Metody Rozpoznawania Obrazów I Podstawy Uczenia Maszynowego

Klątwa wymiaru

Autor: Ryszard Sikora

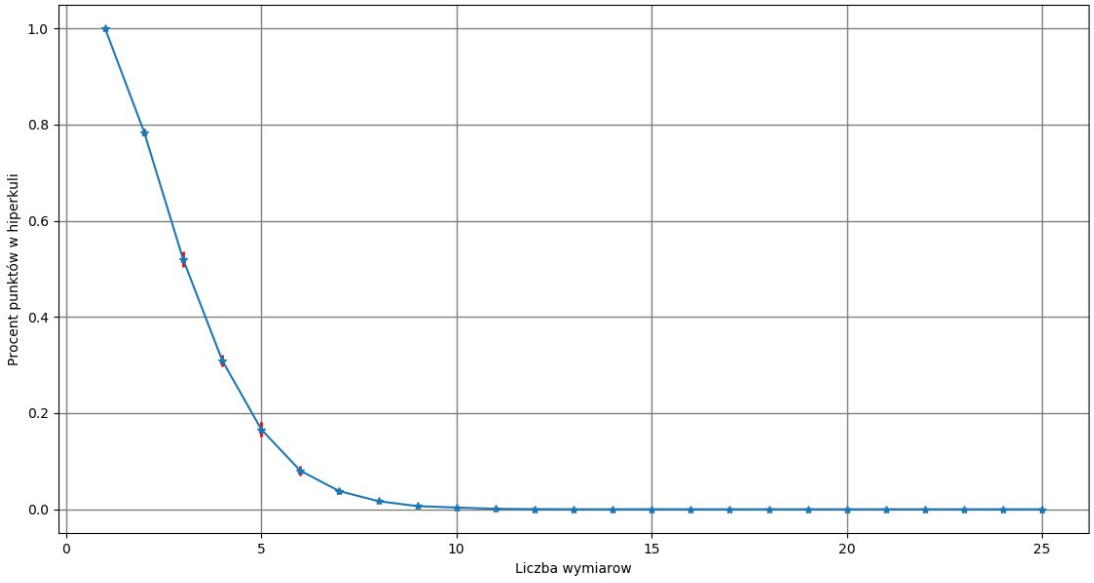
Obrazy w lepszej rozdzielczości znajdują się w repozytorium (https://github.com/rychuhardy/MRO1)

Zadanie A – Stosunek objętości (miarę Lebesque'a) hiperkuli wpisanej w hiperześcian do objętości tego hiperześcianu w zależności od liczby wymiarów

Obliczenie przeprowadzone zostało metodą Monte Carlo. Dla każdego wymiaru eksperyment powtórzony został 20 razy, a liczba losowanych punktów wynosiła 1000. Zakres wymiarów od 1 do 25.

Poniżej przedstawiony jest wykres wyniku eksperymentu. Czerwony kolor oznacza odchylenie standardowe.

Na rysunku błędnie podpisana jest oś Y – nie jest to procent a stosunek.



Obserwacja: Wraz ze wzrostem liczby wymiarów maleje stosunek objętości hiperkuli wpisanej w hiperześcian do objętości tego sześcianu. Przy około 12 wymiarze zaczęło brakować precyzji i stostunek ten spadł do 0.

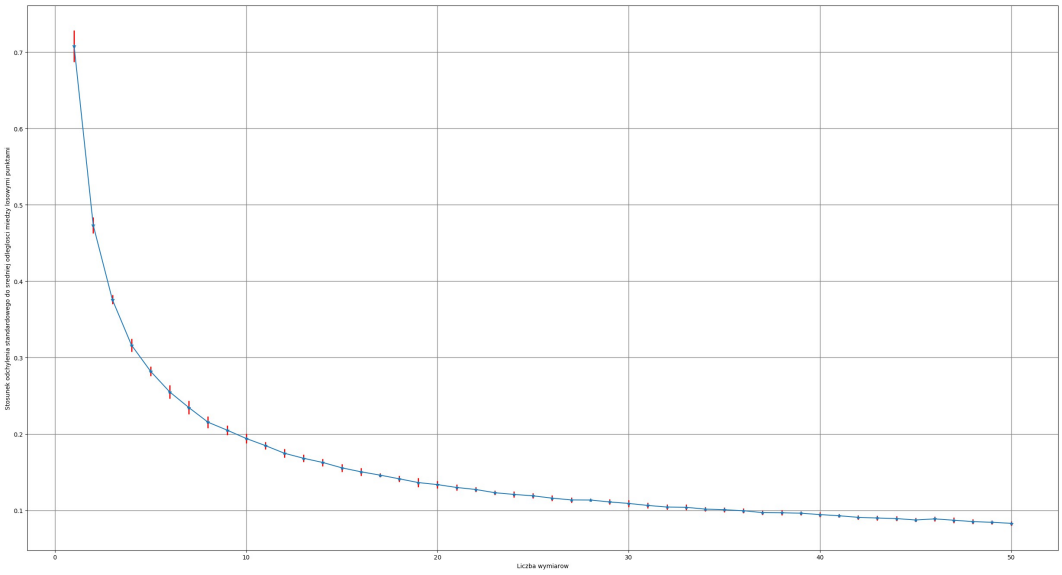
Wydaję mi się, że konsekwencją zaobserwowanego zjawiska jest to, że wraz ze wzrostem liczby wymiarów coraz trudniej jest znaleźć “punkty sąsiadujące” (wektory reprezentujące jakiś obiekt

podobny do wzorca) w zadanej odległości (promieniu) od zadanego punktu (czyli podobieństwo do tego wzorca). Oczywiście zwiększa się też koszt mocy obliczeniowej.

Zadanie B – Stostunek odchylenia standardowego odległości między punktami w hipersześcianie o krawędzi 1 do średniej odległości między tymi punktami

Podobnie jak w poprzednim zadaniu eksperyment został przeprowadzony metodą Monte Carlo. Dla każdego wymiaru wylosowano 50 próbek. Liczba powtórzeń dla wymiaru wynosiła 20. Zakres wymiarów to 1 do 50.

Poniżej przedstawiony jest wykres wyniku eksperymentu. Czerwony kolor oznacza odchylenie standardowe.



Obserwacja: Wraz ze wzrostem liczby wymiarów maleje stosunek odchylenia standardowego do średniej odległości między punktami. Punkty w przestrzeń n wymiarowej, dla dużego n (n > 1000), są losowane zawsze w niemal tej samej średniej odległości.

Wydaję mi się, że konsekwencją może być trudność wytrenowania modelu dla wielu wymiarów z powodu małej wariancji danych.