Klątwa wymiaru - raport

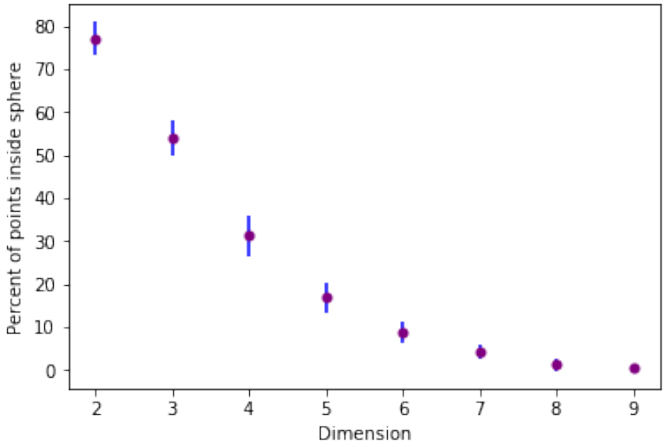
13 października 2017

Grzegorz Borkowski

1. WSTĘP

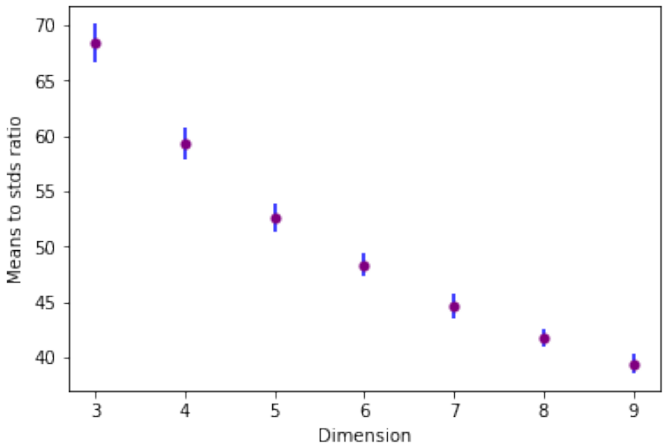
Zadanie zostało zaimplementowane w Pythonie 3.6.1
Do raportu został dołączony Jupyter Notebook z kodem źródłowym programu.

2. ZADANIE 1A



Rysunek 1. Procent punktów wewnątrz kuli dla różnych wymiarów dla 100 punktów i 25 powtórzeń

3. ZADANIE 1B



Rysunek 2. Stosunek odchylenia standardowego odległości między punktami do średniej odległości między nimi dla 100 punktów i 25 powtórzeń

4. KONSEKWENCJE DLA ALGORYTMÓW

Dla większej liczby wymiarów nie jesteśmy w stanie trenować modeli skutecznie, o ile nie mamy ogromnego zbioru treningowego. Spowodowane jest to tym, że liczba możliwych podziałów zbioru treningowego ze względu na cechy rośnie wykładniczo w stosunku do liczby cech. Wielkość zbioru treningowego musi rosnąć wykładniczo wraz ze wzrostem wymiaru cech. W przeciwnym razie model będzie przeuczony i nie będzie dawał wiarygodnych rezultatów na zbiorze testowym.
Stosunek odchylenia standardowego do średniej odległości między punktami maleje wraz ze wzrostem wymiarów. Można stąd wyciągnąć wniosek, że im większy wymiar, tym losowe punkty będą miały zbliżone odległości od siebie. Można sobie wyobrazić, że będą one leżały wewnątrz pewnej hipersfery i nie będziemy losować wielu punktów spoza poza obszarem tej hipersfery. Powoduje to podobny wniosek jak z zadania pierwszego, losowa przestrzeń punktów nie będzie gęsto pokrywała wszystkich możliwych podziałów, co prowadzi do przeuczenia modelów.
Konieczna jest zatem redukcja wymiarów.