AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 3

Klasyfikacja

Indukcja drzew decyzyjnych za pomocą CART

data aktualizacji: 04.04.2023

Cel ćwiczenia

Celem ćwiczenia laboratoryjnego jest zapoznanie się z drzewami decyzyjnymi.

Wprowadzenie

W zadaniu badany będzie algorytm Classification and Regression Trees (CART), aktualnie jedna z najczęściej używanych implementacji. Algorytm działa w oparciu o współczynnik Giniego (mierzący stopień nierównomierności rozkładu, który można w tym kontekście interpretować podobnie jak entropię) i na tej podstawie zachłannie dzieli dane budując finalne drzewo decyzyjne. Algorytm ten ma zestaw parametrów, które mogą krytycznie wpłynąć na skuteczność klasyfikacji wynikowego drzewa decyzyjnego.

Przebieg ćwiczenia

1. Zaczytanie zbiorów danych IRIS, Polish Companies Bankruptcy.
2. Uruchomienie algorytmu drzewa decyzyjnego dla IRIS dla domyślnych parametrów.

3. Wizualizacja drzewa i analiza jakości klasyfikacji wynikowego drzewa decyzyjnego.
4. Wstępne strojenie algorytmów dla zbiorów - sugerowane 4 hiperparametry: criterion, max_depth, min_samples_leaf, cpp_alpha. W szczególności zwróć uwagę na pruning (cpp_alpha).
5. Wizualizacja drzew dla różnych hiperparametrów i analiza jakości klasyfikacji wynikowych drzew decyzyjnych.
6. Użycie walidacji krzyżowej (zwykłej oraz stratyfikowanej).
7. Użycie parametru wagi klasy (class_weight) oraz analiza wyników.
8. Podsumowanie wyników. To jest miejsce na tabelki, wykresy, wnioski – wybieramy prezentowane dane/zestawienia, nie dajemy wszystkich wyników.

Uwaga! Przy tym zadaniu nie używamy *boostingu* – ten mechanizm będzie badany przy okazji jednego z następnych zadań laboratoryjnych.

Punktacja

Przy realizacji zadania student może otrzymać **max 10 punktów** wedle poniższej tabeli.

2	Wczytanie danych, uruchomienie klasyfikatora, wizualizacja wyników.
4	Strojenie hiperparametrów drzewa decyzyjnego oraz zbadanie wpływu na skuteczność. Wizualizacja kilku przykładowych wynikowych drzew (“Jak różnią się wynikowe drzewa pod wpływem różnych zestawów hiperparametrów?”)
2	Zbadanie jak wyniki użycia walidacji stratyfikowanej różnią się od „zwykłej” walidacji krzyżowej
2	Zbadanie jak użycie parametru wagi klasy wpływa na wyniki modelu.

Przy realizacji tego zadania wystarczy prosty raport PDF. Przy wynikach badań należy dać komentarz, podać wnioski i podsumowanie.

Pytania pomocnicze

1. Co znajduje się w liściach drzewa?
2. Czy przycinanie drzewa (*pruning*) jest potrzebne? Na czym polega ten proces?
3. Czy drzewo może być za „duże” lub za „małe”?
4. Dlaczego typ/rozmiar kroswalidacji może mieć duży wpływ na skuteczność modelu?
5. Czy drzewo decyzyjne potrzebuje normalizacji/standaryzacji/dyskretyzacji danych?
6. Czy model można przeuczyć?
7. Na czym polega wagowanie klas?

Literatura

1. Wykłady do przedmiotu autorstwa prof. H.Kwaśnickiej
2. Cichosz P. ”Systemy uczące się”, WNT Warszawa
3. Zasoby Internetu: uczenie maszynowe (machine learning), data mining, klasyfikacja, drzewa decyzyjne, indukcja drzew decyzyjnych, pruning (przycinanie drzewa), generalizacja
4. 1.10. Decision Trees — scikit-learn 1.2.2 documentation
5. sklearn.tree.DecisionTreeClassifier — scikit-learn 1.2.2 documentation
6. Decision tree learning - Wikipedia