Teoria informacji w uczeniu maszynowym – Projekt 1

Twoim zadaniem jest zaimplementowanie dwóch modeli klasyfikacyjnych: drzewa decyzyjnego i lasu losowego. Nie możesz korzystać z żadnych gotowych implementacji tych modeli lub ich elementów.

Drzewo decyzyjne

W implementacji powinny się znaleźć następujące procedury:

- `create_dataset`: wczytaj dane, podziel je losowo na zbiór treningowy, walidacyjny i testowy według proporcji, które są argumentem procedury; domyślny podział $5:1:1$.
- `create_tree`: budujemy drzewo zgodnie z materiałem z wykładu. Stosujemy entropijne kryterium podziału.
- `evaluate_tree`: uruchamiamy wytrenowany klasyfikator na zbiorze przykładów. Procedura powinna zwrócić klasy, jakie przypisał klasyfikator oraz skuteczność (*accuracy*).
- `prune_tree`; funkcja przyjmuje wartość $\varepsilon \geq 0$ – domyślnie $\varepsilon = 0$:
  - Sprawdź wartość accuracy Acc_1 zbudowanego drzewa dla zbioru walidacyjnego.
  - Wybierz węzeł, którego potomkowie są liśćmi. Usuń tych potomków, a jako klasę rozważanego węzła wybierz tę, która ma najwięcej elementów.
  - Sprawdź accuracy Acc_2 dla takiego obciętego drzewa na zbiorze walidacyjnym. Jeśli $Acc_2 \geq Acc_1 - \varepsilon$, to zachowaj drzewo obcięte, w przeciwnym razie zachowaj poprzednie drzewo.
  - Powtarzaj dwa poprzednie kroki dopóki nie da się obciąć liści żadnego węzła.

Las losowy

Zaimplementuj las losowy według opisu z `https://en.wikipedia.org/wiki/Random_forest#Algorithm` (bez podsekcji *ExtraTrees*). W implementacji powinny się znaleźć następujące procedury:

- modyfikacja procedury `create_tree` tak, by pozwalała na użycie *feature bagging* (patrz wiki) podczas tworzenia drzewa.
- `create_random_forest`: budujemy las losowy korzystając z drzew decyzyjnych i *tree bagging*.
- `evaluate_random_forest`: uruchamiamy wytrenowany klasyfikator na zbiorze przykładów. Procedura powinna zwrócić klasy, jakie przypisał klasyfikator oraz skuteczność (accuracy).

Testowanie

Przetestuj swoje implementacje na zbiorze danych *Wine Quality Data Set*, dostępnym pod adresem `https://archive.ics.uci.edu/ml/datasets/wine+quality`. Użyj pliku `winequality-white.csv`.

Drzewo decyzyjne powinno uzyskać co najmniej 50% skuteczności na zbiorze testowym, a las losowy – co najmniej 60%. Oczywiście zbiór testowy nie może być używany przy trenowaniu ani obcinaniu drzew.

Obrona

W ramach obrony projektu będziecie musieli pokazać, że implementacja działa i ją rozumiecie.

Wstępny termin oddania projektów: 7 grudnia, 12:00-15:00, sala 1056 (porozmawiamy na zajęciach, jeśli z jakichś powodów wam nie pasuje to możemy go trochę dostosować). Po tym terminie nie będzie możliwości uzyskania punktów!