

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

- 1 AI, ML, Data Mining
- 2 Metody Local vs Global
- 3 Klątwa wielowymiarowości
- 4 Rzadkość
- 5 Liczba modeli
- 6 Metody radzenia sobie z problemem

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.

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.

Sztuczna Inteligencja - AI

W świecie, gdzie niepewność modelu jest często ograniczeniem w procedurach wnioskowania, ważniejszą stała się predykcja/przewidywanie niż testowanie czy estymacja....

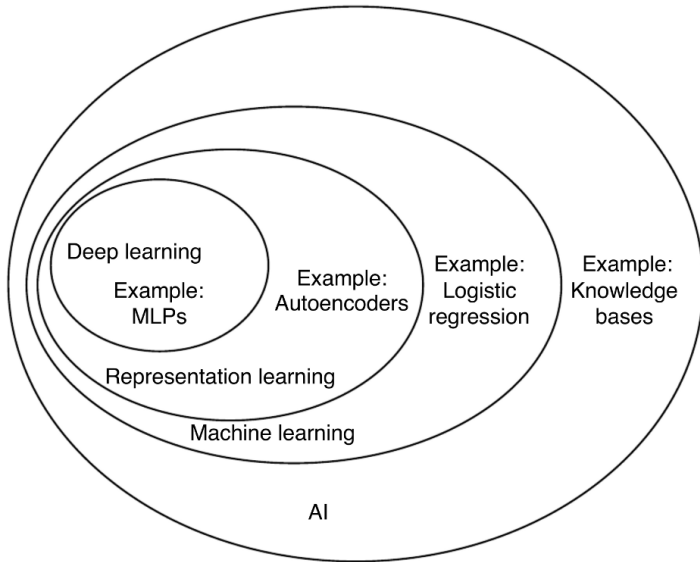
Dział informatyki zajmujący się rozwiązywaniem problemów, które nie są efektywnie algorytmizowalne.

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. Informacje są nieustrukturyzowane i wielorakie.



Model

- Odwzorowanie rzeczywistości
- $s = v \cdot t$

Model

- Odzworowanie rzeczywistości
- $s = v \cdot t$

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: $d(a, b) \leq d(a, c) + d(c, b)$

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:

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

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 Euklidesowa

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

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

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

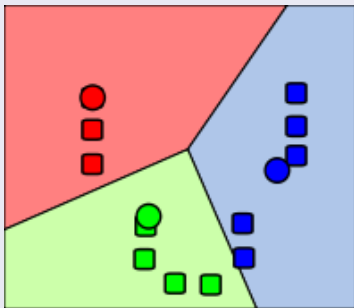
Metryka używana do porównywania zbiorów: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Żeby były spełnione warunki metryki:

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

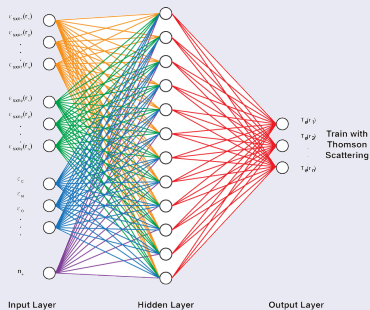
Local Methods

- K-means



Global Methods

- sieci neuronowe



Intuicja

Przy wysokim wymiarze przestrzeni, dane są zbyt rzadkie.

Przy wysokim wymiarze przestrzeni, liczba możliwych modeli do rozważenia rośnie w sposób super-wykładniczy (superexponential).

Z ekstremalnym przypadkiem mamy do czynienia, gdy przestrzeń ma wiele wymiarów p , a danych jest niewiele.

Wprowadzenie pojęć

- Rzadkość (Sparsity)
- Liczba możliwych modeli
- Concurvity

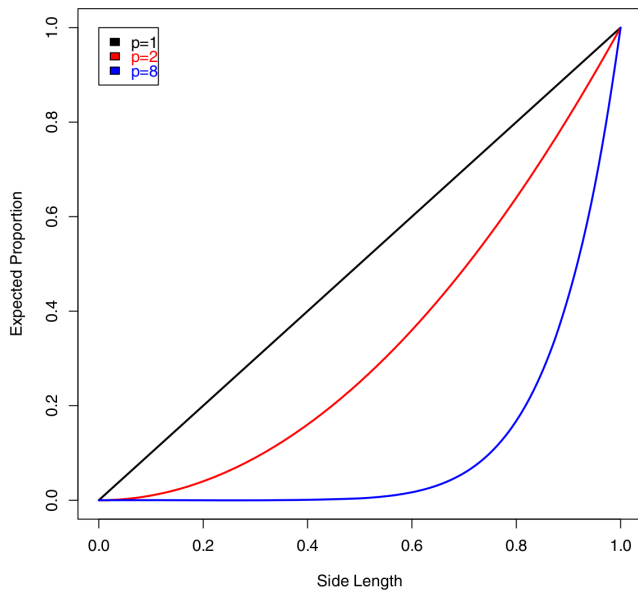
Motywacja

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 .

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



Liczba modeli rośnie super-wykładniczo (superexponential) wraz ze wzrostem rozmiaru.

Przykład

Dla $p = 1$ jest 7 możliwych różnych modeli:

$$\mathbb{E}(Y) = \beta_0,$$

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1,$$

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2,$$

$$\mathbb{E}(Y) = \beta_1 x_1,$$

$$\mathbb{E}(Y) = \beta_0 + \beta_2 x_1^2,$$

$$\mathbb{E}(Y) = \beta_2 x_1^2,$$

$$\mathbb{E}(Y) = \beta_1 x_1 + \beta_2 x_1^2,$$

Dla $p = 2$ liczba możliwości wynosi 63.

Oczywistym jest, że problem się pogarsza dla wielomianów większego rzędu.

PCA

Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd..

PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników.

PCA

Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd..

PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników.

LDA

Używanie w uczeniu maszynowym do znalezienia liniowej kombinacji cech, które najlepiej rozróżniają dwie lub więcej klas obiektów lub zdarzeń. Wynikowe kombinacje są używane jako klasyfikator liniowy lub, częściej, służą redukcji wymiarów do późniejszej klasyfikacji statystycznej.