Mini-AutoML

Marta Balcerzak Michał Dębski
Maciej Koczorowski

Plan prezentacji

- Selekcja modeli
- System AutoML
  - Preprocessing
  - Metryka stabilności
  - Wybór optymalnego modelu
  - Meta-learner

Selekcja modeli

Algorytm	Hiperparametr	Granica dolna	Granica górna
XGBoost	n_estimators	50	501
	learning_rate	0.01	0.51
	max_depth	2	10
	subsample	0.5	1
	colsample_bytree	0.5	1
	gamma	0	5

- AutoGluon
- RandomSearchCV
- Heurystyki
  - 8 RF, 10 XGB, 10 CB, 6 ET, 5 kNN
- Rankingi
- Wybór zbiorów danych
  - Zmienne jakościowe
  - Duże (i małe) zbiory
  - Multiclass

Działanie systemu

- Zbiór treningowy
- Preprocessing
- Wybór podzbioru modeli
- Ranking rozważanych modeli
- Trenowanie najlepszego modelu

Preprocessing

- Zmienne numeryczne
  - Uzupełnianie braków danych – mediana
  - Przekształcenia - standaryzacja
    - Zmienne numeryczne skośne dodatnio są przekształcane za pomocą: *log(x+1)*
    - Zmienne numeryczne skośne ujemnie są przekształcane za pomocą: *log(max(x)-x+1)*
- Zmienne jakościowe
  - Uzupełnianie braków danych - najczęstsze obserwacje
  - Przekształcenia - One-Hot Encoding

Metryka stabilności

Tutaj będzie to funkcja, która będzie przyjmować dwie wartości:

- Wektor wyników po walidacji krzyżowej na bazie miary *"balanced accuracy"*;
- Parametr odpowiedzialny za "karanie" za niestabilność.

Na bazie wyników tej funkcji będziemy tworzyć odpowiedni ranking modeli.

$$f(\theta, x) = \bar{x} - \theta \cdot S_x$$

Wiedza ekspercka

Pakiet z doświadczenia, patrząc na zbiór treningowy, będzie wybierał odpowiedni podzbiór z listy modeli.

Decyzja będzie się opierać przede wszystkim na liczbie zmiennych, obserwacji, zmiennych skośnych dodatnio oraz ujemnie, a także niezrównoważeniu klas.

Wybór modelu

Pracując na podzbiorze modeli za pomocą walidacji krzyżowej otrzymamy wyniki metryki stabilności.

Na podstawie trzech najlepszych modeli rozważymy pewne komitety (voting oraz stacking).

Mając wyniki metryki dla modeli z rozważanego podzbioru oraz dla komitetów wybieramy najlepszy model do treningu.

Meta learning

Zbiór, który będzie odpowiadał za wybieranie podzbioru modeli będziemy tworzyć następująco:

- Rozważamy 10 zbiorów danych o różnych właściwościach;
- Dla każdego zbioru będziemy tworzyć obserwację o zmiennych takich jak: rozmiar zbioru, czy niezrównoważenie klas;
- Ponadto dla każdego zbioru danych rozważamy wszystkie modele z listy i na podstawie metryki stabilności wybieramy 10 najlepszych modeli.