Rozpoznawanie i klasyfikacja pisanych cyfr przy użyciu modeli matematycznych

Anna Zawadzka
Piotr Waszkiewicz

6 listopada 2016

Rozdział 1

Projekt

1.1 Opis

Celem zadania jest porównanie jakości klasyfikacji dla różnych modeli matematycznych, oraz próba minimalizacji ich błędów poprzez ekstrakcję dodatkowych cech obiektów z bazy danych. Podczas realizacji projektu wykorzystane zostaną jedne z najpopularniejszych obecnie klasyfikatorów: maszyny wektorów podpierających (SVM), Lasy Losowe, kNN, model regresji wielomianowej oraz sieci neuronowe. Zbiory danych treningowych oraz testowych zostaną zaczerpnięte z publicznej bazy danych MNIST[1].

1.2 Cel badań

Celem badania jest wskazanie najskuteczniejszego klasyfikatora pod względem czasu uczenia, wydajności i jakości udzielanych odpowiedzi. Oprócz tego badania mają na celu rozszerzenie istniejącego wektora cech o nowe, unikalne wartości które polepszą jakość klasyfikacji. Przykładem takich cech może być liczba przecięć w napisanym symbolu, liczba zakończeń lub procent powierzchni zajmowanej przez narysowany symbol. W trakcie obliczeń podjęta zostanie próba odrzucenia tych cech które przeszkadzają lub pogarszają działanie modeli. Przeprowadzone badania obejmą również wybór optymalnych parametrów dla poszczególnych klasyfikatorów metodą GridSearch[2].

1.3 Zbiory danych

Zbiory danych treningowych oraz testowych pochodzą z publicznej bazy danych MNIST[1]. Każdy element ze zbioru treningowego jest opisany 785 wartościami. Pierwsza liczba określa zakodowaną cyfrę (wartość z przedziału [0, 9]), kolejne

784 wartości są z przedziału [0, 255] i opisują kolory pikseli zeskanowanej cyfry w skali szarości dla obrazka o wymiarach 28x28 pikseli. Zbiór testowy w przeciwieństwie do treningowego nie zawiera informacji o reprezentowanej klasie. Zbiór treningowy i testowy zawierają odpowiednio 60,000 i 10,000 elementów.

1.4 Sposób weryfikacji rozwiązań

Podczas ewaluacji otrzymywanych rozwiązań minimalizowana będzie funkcja błędu opisana wzorem

$$f(M, d) = e(M, d) + t(M, d)$$

gdzie M oznacza model, d zbiór testowy, *e()* współczynnik *Error rate*, czyli miarę określającą stosunek źle zaklasyfikowanych elementów do wszystkich obiektów w zbiorze, oraz *t()* funkcję czasu liczoną jako liczbę sekund potrzebną na realizację obliczeń.

Bibliografia

[1] http://yann.lecun.com/exdb/mnist/index.html

[2] http://scikit-learn.org/stable/modules/grid_search.html