AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe” – laboratorium

Laboratorium 5 - Zespoły klasyfikatorów

Data aktualizacji: 16.05.2024

Cel ćwiczenia

Celem ćwiczenia laboratoryjnego jest zapoznanie się z najpopularniejszymi metodami wykorzystywanymi do rozwiązywania rzeczywistych problemów - zespołami klasyfikatorów (opartymi o drzewa). W szczególności celem zadania będzie przetestowanie modeli takich jak Random Forest [1], Histogram-based Gradient Boosting [2], XGBoost [3], LightGBM[4], CatBoost[5].

Wprowadzenie

Algorytmy oparte o zespoły klasyfikatorów do dnia dzisiejszego są uznawane za state-of-the-art w przypadku problemów na danych tabelarycznych (patrz [6]). Zazwyczaj są one de-facto standardem w praktycznych zastosowaniach w przemyśle i dzięki wielu ich własnością umożliwiają modelowanie różnorodnych problemów. Przykładami takich własności są: natywne wsparcie dla zmiennych kategorycznych, natywne wsparcie dla brakujących danych, wsparcie dla wag, ograniczenia monotoniczności, ograniczenia interackji, własne funkcje straty, modelowanie probabilistyczne dla zmiennych ciągłych i wiele innych. Dla ciekawych więcej tutaj [7,8].

Przebieg ćwiczenia

1. Wybranie własnego zbioru danych zawierającego zmienne numeryczne, kategoryczne oraz brakujące wartości, który nie pojawił się do tej pory na zajęciach. (Gdyby zabrakło inspiracji to tutaj wskazówka: [9]). Dokładna analiza zbioru danych razem z wnioskami.

2. Uruchomienie algorytmów Random Forest [1], Histogram-based Gradient Boosting [2], XGBoost [3], LightGBM[4], CatBoost[5] **bez** szukania hiperparametrów, ale z walidacją krzyżową. Porównanie rezultatów.

3. Wybranie jednego z powyższych algortymów oraz zapoznanie się z dostępnymi hiperparametrami. Przeprowadzenie procesu szukania hiperparametów (dla ambitnych można skorzystać z pakietów do optymalizacji bayesowskiej, e.g., https://optuna.org/). Analiza rezultatów.

4. Wybranie jednego z powyższych algorytmów wspierających jedną z następujących funkcjonalności: ograniczenia monotoniczności, ograniczenia interackji, własne funkcje straty lub inne wybrane. Przeprowadzenie eksperymentu sprawdzającego wpływ (jednej) wybranej funkcjonalności na wyniki oraz zachowanie modelu.

Punktacja

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

2	Realizacja ćwiczenia 1.
2	Realizacja ćwiczenia 2.
3	Realizacja ćwiczenia 3.
3	Realizacja ćwiczenia 4.

Literatura

1. RandomForest - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier
2. HistGradientBoosting - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html#sklearn.ensemble.HistGradientBoostingClassifier
3. XGBoost - https://xgboost.readthedocs.io/en/stable/
4. LightGBM - https://lightgbm.readthedocs.io/en/stable/
5. CatBoost - https://catboost.ai/
6. Why do tree-based models still outperform deep learning on tabular data? - https://arxiv.org/pdf/2207.08815
7. Ensembles in scikit-learn - https://scikit-learn.org/stable/modules/ensemble.html
8. Introduction to Boosted Trees - https://xgboost.readthedocs.io/en/stable/tutorials/model.html
9. Zbiór danych Titanic.