Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

AutoML
Homework2

Piotr Robak, Agata Węglerska

1 Wstęp

W ramach zadania głównym celem było opracowanie efektywnej metody klasyfikacji dla sztucznie wygenerowanego zbioru danych oznaczonego jako *artificial*. Zadanie polegało na skonstruowaniu modelu, który osiągnie jak najwyższą moc predykcyjną, a jego skuteczność będzie oceniana za pomocą miary zrównoważonej dokładności (balanced accuracy), która jest zdefiniowane jako:

$$\text{Miara zrównoważonej dokładności} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right)$$

gdzie:

TP – liczba prawdziwie pozytywnych (True Positives),
TN – liczba prawdziwie negatywnych (True Negatives),
FP – liczba fałszywie pozytywnych (False Positives),
FN – liczba fałszywie negatywnych (False Negatives).

Modele klasyfikacyjne zostały przygotowane w dwóch wariantach. Pierwszy z nich to podejście manualne, gdzie konstrukcją modelu zajęto się ręcznie. Drugim wariantem było wykorzystanie frameworków AutoMLowych, czyli narzędzi automatyzujących proces tworzenia modeli maszynowych.

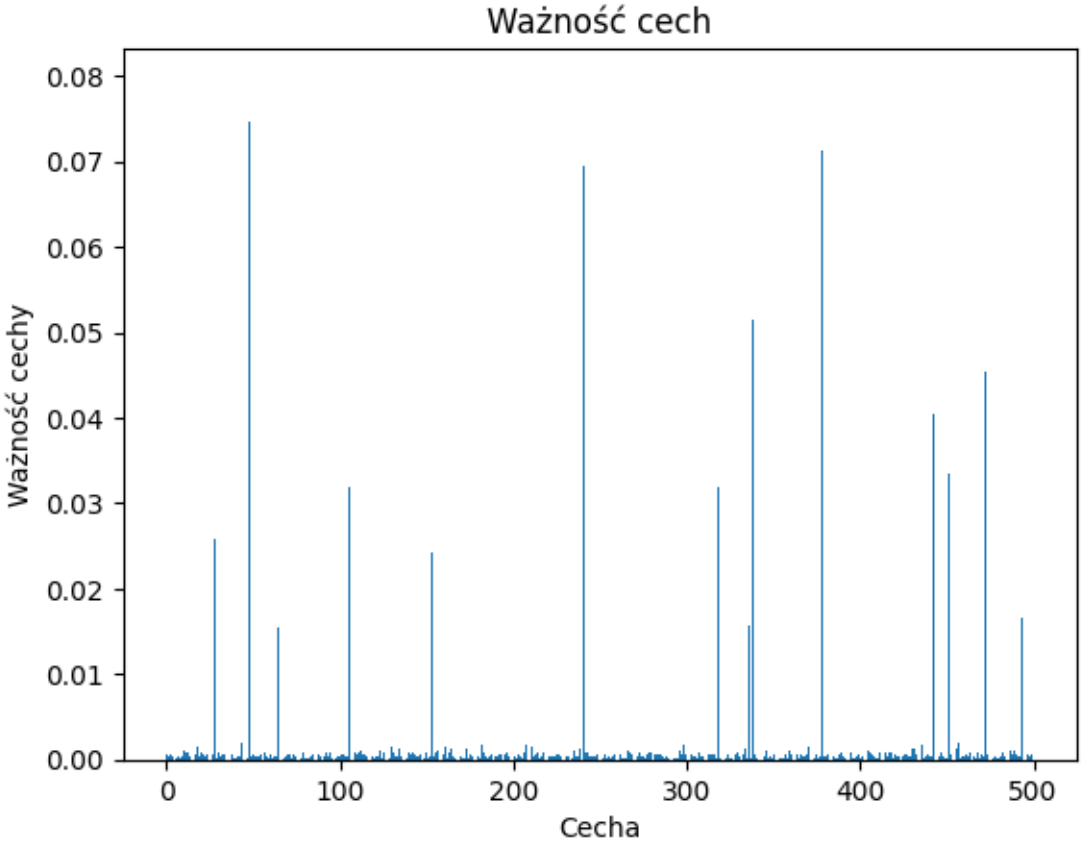
2 Preprocessing zbioru

W fazie pierwszej zajęto się zbadaniem zbioru, na którym modele będą uczone. W tym celu sprawdzono ilość komórek pustych (aby sprawdzić czy nie ma braków w danych), których liczba wyniosła zero. Kolejno sprawdzono korelacje zachodzące pomiędzy poszczególnymi cechami danych. Okazało się, że są one znikome. Sprawdzono również zrównoważoność zbioru, tzn ile elementów w zbiorze labels jest równych -1, a ile 1 (przynależności do dwuch różnych klas), z wynikiem pozytywnym, tzn ze zbiór jest zrównoważony idealnie. Ponad to określono liczbę cech, wynoszącą 500.

Następnie dostarczony zbiór z pliku *artificial_train.data* oraz analogicznie z *artificial_train.labels* podzielono w skali 7:3 na zbiory treningowy oraz testowy.

Bazując na przeprowadzonym badaniu właściwości postanowiono zająć się ograniczeniem ilości jego cech o ile to możliwe. Wykorzystano w tym celu drzewa decyzyjne, które wyuczone na zbiorze treningowym pozwoliły określić

ważność cech, które brały pod uwagę co widać na wykresie. Wybranych zostało z nich 20 najbardziej znaczących.



3 Modele wybrane manualnie

Jako klasyfikatory binarne wybrano regresję logistyczną, las losowy oraz wektory nośne dla klasyfikacji (SVC) ze względu na różnorodne działanie każdego z algorytmów oraz małą, który z nich mógłby się sprawdzić najlepiej.

Po przeprowadzeniu kilkudziesięciu prób wybrano działanie lasów losowych jako najbardziej skuteczne, dające najlepsze wyniki.

Model	Balanced accuracy na zbiorze treningowym	Balanced accuracy na zbiorze testowym
Regresja logistyczna	0.592	0.612
SVC	0.860	0.89
Las losowy	0.897	1.00

Z tabeli można odczytać miarę zrównoważonej dokładności dla każdego z modeli. Wytrenowanie lasów losowych odznacza się przetrenowaniem i niestety nie udało się tego zmienić.

Następnie dla wybranego lasu losowego sprawdzono jak model sobie radzi trenując oraz testując go z różnym podziałem na te dwa zbiory. Testowanie okazało się stabilne. W związku z tym wytrenowano model na całym zbiorze *artificial_train.data* oraz przeprowadzono predykcję z wyznaczeniem prawdopodobieństwa przynależności do klasy 1 dla zbioru *artificial_test.data*.

4 Model wybrany przez framework AutoMLowy

Dwa frameworki AutoMLowe zostały wykorzytsane do wyznaczenia modelu, robiącego dobre predykcje dla zbioru 'artificial'. Pierwszy z nich to AutoSklearnClassifier oraz drugi - AutoSklearn2Classifier. Obydwa wybierały modele w przeciągu 30 minut z metryką mierzącą jakość predykcji ustawioną na balanced accuracy. Dane na których modele były uczone oraz testowane, a w zasadzie podział opisany został w części dotyczącej preprocessingu.

Predykcje otrzymane przez AutoSklearnClassifier okazały się być nieco lepsze z wynikiem 0.870 mierzonym za pomocą balanced accuracy oraz przez wybrany model Bernoulli Naive Bayes classifier z przygotowanymi na początku danymi przechodzącymi przez preprocessing na całym zbiorze, a także jego cechach.