Lasy losowe

Zadanie 1. Celem zadania jest zapoznanie się z metodami bazującymi na drzewach decyzyjnych (ang. decision trees). Pracować będziemy na zbiorze `adult`. Brakujące wartości cech w tym zbiorze oznaczone są symbolem `?`. Wytrenuj klasyfikatory RandomForest oraz ExtraTrees (ang. extremely randomized trees) i dokonaj oceny ich działania.

Na działanie klasyfikatorów RandomForest i ExtraTrees wpływa kilka parametrów, m.in.:

- liczba trenowanych drzew,
- liczba cech branych pod uwagę przez drzewo przy podziale węzła
- minimalny rozmiar podzielnego węzła drzewa,
- minimalny rozmiar liścia,

Wypróbuj różne kombinacje tych parametrów, niekoniecznie wszystkie, jeżeli uczenie trwa zbyt długo. Skorzystaj z techniki zagnieżdżonej kroswalidacji (ang. nested cross-validation) aby ustalić nieobciążoną (ang. unbiased) dokładność (ang. accuracy), precyzję (ang. precision) i czułość (ang. recall) oraz ich odchylenie standardowe [1].

Powyższe klasyfikatory pozwalają na ustalenie, które cechy są przydatne do klasyfikacji (a które nie) na podstawie tego, jak często były wykorzystywane do tworzenia węzłów drzewa (i jaki procent informacji zawartej w zbiorze tłumaczyły te węzły). W bibliotece scikit-learn informacje te zwracane są w zmiennej `feature_importances_`. Które cechy naszych obserwacji zostały uznane za najważniejsze, a które za najmniej ważne? Czy jest to zgodne z intuicją?

Klasyfikatore te pozwalają też na ocenę pewności siebie co do dokonanej predykcji, wyrażoną jako odsetek wewnętrznych drzew głosujących za daną klasą. To oznacza, że może mieć on zmienną (i ustalaną przez użytkownika) czułość: np. potrzebować tylko 35% głosów za daną klasą, albo aż 70%. Wpływ wybranego progu czułości na zachowanie klasyfikatora (w tym jego precyzję i czułość) można zobrazować z użyciem tzw. krzywej ROC. Przedstaw wykresy tej krzywej dla powyższych klasyfikatorów. Które punkty na nich przedstawiają użyteczne konfiguracje końcowe?

Ww wnioskach sprawozdania omów następujące kwestie:

[1] `https://github.com/rasbt/model-eval-article-supplementary/blob/master/code/nested_cv_code.ipynb`

- Czy klasyfikatory były czułe na rozważane parametry?
- Czy skuteczność oszacowana na etapie walidacji była zgodna ze skutecznością ustaloną na etapie testowania?
- Czy istnieje wyraźne powiązanie między użytmi cechami, a wynikiem predykcji?
- Czy łatwo ocenić jakość klasyfikatora?
- Co sprawiało największy problem?

Literatura

[1] Sebastian Raschka, Model Evaluation, Model Selection and Algorithm Selection in Machine Learning, 2018.