Podstawy uczenia maszynowego

25.03.2024

Laboratorium 4

Lasy losowe

Łukasz Stępień, Kacper Fus

1. Cel zadania

Celem zadania jest zapoznanie się z metodami bazującymi na drzewach decyzyjnych. W trakcie nauki wykorzystano zestaw danych „adult”.

1. Implementacja

2.1 Preprocesing danych

* załadowano dane z pliku adult-all.csv
* brakujące dane uzupełniono najczęściej występującymi w danej kolumnie
* zmienne numeryczne poddano standaryzacji
* zmienne kategoryczne poddano etykietowaniu
  1. Klasyfikatory
* Utworzono dwa klasyfikatory: RandomForest oraz ExtraTrees
* Przetestowano różne konfiguracje parametrów:

|  |  |  |
| --- | --- | --- |
| Parametr | Opcja 1 | Opcja 2 |
| Liczba trenowanych drzew | 100 | 300 |
| Minimalny rozmiar podzielnego węzła drzewa | 2 | 10 |
| Minimalny rozmiar liścia | 1 | 10 |

1. Wyniki

* Random Forest

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

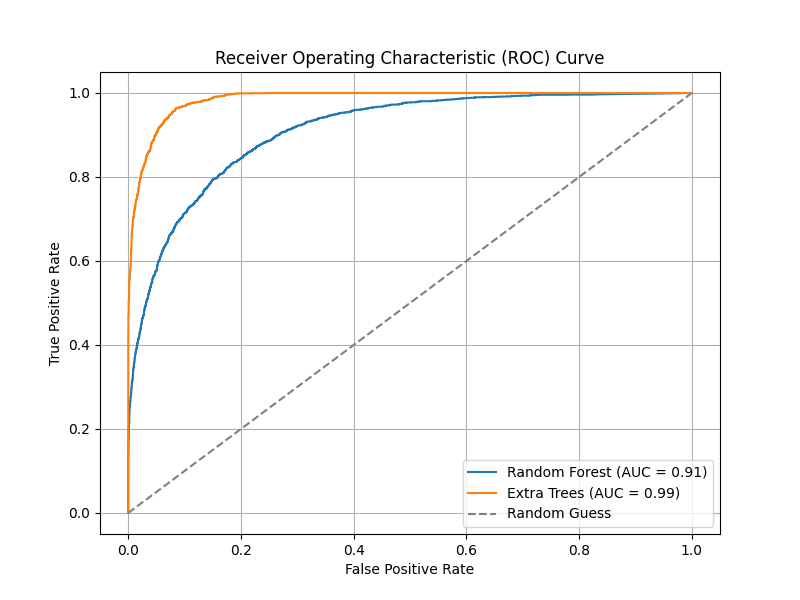
* A screenshot of a computer

  Description automatically generatedExtra Trees

A screenshot of a computer

Description automatically generated

* ROC



1. Wnioski:

* Najlepszy zestaw parametrów w obu klasyfikatorach był jednakowy (zaznaczony na zielono w tabelce), lecz klasyfikatory nie były zbytnio czułe na rozważane parametry.
* Skuteczność klasyfikatorów na etapie walidacji jest do siebie zbliżona, lecz na etapie testowania (ROC) Extra Trees jest znacząco lepsze od Random Forest.
* W obu przypadkach przydatność cech do klasyfikacji jest do siebie zbliżona. W czołówce występują: relationship, capitalgain, education\_num, age. Najmniej ważne były: race, sex,workclass, native\_country.
* Jakość klasyfikatora można ocenić biorąc pod uwagę wiele czynników, takich jak dokładność predykcji, precyzja, czułość, specyficzność, a także krzywa ROC czy obszar pod krzywą ROC (AUC-ROC).
* Największym problemem był czas oczekiwania na krosową walidacje klasyfikatorów dla różnych zestawów parametrów.