WSI - ćwiczenie 6.
Regresja i klasyfikacja
grupa 101

28 kwietnia 2021

1 Sprawy organizacyjne

1. Ćwiczenie realizowane jest samodzielnie.
2. Ćwiczenie wykonywane jest w jezyku R lub Python.
3. Ćwiczenie powinno zostać wykonane do 13.05.2021 23:59. Do tego czasu na adres mailowy jakub.lyskawa.dokt@pw.edu.pl należy przesłać plik .zip albo .tar.gz zawierający kod, dokumentację oraz skan lub zdjęcie podpisanego oświadczenia o pracy zdalnej.
4. Dokumentacja powinna być w postaci pliku .pdf, .html albo notebooka jupyterowego. Szczegółowe informacje co dokumentacja powinna zawierać oraz na co będzie zwracana uwaga podczas oceniania znajdują się na stronie
`http://staff.elka.pw.edu.pl/~rbiedrzy/WSI/index.html`
5. Wzór oświadczenia o pracy zdalnej jest załącznikiem do zarządzenia
`https://www.bip.pw.edu.pl/var/pw/storage/original/application/9bfa38aad48ba019ab4cd5449ef209b6.pdf`
6. W przypadku pytań lub wątpliwosci zachęcam do pisania na adres mailowy jakub.lyskawa.dokt@pw.edu.pl albo na platformie MS Teams (konto powiązane z powyższym adresem email).

2 Zadanie

W ramach szóstego ćwiczenia należy zaimplementować drzewo decyzyjne indukowane algorytmem ID3.

Należy umożwliwić ustawienie maksymalnej głębokości drzewa podczas jego tworzenia. Jeżeli węzeł na tej głębokości nie pozwala na jednoznaczną klasyfikację, powinien stać się liściem zawierającym najczęstszą klasę.

Następnie należy przetestować zaimplementowany klasyfikator z użyciem zbioru danych titanic `https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv`. Należy uwzględnić podział zbioru na treningowy, walidacyjny i testowy.

Uwaga! Niektóre atrybuty mogą się nie nadawać do zastosowania (na przykład imię i nazwisko), a niektore atrybuty mają wartości ciągłe (na przykład wiek), które należy podzielić na zakresy i potraktować te zakresy jako atrybuty dyskretne.

3 Wskazówki

- W implementacji nie powinno być magicznych stałych, parametry algorytmu powinny być przekazywane np. jako parametry funkcji która ten algorytm implementuje, nie powinny być również przekazywane jako zmienne globalne
- Implementacje powinny być ogólne. Należy unikać pisania osobnej implementacji algorytmu dla każdego problemu.
- Przełączanie wariantów implementacji poprzez komentowanie fragmentów kodu nie jest dobrą praktyką.
- W miarę możliwości warto korzystać z gotowych implementacji np. operacji macierzowych i wektorowych (oczywiście wskazane w poleceniach algorytmy należy zaimplementować samodzielnie).
- Dokumentacja powinna zawierać opis przeprowadzonych eksperymentów, prezentować w jakiejś formie ich wyniki oraz zawierać komentarz do tych wyników.