AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 3 - Drzew Decyzyjne

Data aktualizacji: 10.04.2024

Cel ćwiczenia

Celem ćwiczenia laboratoryjnego jest zapoznanie się z algorytmem drzew decyzyjnych oraz dalsza praktyka z analizą oraz przygotowaniem danych.

Wprowadzenie

W zadaniu badany będzie algorytm Classification and Regression Trees (CART), aktualnie jedna z najczęściej używanych implementacji drzew decyzyjnych. 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. Wczytanie zbioru danych Secondary Mushroom [1].
2. Dokładna analiza zbioru danych. Z uwagi na dużą ilość zmiennych warto wykorzystać narzędia typu AutoEDA, np. [4]. Napisać wnioski.

3. Przygotowanie danych do modelowania (patrz lista numer 1, np. wartości brakujące, wartości odstające, normalizacja). Pamiętać o odpowiednim podziale danych na zbiór treningowy oraz testowy.

4. Wytrenowanie modeli drzew decyzyjnych [2, 3] z wykorzystaniem przeszukiwania hiperparametrów (np. GridSearch [5]) oraz dobraniem odpowiedniej miary klasyfikacji. Sugerowane 4 hiperparametry: criterion, max_depth, min_samples_leaf, cpp_alpha. W szczególności zwróć uwagę na pruning (cpp_alpha). Analiza wpływu hiperparametrów na jakość wyników.

5. Wizualizacja drzewa oraz analiza drzew dla różnych hiperparametrów. ("Jak różnią się wynikowe drzewa pod wpływem różnych zestawów hiperparametrów?")

6. Wykorzystanie parametru wagi klasy (class_weight) oraz analiza wyników.

Uwaga! Przy tym zadaniu nie używamy zespołów klasyfikatorów, np., AdaBoost, XGBoost, Random Forest. Ten mechanizm będzie badany przy okazji następnych zadań laboratoryjnych.

Punktacja

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

1	Analiza zbioru danych.
2	Przygotowanie danych do modelowania.
1	Trening modeli z wykorzystaniem przeszukiwania hiperparametrów modelu.
2	Analiza wyników wpływu hiperparametrów.
2	Analiza wynikowych drzew z różnymi hiperparametrami.
2	Zbadanie jak użycie parametru wagi klasy wpływa na wyniki modelu.

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. Czy drzewo decyzyjne potrzebuje normalizacji/standaryzacji/dyskretyzacji danych?
5. Czy model można przeuczyć?
6. Na czym polega wagowanie klas?
7. Na czym polega walidacja krzyżowa (ang. cross validation) w algorytmie przeszukiwania hiperparametrów?

Literatura

1. Zbiór danych - https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset
2. 1.10. Decision Trees — scikit-learn 1.2.2 documentation
3. sklearn.tree.DecisionTreeClassifier — scikit-learn 1.2.2 documentation
4. Przykład AutoEDA - https://github.com/ydataai/ydata-profiling
5. 3.2. Tuning the hyper-parameters of an estimator — scikit-learn 1.4.2 documentation