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”.

2. 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

2.2 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

3. Wyniki

- Random Forest

```
Random Forest:
Accuracy: 0.8639 (+/- 0.0040)
Precision: 0.7649 (+/- 0.0097)
Recall: 0.6214 (+/- 0.0141)
min_samples_leaf: 1
min_samples_split: 10
n_estimators: 100
```

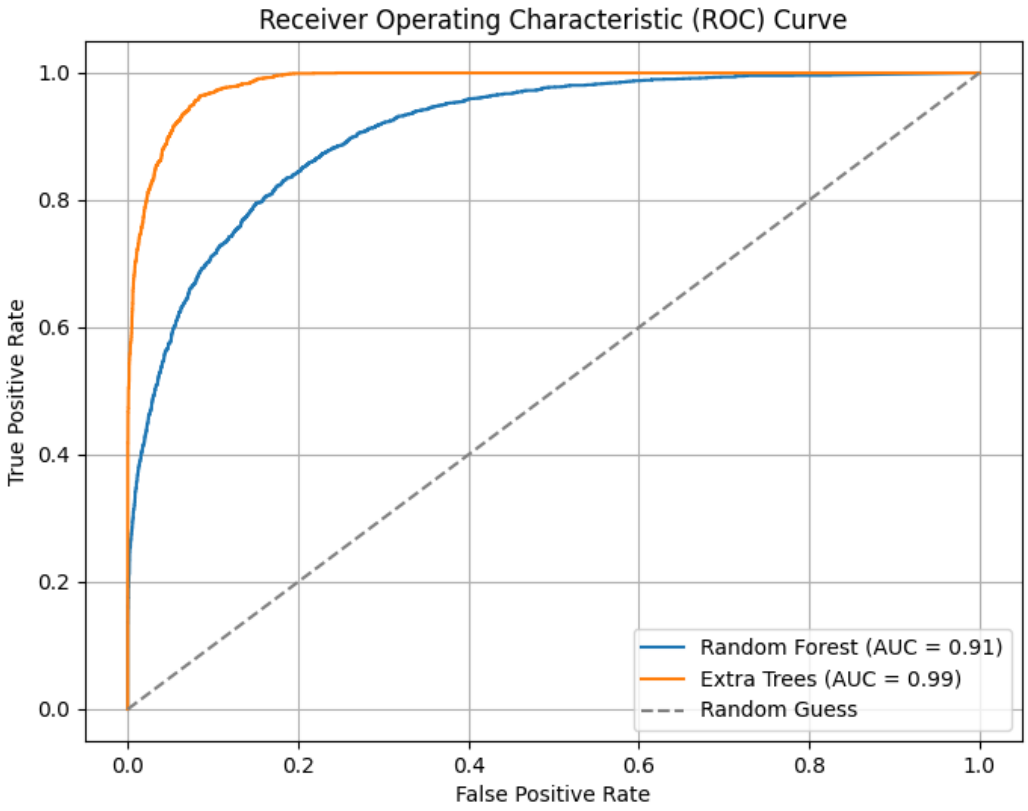
```
Random forest features (sorted):
1. capitalgain: 0.15
2. relationship: 0.14
3. age: 0.13
4. education-num: 0.11
5. fnlwgt: 0.11
6. marital-status: 0.08
7. hoursperweek: 0.07
8. occupation: 0.05
9. capitalloss: 0.04
10. education: 0.04
11. workclass: 0.03
12. native-country: 0.01
13. sex: 0.01
14. race: 0.01
```

- Extra Trees

```
Extra trees:
Accuracy: 0.8575 (+/- 0.0044)
Precision: 0.7498 (+/- 0.0123)
Recall: 0.6055 (+/- 0.0118)
min_samples_leaf: 1
min_samples_split: 10
n_estimators: 300
```

```
Extra trees features (sorted):
1. relationship: 0.14
2. capitalgain: 0.13
3. education-num: 0.13
4. marital-status: 0.12
5. age: 0.1
6. hoursperweek: 0.07
7. occupation: 0.06
8. fnlwgt: 0.05
9. education: 0.05
10. sex: 0.05
11. capitalloss: 0.04
12. workclass: 0.04
13. native-country: 0.01
14. race: 0.01
```

- ROC



4. 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.