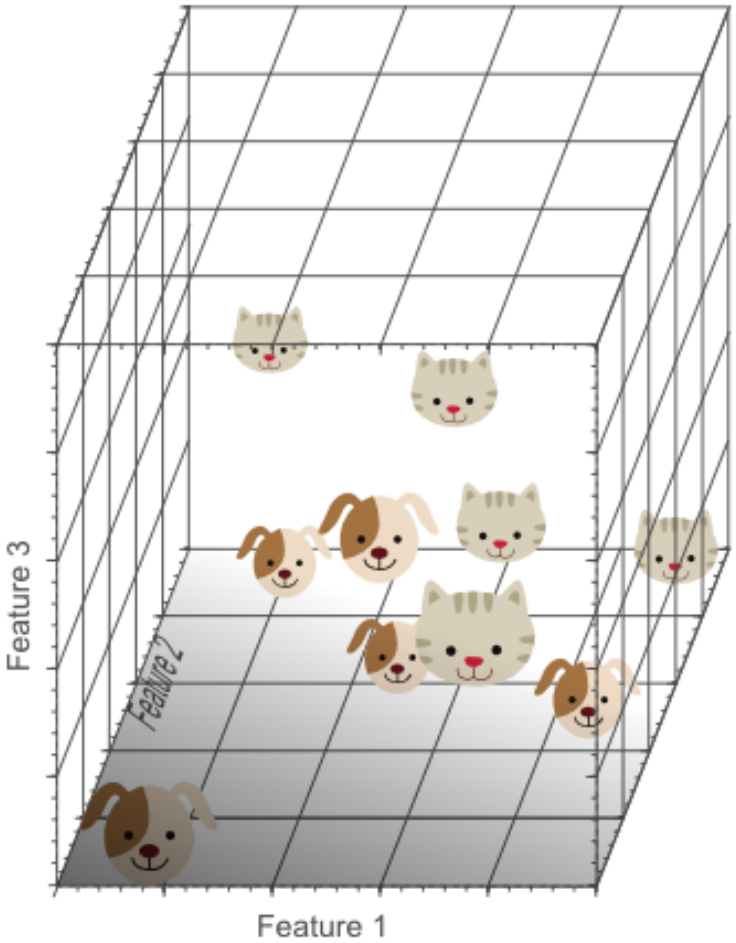


Rysunek 2: źródło
`http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`



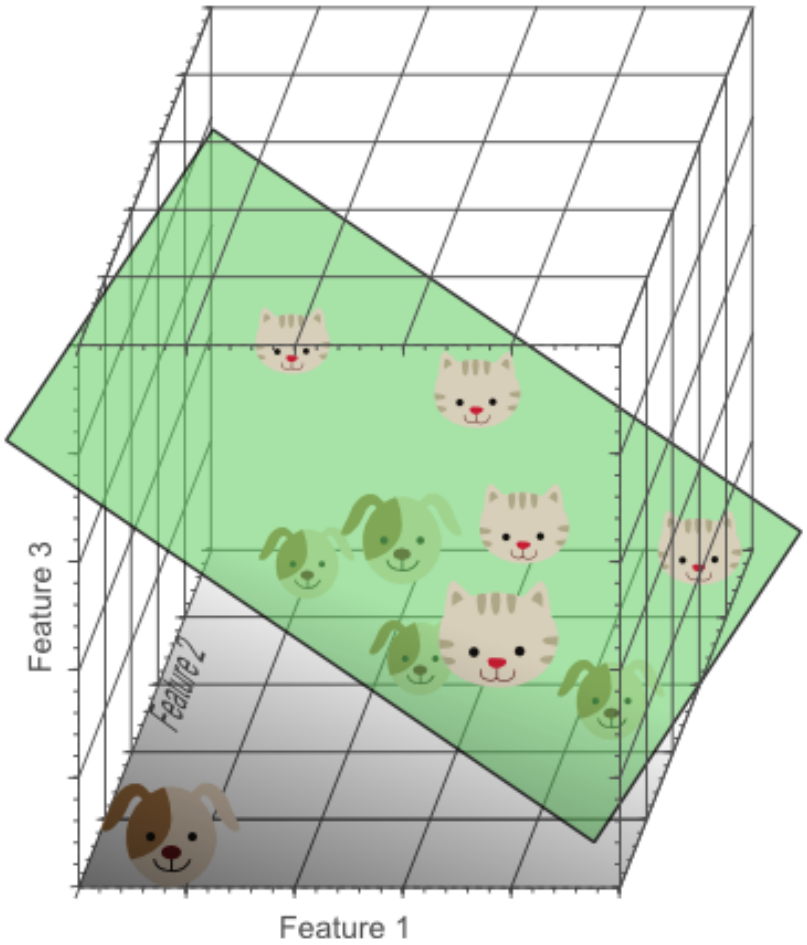
Rysunek 3: źródło:
`http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`

- Dodajmy kolejną cechę
  - średnia wartość koloru czerwonego
  - średnia wartość koloru zielonego
- 2D
- Nadal nie otrzymujemy widocznego podziału.
- Nie ma możliwości przeprowadzenia linii, która idealnie rozdzieli przykłady na interesujące nas klasy.



Rysunek 4: źródło:
http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

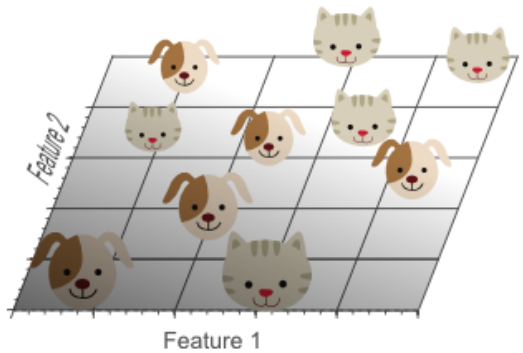
- Dodajmy kolejną cechę
  - średnia wartość koloru czerwonego
  - średnia wartość koloru zielonego
  - średnia wartość koloru niebieskiego
- 3D
- Możemy zaobserwować, że istnieje płaszczyzna doskonale rozdzielająca przykłady z naszego zbioru uczącego.

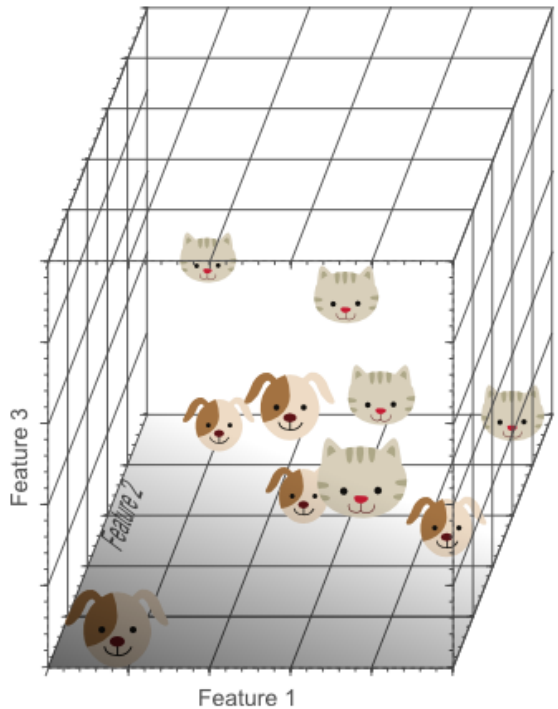


Rysunek 5: źródło:
http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

- Dodajmy kolejną cechę
  - średnia wartość koloru czerwonego
  - średnia wartość koloru zielonego
  - średnia wartość koloru niebieskiego
- 3D
- Możemy zaobserwować, że istnieje płaszczyzna doskonale rozdzielająca przykłady z naszego zbioru uczącego.

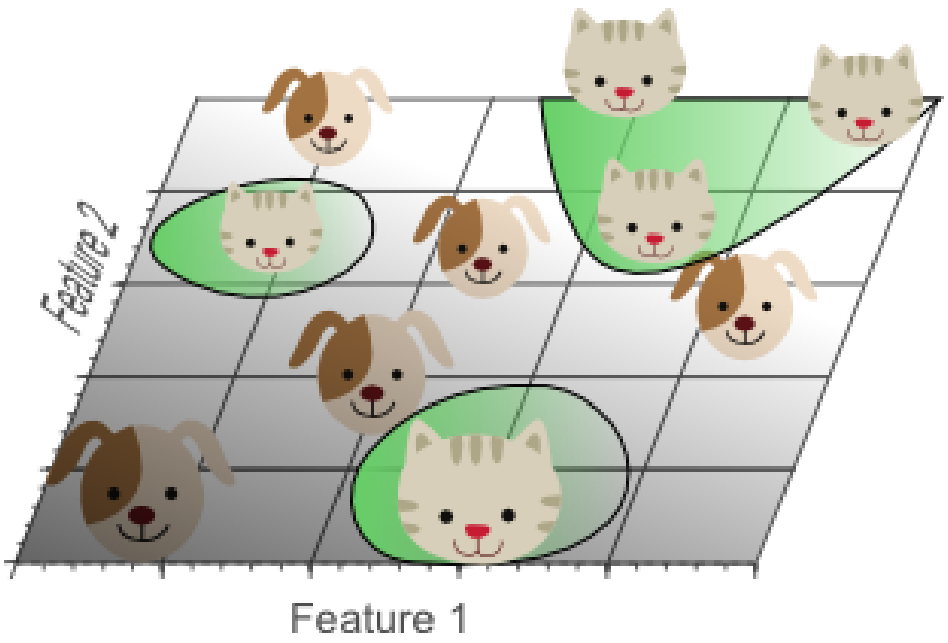
- Dalsze zwiększanie liczby wymiarów doprowadziłoby do obniżenia wydajności klasyfikatora z powodu zmniejszenia gęstości danych.
  - 1D: $10/5 = 2$ (2 przykłady na przedział)
  - 2D: $10/(5 \times 5) = 0.4$ (0.4 przykładu na kwadrat)
  - 3D: $10/(5 \times 5 \times 5) = 0.08$ (0.08 przykładu na sześcian)





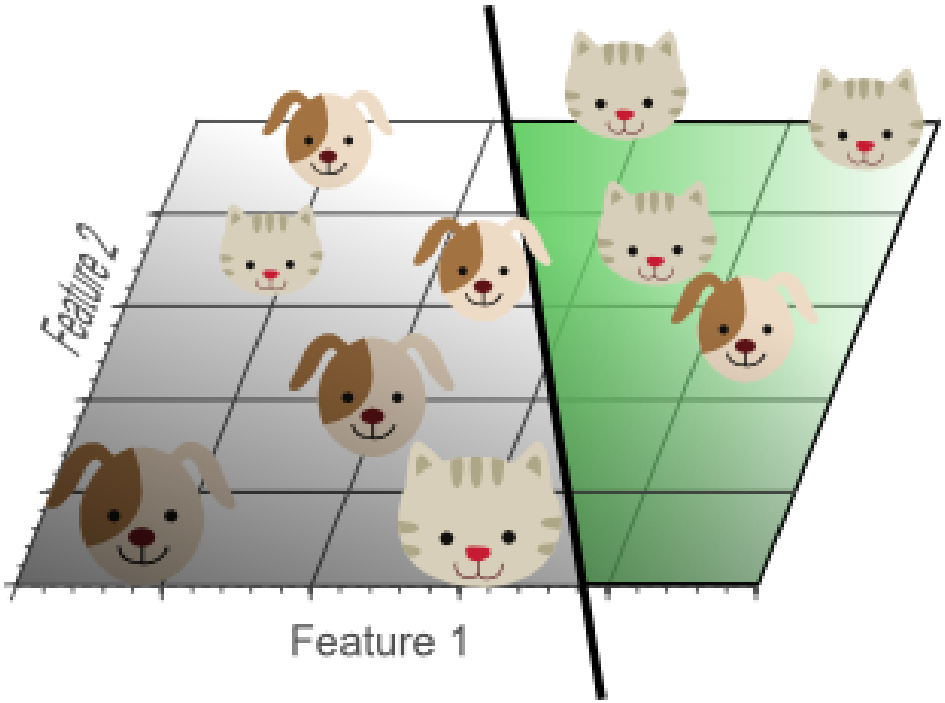
źródło:
`http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`

- Ale dlaczego rzadkość danych jest problemem? Jeśli mamy niewiele przykładów rozlokowanych w dużej przestrzeni powinno być łatwo dopasować hiperpłaszczyznę, która rozdzieli dane.
- Tak, jest łatwo, ZBYT łatwo.
- Zobaczmy więc jak wygląda wynik klasyfikacji w przestrzeni 3D zrzutowany na płaszczyznę 2D...



Rysunek 6: źródło:
`http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`

- Rysunek pokazuje zjawisko przetrenowania klasyfikatora polegające na nadmiernym dopasowaniu do danych uczących.
- Tak wytrenowany klasyfikator nie będzie potrafił generalizować na nowe przykłady ze zbioru testowego.
- Nadmierne dopasowanie jest bezpośrednim następstwem problemu przekleństwa wymiarowości.



Rysunek 7: źródło:
`http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`

- Zobaczmy jak wyglądałby wynik klasyfikacji przeprowadzonej przez klasyfikator wytrenowany na 2 cechach zamiast 3.
- Podział na zbiorze uczącym nie jest idealny, jednak ten klasyfikator osiągnie lepsze wyniki na nieznanych sobie przykładach ponieważ nie wyuczył się konkretnych zależności, które przypadkowo były obecne w zbiorze uczącym.
- W ten sposób uniknęliśmy przekleństwa wymiarowości oraz przetrenowania klasyfikatora.

Inne sposoby redukcji wymiarowości

- wybór najbardziej informatywnych cech (feature selection)
- zbudowanie nowego zbioru cech na podstawie istniejącego zbioru, gdzie nowe cechy są kombinacją cech początkowych, a także informacje w nich zawarte nie są skorelowane (feature extraction)
- podział zbioru przykładów na podzbiory, przeprowadzenie uczenia na niektórych z nich, a następnie wykorzystanie reszty do potwierdzenia wiarygodności uzyskanych wyników (cross-validation / walidacja krzyżowa)
- ...

Jak się to robi w głębokim uczeniu maszynowym?

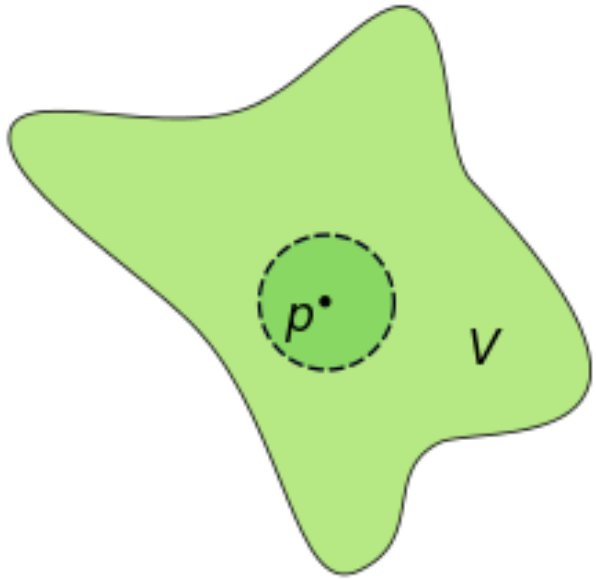
Rozmaitość (*Wikipedia*)

Obiekt geometryczny, który lokalnie ma strukturę przestrzeni R^n (przestrzeni euklidesowej). Pojęcie to uogólnia na dowolną liczbę wymiarów pojęcia krzywej i powierzchni.

- Przestrzeń topologiczna jest lokalnie euklidesowa, gdy otoczenie każdego jej punktu można przekształcić w jakiś podzbiór przestrzeni euklidesowej (n-tego wymiaru) przez rozciąganie, ściskanie, lub skręcanie.
- Np. fragment sfery można przekształcić we fragment płaszczyzny za pomocą odpowiedniej deformacji.

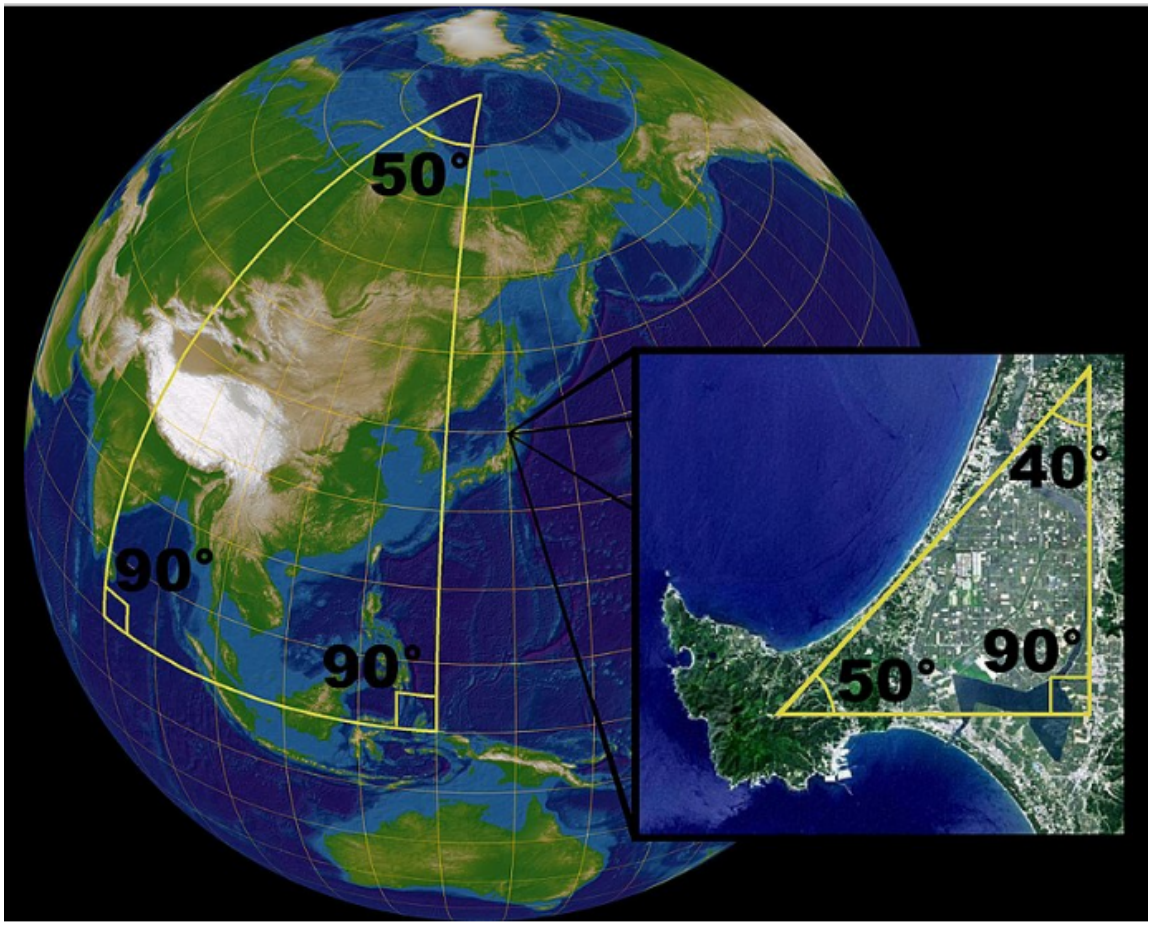
Otoczenie punktu (*Wikipedia*)

Dowolny zbiór, który zawiera zbiór otwarty zawierający dany punkt.



Rysunek 8: źródło: `https://pl.wikipedia.org/wiki/Otoczenie_(matematyka)`

Zbiór V na płaszczyźnie jest otoczeniem punktu p jeżeli istnieje koło (bez brzegu) zawierające p i zawarte w V.



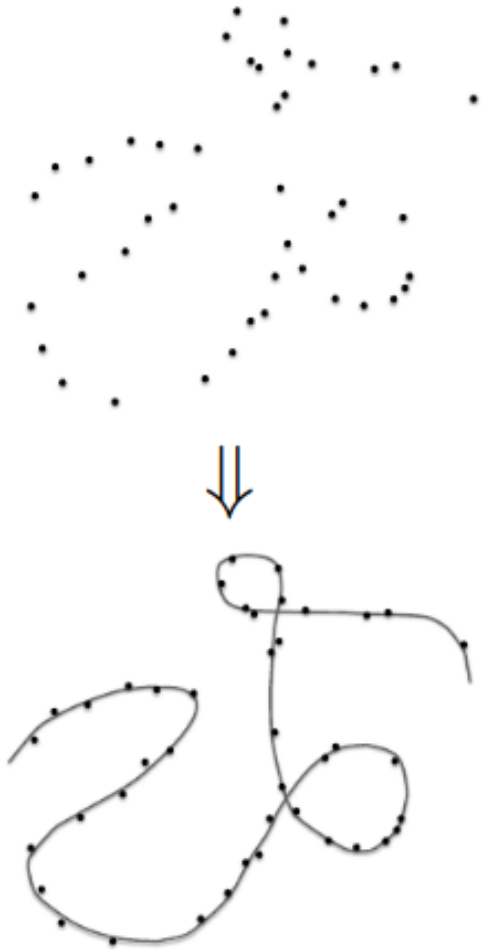
Rysunek 9: źródło: `https://pl.wikipedia.org/wiki/Rozmaito%C5%9B%C4%87`

- Sfera to dwuwymiarowa rozmaitość:
  - w dużej skali: geometria nieeuklidesowa (suma kątów dużego trójkąta $> 180°$)
  - lokalnie: geometria euklidesowa (suma kątów małego trójkąta $= 180°$)

- Powierzchnia Ziemi jest 2-wymiarową rozmaitością 'owiniętą' wokół sfery w 3D.
- Ziemia istnieje w przestrzeni trójwymiarowej, więc moglibyśmy lokalizacje takie jak miasta opisywać za pomocą 3 wymiarów (cech). Jednak nie mamy problemu z wykorzystaniem jedynie dwóch wymiarów (długości i szerokości geograficznej).

Rysunek 10: źródło: `https://medium.freecodecamp.org/the-curse-of-dimensionality-how-we-can-save-big-data-from-itself-d9fa0f872335`

- Rysunek obok przedstawia zbiór punktów (przykładów uczących) w przestrzeni wielowymiarowej (2D dla wizualizacji).
- Drogi są rozmaitościami 1D zagnieżdżonymi w przestrzeni 3D. Punktami w tej rozmaitości są adresy pojedynczych domów wzdłuż drogi.
- Oczywiście rozmaitości mogą być bardziej złożone oraz obejmować więcej wymiarów niż w przytoczonych przykładach.



Rysunek 11: źródło: `http://www.deeplearningbook.org/version-2015-10-03/contents/manifolds.html`

Hipoteza rozmaitości

Dane o wielu wymiarach w rzeczywistości leżą na rozmaitości o mniejszej liczbie wymiarów osadzonej w przestrzeni wielowymiarowej.

- Problemem jest tutaj znalezienie rozmaitości w przestrzeni wielowymiarowej.
- Jest to przedmiotem eksploracji metod głębokiego uczenia maszynowego.
- Wyciągnięcie współrzędnych rozmaitości stanowi wyzwanie, ale pozwala liczyć na poprawę wielu algorytmów dla systemów uczących się.

Bibliografia

- Ian Goodfellow, Yoshua Bengio, Aaron Courville. *Deep Learning. Systemy uczące się.* Wydawnictwo Naukowe PWN S.A. 2016.
- The Curse of Dimensionality, https://medium.freecodecamp.org/the-curse-of-dimensionality-how-we-can-save-big-data-\from-itself-d9fa0f872335
- The Curse of Dimensionality in classification, http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/
- The Manifold Perspective on Representation Learning, http://www.deeplearningbook.org/version-2015-10-03/contents/manifolds.html
- Histogram obrazu, http://analizaobrazu.x25.pl/articles/12
- Feature extraction, https://en.wikipedia.org/wiki/Feature_extraction
- Walidacja krzyżowa, https://pl.wikipedia.org/wiki/Sprawdzian_krzy%C5%BCowy
- Rozmaitość, https://pl.wikipedia.org/wiki/Rozmaito%C5%9B%C4%87
- Otoczenie punktu, https://pl.wikipedia.org/wiki/Otoczenie_(matematyka)