

WSI - ćwiczenie 5.

Modele bayesowskie

grupa 101

14 kwietnia 2021

1 Sprawy organizacyjne

1. Ćwiczenie realizowane jest samodzielnie.
2. Ćwiczenie wykonywane jest w języku R lub Python.
3. Ćwiczenie powinno zostać wykonane do 29.04.2021 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 wątpliwości 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 piątego ćwiczenia należy zaimplementować naiwny klasyfikator Bayesa.

Korzystając z tego klasyfikatora należy zbadać, który atrybut ze zbioru danych *wine* (<https://archive.ics.uci.edu/ml/datasets/wine>) pozwala osiągnąć najlepszą dokładność klasyfikacji. W celu oceny zbioru należy wykorzystać algorytm n-krotnej walidacji krzyżowej.

Następnie należy zbadać, czy dodanie atrybutu o numerze $N + 1$ pozwala poprawić jakość klasyfikacji, gdzie N oznacza ostatnią cyfrę numeru indeksu (czyli student, którego numer indeksu kończy się na 0 powinien dodać atrybut o numerze 1). Jeżeli atrybut $N + 1$ jest atrybutem znalezionym w poprzednim etapie, dodany powinien zostać atrybut o numerze $N + 2$.

Numeracja atrybutów jest zgodna z informacjami pod powyższym linkiem, tj.

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

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 algorytmu należy zaimplementować samodzielnie).

- Dokumentacja powinna zawierać opis przeprowadzonych eksperymentów, prezentować w jakiejś formie ich wyniki oraz zawierać komentarz do tych wyników.