

Projekt 4 Super k-NN

1. Zbiór danych: TNG, ok 18000 próbek, 20 klas. Zbiór danych MNIST (70000 próbek 10 klas). Odnosnie TNG wykorzystujemy gotowe dane reprezentujące tekst blogów w postaci wektorów (dostarcza prowadzący). Dane dekorrelujemy wykorzystując transformatę PCA.
2. Celem projektu jest stworzenie zupełnie nowego zespołowego klasyfikatora k-NN i porównania jego jakości, czasów jego uczenia i odpowiedzi ze standardowym klasyfikatorem SVM.
3. Z jednego zbioru danych tworzymy kilka sub-zestawów danych ($\geq 5 \leq 10$) na różnych zestawach cech (maski mogą być losowane w sposób random, ale nie powinny być gęste). Maski mogą mieć różną długość. Prawdopodobieństwo wystąpienia cechy w zestawie może być proporcjonalne do jej istotności (np. mierzonej wielkością wartości własnych po transformacji PCA). Jednak nie może być takiej cechy, która nie dostała się do żadnego zestawu.
4. Liczymy średnią przynależność każdej próbki do danej klasy na bazie klasyfikatora k-NN dla każdego sub-zestawu danych. Dokonujemy fuzji wyników klasyfikacji (jakiej?) dla każdej próbki po sub-zestawach danych.
5. Jak zmieni się jakość klasyfikatora w zależności od k.
6. Wyniki jakości klasyfikatorów oceniać na bazie krzyżowej-walidacji, (accuracy - Loss-błąd, Krzywa ROC, Precision-Recall, (pole pod krzywymi) F1).
7. Należy sporządzić raport z projektu.

W nawiasie ocena maksymalna.