Klątwa wielowymiarowości
The Curse of Dimensionality

Agnieszka Pocha
Michał Kowalik

11 marca 2015

na podstawie książki:
Bertrand Clarke, Ernest Fokoue, Hao Helen Zhang
Principles and Theory for Data Mining and Machine Learning

Agenda

Sztuczna Inteligencja - AI

W świecie, gdzie niepewność modelu jest często ograniczeniem w procedurach wnioskowania, ważniejszym stała się predykcja/przewidywanie niż testowanie czy estymacja....
Dział informatyki zajmujący się rozwiązywaniem problemów, które nie są efektywnie algorytmizowalne

Uczenie Maszynowe - Machine Learning

Pojęcie odnosi się do użycia formalnych struktur (maszyny) do wnioskowania (uczenie) - MODELOWANIE. Informacja tutaj pomaga zmniejszyć niepewność.

Data Mining

Odnosi się do przeszukiwania ogromnych, wielowymiarowych, wielotypowych zbiorów danych. Dane są nieustrukturyzowane i wielorakie.

Przestrzeń

Przestrzeń – zbiór, w którym określone są rozmaite relacje i działania pomiędzy jego elementami

Metryka

Metryką (w zbiorze X) nazywa się funkcję:
$d : X \times X \rightarrow [0, +\infty)$, która dla dowolnych elementów a, b, c tego zbioru spełnia następujące warunki:

- identyczność nierozróżnialnych: $d(a, b) = 0 \iff a = b$
- symetria: $d(a, b) = d(b, a)$
- warunek trójkąta: $d(a, b) \leqslant d(a, c) + d(c, b)$

Gdy d jest metryką w zbiorze X, to para (X, d) nazywana jest przestrzenią metryczną

Metryka Euklidesowa

Ogólnie, w przestrzeni $\mathbb{R}^n$ metrykę euklidesową definiuje się wzorem:

$$d_e(\mathbf{x}, \mathbf{y}) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2}$$

tzn. jako pierwiastek euklidesowego iloczynu skalarnego różnicy dwóch wektorów przez siebie:

$$d_e(\mathbf{x}, \mathbf{y}) = \sqrt{\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle}$$

Metryka Jaccarda

Metryka używana do porównywania zbiorów: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Żeby były spełnione warunki metryki:

$$d_J(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Local Methods

- K-means

Global Methods

- sieci neuronowe

'Definicja' - Curse of Dimensionality

Rzeczywisty wymiar modelu jest skończony, ale rozmiar przestrzeni w której się znajduje może być nieograniczony.
Trudność szacowania rośnie w sposób wykładniczy względem wymiaru.

Ekstremalny przypadek problem jest, gdy duże p, małe n, gdzie p - rozmiar przestrzeni, n - ilość danych.

Intuicja

Przy wysokim wymiarze przestrzeni, dane są zbyt rzadkie.
Przy wysokim wymiarze przestrzeni, liczba możliwych modeli do rozważenia rośnie w sposób wykładniczy (superexponential).

Sparsity

Concurvity

Liczba modeli

Motywacja

Jeśli nie ma zbyt wiele obiektów do porównania w sąsiedztwie jakiegoś punktu x, wtedy trudno określić jak powinna wyglądać funkcja $f(x)$.

Gdy liczba wymiarów p rośnie, liczba danych lokalnych maleje do 0.

Objętość kuli o promieniu r maleje do 0, wraz ze wzrostem wymiaru p.

$$V_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \cdot r^n = \begin{cases} \dfrac{\pi^k}{k!} \cdot r^n & \text{dla } n = 2k, \\ \dfrac{2^k \pi^{k-1}}{n!!} \cdot r^n & \text{dla } n = 2k-1, \end{cases}$$

Concurvity

Liczba modeli

PCA

LDA