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

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

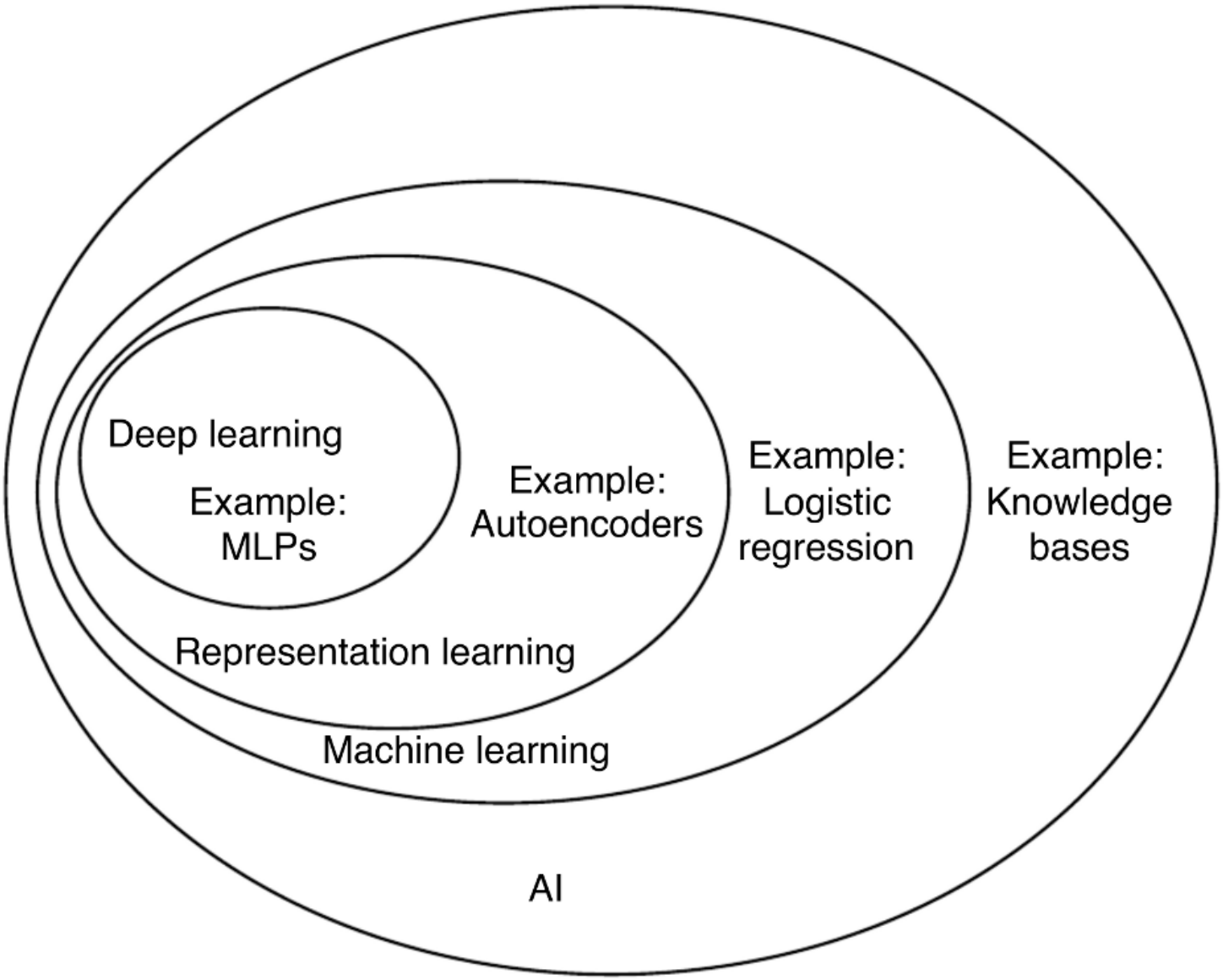
W świecie, gdzie niepewność modelu jest często ograniczeniem w procedurach wnioskowania, ważniejszym stała się predykcja/przewidywanie niż testowanie czy estymacja....
Dział informatyki zajmujący się rozwiązywaniem problemów, które nie są efektywnie algorytmizowalne

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. Informacje są nieustrukturyzowane i wielorakie.


Deep learning
Example:
MLPs
Example:
Autoencoders
Example:
Logistic
regression
Example:
Knowledge
bases
Representation learning
Machine learning
AI

Model

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

Model

- Odwzorowanie 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: $d(a, b) \leqslant d(a, c) + d(c, b)$

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 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

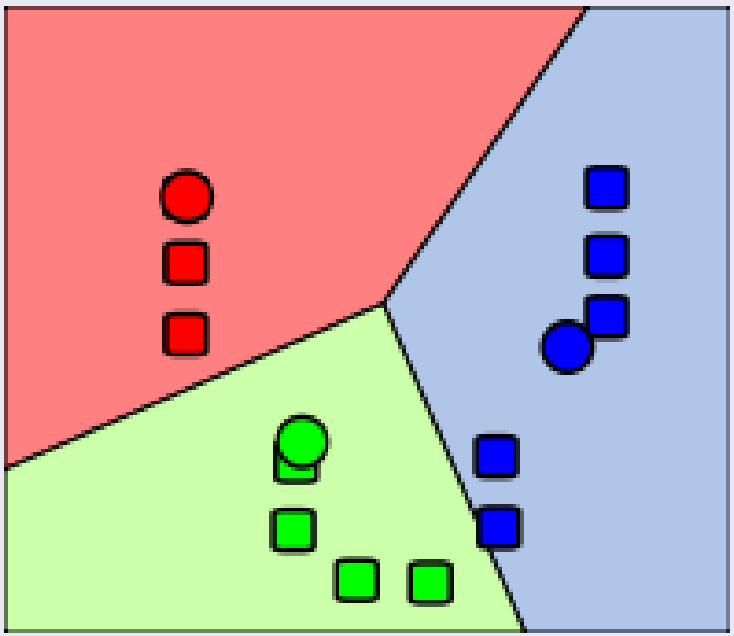
Metryka używana do porównywania zbiorów: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Żeby były spełnione warunki metryki:

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

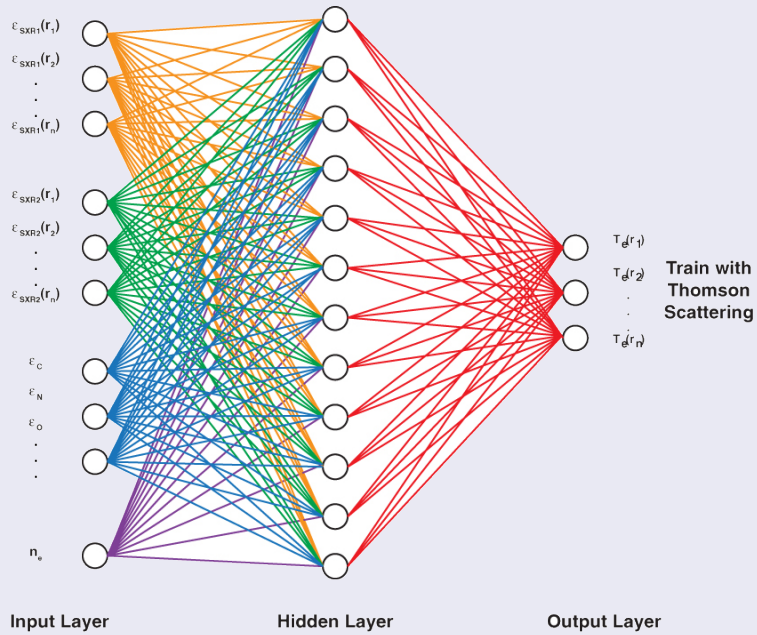
Local Methods

- K-means



Global Methods

- sieci neuronowe



'Definicja'

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.

'Definicja'

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 super-wykładniczy (superexponential).

Wprowadzenie pojęć

- Sparsity
- Liczba możliwych modeli
- Concurvity

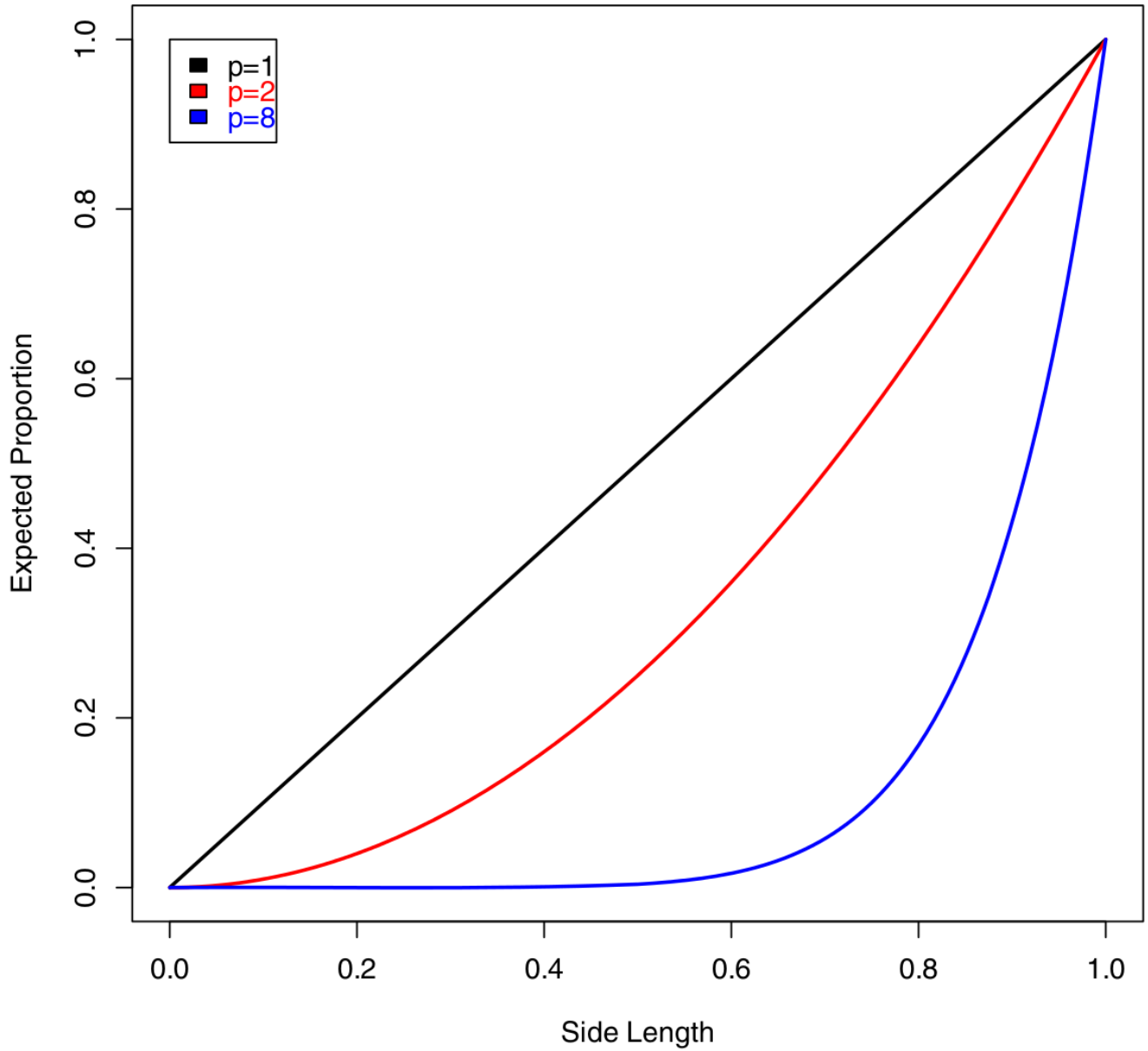
Motywacja

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.

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


p=1
p=2
p=8
Expected Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Side Length
0.0
0.2
0.4
0.6
0.8
1.0

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

Przykład

Dla $p = 1$ jest 7 możliwych różnych modeli:
$\mathbb{E}(Y) = \beta_0$,
$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1$,
$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$,

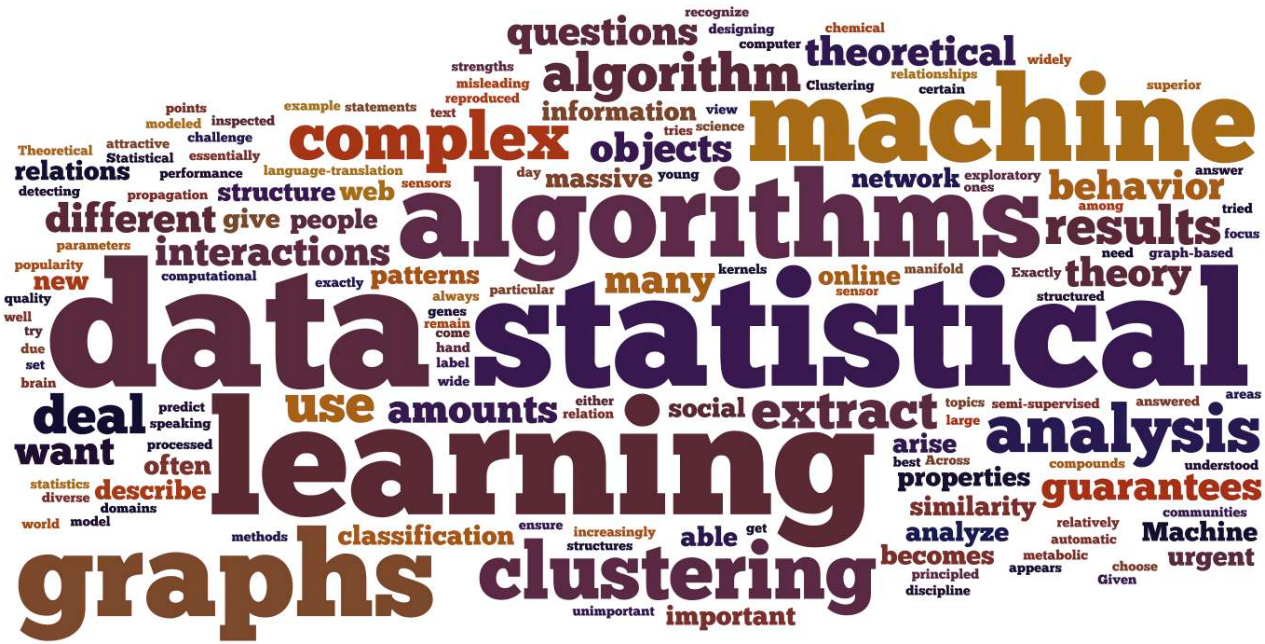
$\mathbb{E}(Y) = \beta_1 x_1$,
$\mathbb{E}(Y) = \beta_0 + \beta_2 x_1^2$,

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

Dla $p = 2$ liczba możliwości wynosi 63.

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

Concurvity


recognize
questions
designing
computer
chemical
theoretical
widely
algorithm
strengths
misleading
reproduced
Clustering
relationships
certain
superior
machine
points
example
statements
text
information
view
science
modeled
inspected
challenge
tries
complex
objects
Theoretical
attractive
Statistical
essentially
relations
performance
language-translation
sensors
day
massive
young
network
exploratory
ones
answer
behavior
detecting
propagation
structure
web
among
tried
different
give
people
algorithms
results
focus
parameters
interactions
need
graph-based
popularity
new
computational
exactly
patterns
kernels
many
online
manifold
Exactly
theory
quality
always
particular
sensor
structured
well
genes
data
statistical
try
remain
come
due
hand
set
label
brain
wide
areas
deal
predict
use
amounts
either
relation
social
extract
topics
semi-supervised
answered
large
analysis
want
speaking
processed
arise
often
best
Across
compounds
understood
properties
guarantees
statistics
describe
learning
diverse
similarity
communities
domains
world
model
ensure
increasingly
relatively
Machine
methods
classification
structures
able
get
analyze
automatic
becomes
metabolic
urgent
graphs
clustering
principled
appears
choose
discipline
Given
unimportant
important

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.