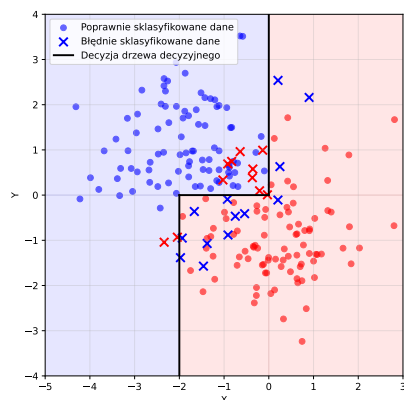




Ilustracja zasady działania drzewa decyzyjnego w przestrzeni dwóch cech. Łamana czarna granica odpowiada sekwencji pytań decyzyjnych: (1) czy $y > 0$; (2a) jeśli tak, czy $x < 0$; (2b) jeśli nie, czy $x < -2$. Krzyżyki oznaczają błędne klasyfikacje

Kaggle to popularna platforma do nauki programowania oraz uczenia maszynowego. Oprócz zadań dla początkujących organizowane są tam również rzeczywiste wyzwania – często z nagrodami finansowymi – polegające na zastosowaniu metod uczenia maszynowego do problemów, które nie zostały jeszcze w pełni rozwiązane lub zoptymalizowane.

Konkurs został zorganizowany przez zespół ATLAS przy eksperymencie LHC, który udostępnił publicznie symulacje wykorzystywane wcześniej w poszukiwaniach bozonu Higgsa. Zadanie polegało na opracowaniu algorytmu uczenia maszynowego zdolnego do skutecznego rozróżniania zdarzeń zawierających bozon Higgsa od tła, czyli zdarzeń imitujących interesujący fizyków sygnał.

Zainteresowanie wyzwaniem okazało się ogromne. Złożono ponad trzydzieści pięć tysięcy rozwiązań, co robi wrażenie nawet z dzisiejszej perspektywy – mimo znacznie większej popularności tej tematyki przeciętne konkursy na platformie Kaggle przyciągają zwykle około tysiąca zgłoszeń. Wydarzenie to jednak zapisało się w pamięci uczestników nie tylko ze względu na skalę zainteresowania, lecz także – a być może przede wszystkim – jako moment spopularyzowania niewielkiej wówczas biblioteki o nazwie XGBoost.

XGBoost jest algorytmem opartym na zespołach drzew decyzyjnych. Same drzewa decyzyjne działają poprzez zadawanie sekwencji prostych pytań, które stopniowo dzielą zbiór danych na podzbiory o podobnych własnościach. Dla przykładu, pytanie „czy $x > 0$?” pozwala rozróżnić liczby dodatnie i ujemne. W praktyce drzewa decyzyjne stosują wiele takich rozgałęzień jednocześnie, opierając się na zestawie parametrów wejściowych używanych do przewidywania wyniku. W przypadku danych z LHC mogłoby to być na przykład pytanie: „czy energia cząstki jest większa niż 20 GeV?”. W swojej klasycznej postaci algorytm ten analizuje możliwe podziały przestrzeni parametrów i wybiera te, które prowadzą do najlepszego dopasowania modelu do danych.

Drzewa decyzyjne stanowią podejście odmienne od powszechnie dziś stosowanych metod opartych na sieciach neuronowych. Choć są rozwiązaniem prostszym, często osiągają porównywalne, a niekiedy nawet lepsze wyniki. Czytelnikom *Delty*, którzy chcą rozpocząć naukę uczenia maszynowego, proponuję pokrótce zapoznać się z drzewami decyzyjnymi, zanim sięgną po bardziej złożone modele.

XGBoost

Jeszcze w latach dziewięćdziesiątych CERN (Europejska Organizacja Badań Jądrowych) był pionierem w rozwoju technologii informatycznych. To właśnie tam powstał i został przetestowany system WWW (World Wide Web), zanim został udostępniony publicznie.

Również w CERN-ie już w tamtym okresie szeroko stosowano metody, które dziś zaliczamy do uczenia maszynowego. Sieci neuronowe wykorzystywano między innymi w analizie danych z akceleratora LEP – bezpośredniego poprzednika LHC. CERN należał wówczas do jednych z pierwszych ośrodków prowadzących badania w tym zakresie, co – biorąc pod uwagę ogromne ilości przetwarzanych danych – nie powinno dziwić. Obecnie sytuacja wygląda jednak inaczej: CERN w większym stopniu korzysta z postępów w dziedzinie uczenia maszynowego, niż sam je wyznacza.

Kiedy zatem CERN dał się wyprzedzić środowisku komputerowych geeków? Choć na to pytanie nie ma jednoznacznej odpowiedzi, za symboliczną datę można uznać 12 maja 2014 roku. Tego dnia na platformie Kaggle rozpoczęło się wyzwanie *Higgs Boson Machine Learning Challenge*.

Ta pozornie prosta metoda jest dodatkowo wzmocniana za pomocą tzw. przyspieszenia gradientowego (ang. *gradient boosting*). Polega ono na sekwencyjnym trenowaniu wielu drzew decyzyjnych, przy czym w każdej kolejnej iteracji proces uczenia koncentruje się na obszarach przestrzeni danych, które zostały najslabiej opisane przez poprzednie modele. Do poruszania się po przestrzeni parametrów wykorzystuje się przybliżoną funkcję straty oraz jej gradient i krzywiznę. Ostateczna predykcja powstaje jako średnia przewidywań wszystkich drzew wchodzących w skład modelu.

Algorytm ten stanowił podstawę większości skutecznych rozwiązań zgłoszonych w ramach *Higgs Boson Machine Learning Challenge*, choć żadne z nich nie zajęło ostatecznie miejsca na podium. Jedynym oficjalnie wyróżnionym twórcą został Tianqi Chen – autor algorytmu XGBoost – który otrzymał nagrodę specjalną. Co jednak istotniejsze, konkurs ten znacząco przyczynił się do spopularyzowania XGBoost w środowisku zajmującym się uczeniem maszynowym.

Obecnie XGBoost uchodzi za jedno z najlepszych narzędzi w swojej klasie, szczególnie dobrze sprawdzające się w zadaniach opartych na danych tabularycznych. Do jego głównych zalet należą bardzo szybka implementacja oraz stosunkowo wysoka – w porównaniu z sieciami neuronowymi – interpretowalność wyników. W 2022 roku biblioteka ta była trzecią najczęściej używaną biblioteką uczenia maszynowego, ustępując jedynie pakietom Scikit-learn i Keras, czyli rozbudowanym platformom służącym do budowy i trenowania różnorodnych modeli, w tym sieci neuronowych.

Oprócz XGBoost-a dużą popularnością cieszą się obecnie również biblioteki LightGBM oraz CatBoost. Opierają się one na tych samych zasadach, a wybór konkretnego algorytmu zależy zazwyczaj od charakteru i rozmiaru zbioru danych, na którym trenowany jest model.

A CERN? W rzeczywistości instytucja ta nadal prowadzi intensywne badania z zakresu uczenia maszynowego, lecz nie pełni już roli głównego motoru rozwoju algorytmów. Pałeczka pierwszeństwa znajduje się dziś gdzie indziej...

Wiktor MATYSZKIEWICZ