

WSI LISTA 4

Bohdan Tkachenko 256630

June 12, 2023

1 Zadanie 1: Las decyzyjny dla bazy danych MNIST

Celem tego zadania było stworzenie i wytrenowanie lasu decyzyjnego do rozpoznawania ręcznie pisanych cyfr z bazy danych MNIST za pomocą 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. Następnie stworzono i wytrenowano las decyzyjny za pomocą funkcji RandomForestClassifier z biblioteki scikit-learn. Las składał się 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 pokazują, że las decyzyjny może być skutecznym narzędziem do rozpoznawania ręcznie 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 bazy danych MNIST za pomocą DBSCAN

Celem tego zadania było zastosowanie algorytmu DBSCAN do klasteryzacji bazy danych MNIST zawierającej ręcznie pisane cyfry.

2.1 Metodologia

Dane MNIST zostały wczytane i przygotowane do analizy. Każdy obrazek (28x28 pikseli) został przekształcony do jednowymiarowej tablicy o długości 784. Następnie dane zostały przeskalowane do średniej 0 i odchylenia standardowego 1 za pomocą funkcji `StandardScaler`.

Zastosowano również redukcję wymiarowości za pomocą analizy głównych składowych (PCA), redukując wymiarowość danych z 784 do 2.

Następnie zastosowano algorytm DBSCAN z parametrami `'eps=0.5'` i `'min_samples = 10'`. *Algorytm został wytrenowany na danych treningowych, a następnie użyto go do przewidywania klastrow na zbiorze testowym.*

2.2 Wyniki

Algorytm DBSCAN zidentyfikował 26 klastrow w danych. Liczba punktów szumu wynosiła 1433, co stanowiło 16.3% wszystkich punktów.

Dokładność modelu, mierzona jako procent poprawnie sklasyfikowanych obrazków, wyniosła 0.0294. Indeks Rand'a po dostosowaniu (ARI), miara podobieństwa między prawdziwymi etykietami a etykietami przypisanymi przez algorytm, wyniósł 0.0207.

2.3 Wnioski

Wyniki pokazują, że algorytm DBSCAN może być używany do klasteryzacji danych MNIST, ale dokładność i ARI są stosunkowo niskie, co sugeruje, że klasteryzacja może nie być idealna. Możliwe, że inne algorytmy klasteryzacji lub dostosowanie parametrów DBSCAN mogłyby poprawić wyniki.