k–Nearest Neigbors

1 Klasyfikator k-NN

1. Wczytujemy zbiór treningowy i wybieramy dodatnią liczbę całkowitą k.
2. Dany jest przykład do zaklasyfikowania.
3. Ze zbioru treningowego wybieramy k przykładów o najmniejszym dystansie do klasyfikowanego przykładu.
4. Klasyfikujemy przykład zgodnie z klasą, która wystąpiła najczęściej wśród wybranych przykładów treningowych.

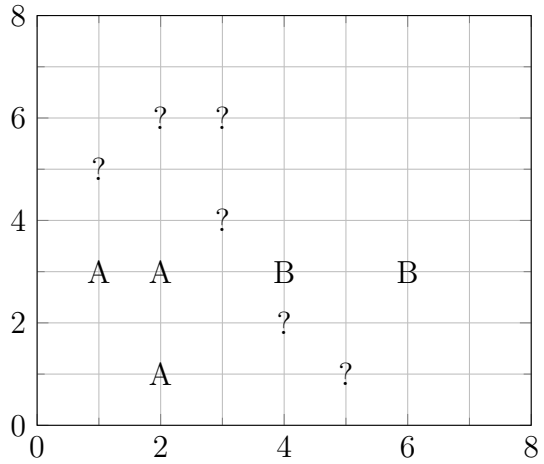
Zadania

Zadanie 1.

Przypisz następujące przykłady do klasy A lub B przy użyciu metody k-NN z $k = 3$.
Zbiór treningowy: A(1, 3), A(2, 1), A(2, 3), B(4, 3), B(6, 3).

Przykłady do zaklasyfikowania:

- $(1, 5)$
- $(3, 6)$
- $(2, 6)$
- $(4, 2)$
- $(3, 4)$
- $(5, 1)$



Zadanie 2.

Podobnie jak w Zadaniu 1., klasyfikuj poniższe przykłady na podstawie zbioru treningowego ($k = 3$):

Zbiór treningowy: A(5, 4, 1), A(4, 3, 0), B(1, 2, 3), B(2, 0, 4), C(6, 1, 1), C(5, 0, 1).

Przykłady do zaklasyfikowania: (4, 4, 0), (1, 1, 5), (6, 0, 0).

Mini-projekt: k-NN

Celem jest napisanie programu, ktory pobiera następujące argumenty:

`k`: dodatnia liczba naturalna będąca hiperparametrem k-NN.

`train-set`: nazwa pliku zawierającego zbiór treningowy w postaci csv – ostatnia kolumna zawiera atrybut decyzyjny.

`test-set`: nazwa pliku zawierajacego zbiór testowy w postaci csv – ostatnia kolumna zawiera atrybut decyzyjny.

Wymagania:

- Program powinien dokonać klasyfikacji k-NN wszystkich obserwacji z pliku `test-set` na podstawie pliku `train-set` oraz podać dokladność (accuracy) tej klasyfikacji (proporcję poprawnie zaklasyfikowanych przykładów testowych).
- Program ma też dostarczać testowy interfejs (niekoniecznie graficzny), który umożliwia (zapętlone) podawanie przez użytkownika pojedynczych wektorów do klasyfikacji i podaje ich etykietę k-NN na podstawie `train-set`.
- Przetestować na danych ze zbiorów treningowego i testowego znajdujących się w plikach `iris.data` i `iris.test.data`.
- **Uwaga:** Program powinien przyjąć dowolny zbiór danych (w formacie podobnym do `iris.data`) i dostosować się do dowolnej liczby wymiarów.
- **Opcjonalnie (dodatkowy punkt za aktywność):** dowolną techniką (excel, python, etc.) zrobić wykres zależności dokładności (accuracy) od wartości k.
- **Opcjonalnie (dodatkowy punkt za aktywność):** Testować także zbiór *WDBC* w `wdbc.data` i `wdbc.test.data` [Źródło].