WSI LISTA 4

Bohdan Tkachenko 256630

June 13, 2023

1 Zadanie 1: Las decyzyjny dla bazy danych MNIST

Celem tego zadania było stworzenie i wytrenowanie lasu decyzyjnego do rozpoznawania recznie pisanych cyfr z bazy danych MNIST za pomoca biblioteki TensorFlow.

1.1 Metodologia

Dane MNIST zostały wczytane i przygotowane do analizy. Każdy obrazek (28x28 pikseli) został przekształcony do jednowymiarowej tablicy o długości 784. Nastepnie stworzono i wytrenowano las decyzyjny za pomoca funkcji RandomForestClassifier z biblioteki scikit-learn. Las składał sie z 100 drzew decyzyjnych.

1.2 Wyniki

Dokładność modelu została oceniona na zbiorze testowym. Dokładność, mierzona jako procent poprawnie sklasyfikowanych obrazków, wyniosła 0.9705. Oznacza to, że las decyzyjny poprawnie rozpoznał 97.05% cyfr w zbiorze testowym.

1.3 Wnioski

Wyniki pokazuja, że las decyzyjny może być skutecznym narzedziem do rozpoznawania recznie pisanych cyfr. Jednakże, wyniki te powinny być porównane z wynikami innych modeli, takich jak sieci neuronowe, aby dokładnie ocenić ich skuteczność.

2 Zadanie 2: Klasteryzacja zbioru danych MNIST za pomoca algorytmu DBSCAN

Celem tego zadania było zastosowanie algorytmu DBSCAN do klasteryzacji zbioru danych MNIST, który zawiera obrazy recznie pisanych cyfr. DBSCAN

został wybrany ze wzgledu na jego zdolność do identyfikacji klastrów o dowolnym kształcie i rozmiarze, co jest szczególnie przydatne w przypadku danych MNIST, gdzie różne style pisania moga prowadzić do znacznej zmienności w obrebie tej samej klasy.

Dane MNIST zostały najpierw przekształcone do jednowymiarowych wektorów, a nastepnie znormalizowane. Nastepnie zastosowano redukcje wymiarowości za pomoca PCA, aby zmniejszyć wymiarowość danych do 50 głównych składowych.

Parametry DBSCAN, takie jak eps (maksymalna odległość miedzy dwoma próbkami, aby były uważane za sasiadów) i min_samples (minimalna liczba próbek w sasiedztwie, aby punkt danych był uważany za punkt centralny), zostały dobrane eksperymentalnie. Ostatecznie, eps zostało ustawione na 10.0, a min_samples na 5.

Algorytm DBSCAN zidentyfikował 30 klastrów, co jest równo z maksymalna liczba 30 sugerowana w zadaniu. Zidentyfikowano również 10976 punktów szumu, co stanowi mniejsza cześć danych w porównaniu do poprzednich prób.

Dokładność klasyfikacji, obliczona jako procent obrazów, które zostały przypisane do klastra reprezentujacego ich prawdziwa klase, wyniosła 0.16. Jest to poprawa w porównaniu do poprzednich prób, ale nadal jest to dość niska wartość, co sugeruje, że wiele obrazów zostało niepoprawnie sklasyfikowanych.

Wnioski: DBSCAN, mimo że jest potężnym algorytmem klasteryzacji, może nie być najlepszym wyborem dla danych o wysokiej wymiarowości, takich jak MNIST. Wyniki moga być poprawione przez dalsze dostrojenie parametrów DBSCAN, zastosowanie innej metody redukcji wymiarowości, lub zastosowanie innego algorytmu klasteryzacji.