Programowanie Obiektowe

Laboratorium 6 – (5 kwietnia 2023)

mgr inż. Damian Mroziński

Zadanie dodatkowe
Poniższe zadania ułożył mgr inż. Paweł Majewski.

Zadanie 6.1 (+5% pkt.)

Zadanie polega na implementacji klasyfikatora knn (k-nearest neighbors) oraz sprawdzeniu jego działania dla przykładowych zbiorów danych. W tym celu zdefiniuj klasę **KNeighborsClassifier**, która będzie zawierała metody:

metoda	argumenty	odpowiedzialności
fit(X_train, Y_train)	tablica X_train o rozmiarze n x k, zawierająca n-próbek treningowych (kolejne rzędy tablicy) oraz k-cech (kolejne kolumny tablicy), tablica Y_train o rozmiarze n, zawierająca n-etykiet dla każdej próbki treningowej np. 0 oraz 1 dla klasyfikacji binarnej	zdefiniowanie referencyjnej przestrzeni cech etykietowanych próbek, w celu wykorzystania jej podczas predykcji
predict(X_test)	tablica X_test o rozmiarze m x k, zawierająca m-próbek testowych oraz k-cech	predykcja etykiet dla danych testowych zawartych w tablicy X_test
calculate_accuracy(X_test, Y_test)	tablica X_test zawierająca próbki testowe (opis podobny jak dla metody predict()) oraz tablica Y_test zawierają etykiety dla danych testowych	obliczenie dokładności modelu jako stosunek właściwych predykcji do wszystkich predykcji

Tabela 1: Metody w klasie KNeighborsClassifier.

W konstruktorze klasy **KNeighborsClassifier** należy uwzględnić parametry:

- **n_neighbors** - liczba sąsiadów brana pod uwagę podczas predykcji,
- **metric** - rodzaj metryki użytej do obliczania odległości między próbkami np. *euclidean*, *manhattan*.

Predykcja dla modelu knn polega na znalezieniu k-najbliższych sąsiadów w przestrzeni cech (zdefiniowanej na podstawie danych treningowych) dla każdej próbki testowej, z użyciem odpowiedniej metryki, określającej dystans pomiędzy próbkami. Na podstawie etykiet k-najbliższych sąsiadów określana jest klasa dla próbki testowej, która ta klasa najczęściej występuje wśród rozważanych sąsiadów.

Dla wczytywania danych do tablic zdefiniuj dodatkowo klasę **DataReader**, która będzie odpowiedzialna za wczytywanie danych treningowych i testowych do tablic. Użyj tablic dynamicznych przy wczytywaniu danych.

W celu weryfikacji poprawnej implementacji przetestuj działanie modelu knn dla dwóch przykładowych zbiorów danych (*iris* oraz *breast_cancer*). Wczytaj dane treningowe i testowe znajdujące się w plikach *.txt. Sprawdź działanie modelu knn dla następujących parametrów modelu $n\_neighbors \in \{3, 5, 7, 9, 11\}$ oraz $metric \in \{'euclidean', 'manhattan'\}$. Oblicz dokładność modelu dla każdego zbioru danych i określonych parametrów modelu. Zapisz wyniki do pliku *.txt. Napisz testy jednostkowe dla metod z klasy **KNeighborsClassifier**, definiując syntetyczne zbiory danych (np. próbki posiadające 2 cechy należące do dwóch separowalnych klas).

Opis zbiorów danych:

1. *iris* - 4 cechy, 3 klasy, etykieta w kolumnie species
2. *cancer* - 30 cech, 2 klasy, etykieta w kolumnie diagnosis (pomiń kolumnę zawierającą id pacjentów).

Dodatkowe informacje możesz odnaleźć na stronach:

- `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`,
- `https://en.wikipedia.org/wiki/Taxicab_geometry`,
- `https://en.wikipedia.org/wiki/Iris_flower_data_set`,
- `https://www.kaggle.com/uciml/breast-cancer-wisconsin-data`