Lasy losowe

Pola Lubińska

BAGGING

Jeżeli:

$Z_1, \ldots, Z_n$ - niezależne zmienne losowe z wariancją σ^2

To wtedy:

$\overline{Z}$ - średnia z tych obserwacji ma wariancję σ^2/n

Uśrednienie zbioru obserwacji pozwala zredukować wariancję.

BAGGING

Idea:

Policzyć $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ używając B osobnych zbiorów treningowych.

Dokońać uśrednionej predykcji:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x)$$

BOOTSTRAP

W jaki sposób uzyskać B osobnych zbiorów treningowych?

Użyć bootstrapu.

Wygeneruj B osobnych zbiorów treningowych losując B razy z powtórzeniami obserwacje ze zbioru treningowego.

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

BOOTSTRAP

Uwagi:

- Nie ucinamy drzew, gdy budujemy drzewa na próbkach bootstrapowych.
- Gdy dokonujemy klasyfikacji to zmiast brać średnią, bierzemy najczęściej przewidywaną klasę.
- Wybieramy relatywnie dużą liczbę (B=100).

ESTYMACJA BŁĘDU OUT OF BAG

Uśredniając, w baggingu do budowy drzewa używa się **2/3** obserwacji.

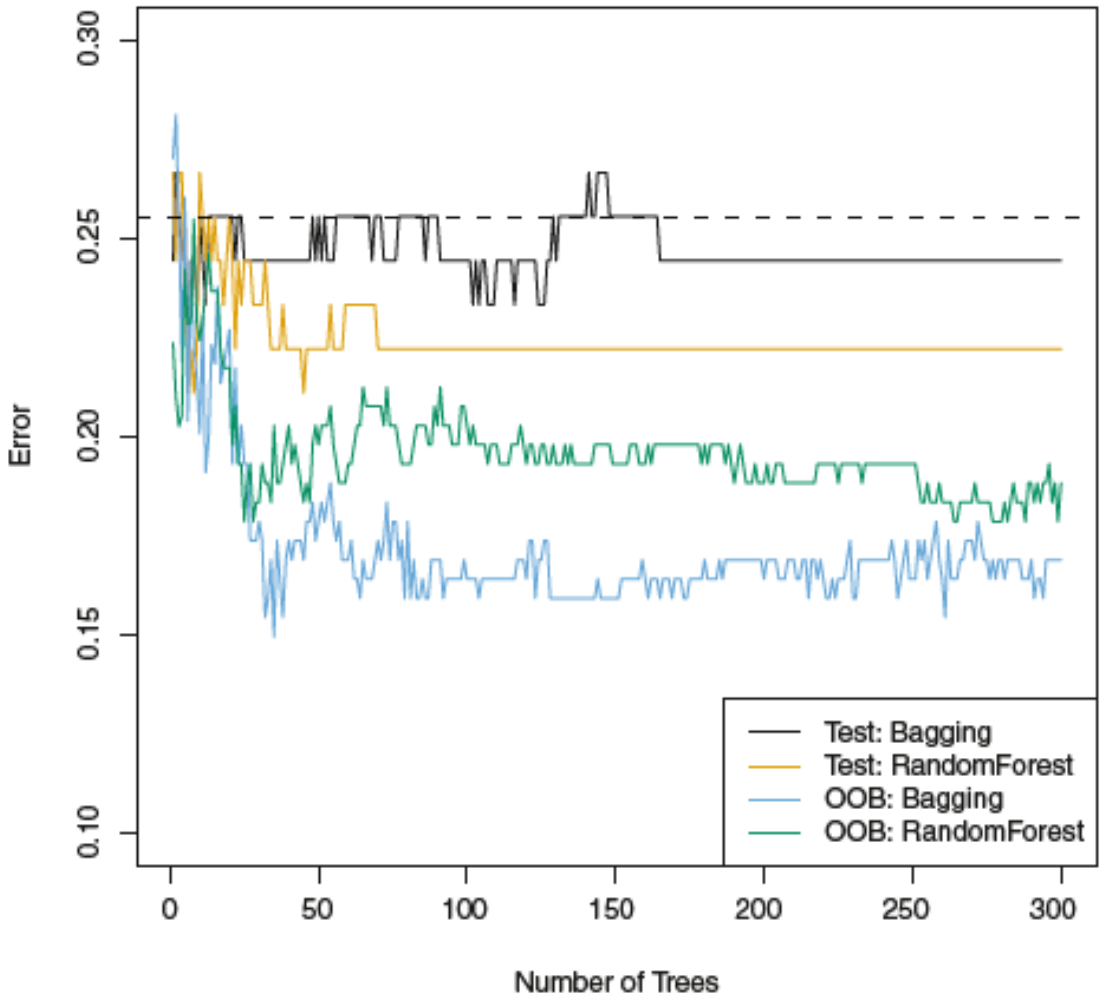
Dlatego możmy dokonać predykcji dla danego punktu, używając drzew, dla których ten punkt danych był **out of bag**.

Uśredniając, dla każdej obserwacji mamy około **B/3** drzew, które nie były trenowane na tej obserwacji.

Dlatego dla każdej obserwacji możemy dostać predykcję, biorąc średnią z drzew dla których ta obserwacja nie była w zbiorze treningowym.

Mając te estymacje możemy policzyć **MSE** lub błąd klasyfikacji.

PRZYKŁAD


Error
0.10
0.15
0.20
0.25
0.30
0
50
100
150
200
250
300
Number of Trees
Test: Bagging
Test: RandomForest
OOB: Bagging
OOB: RandomForest

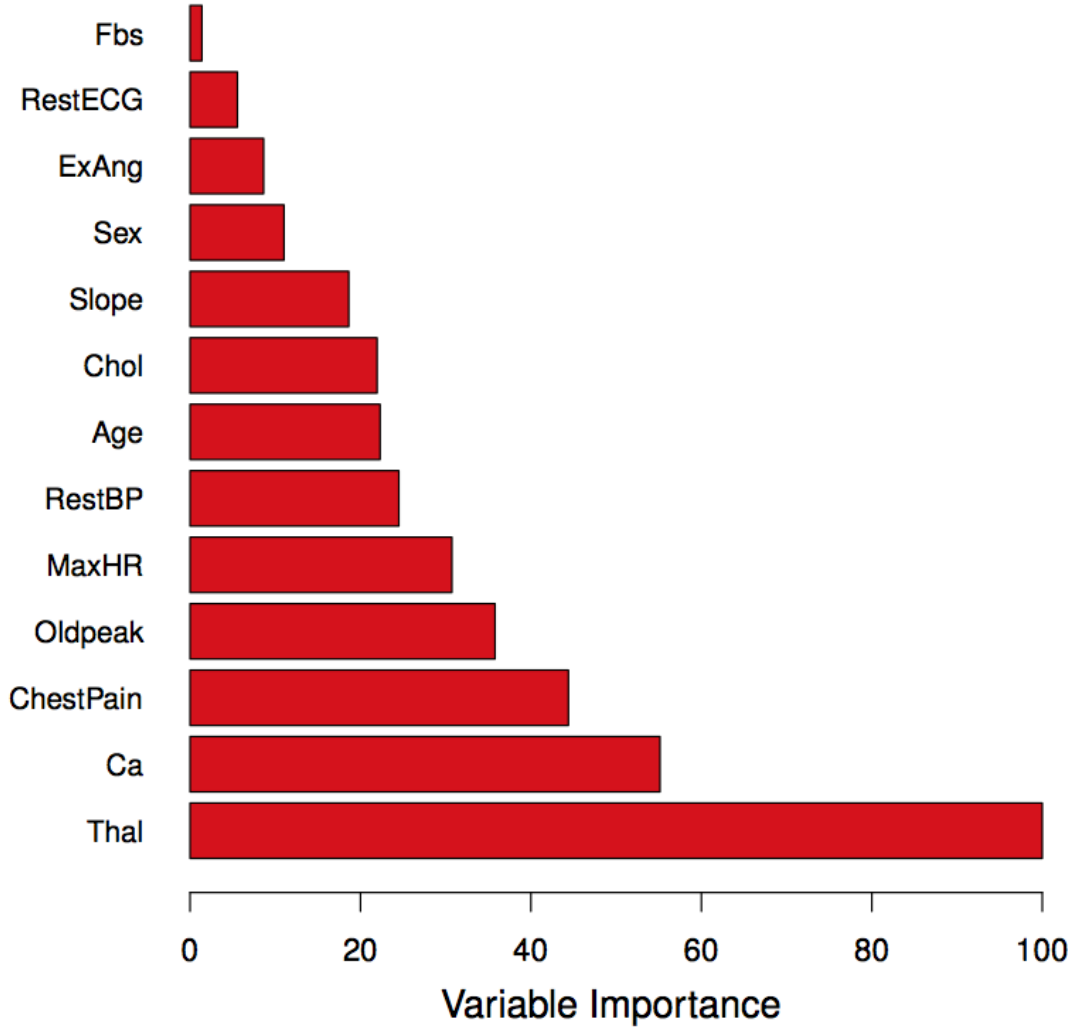
VARIABLE IMPORTANCE MEASURE

Używając baggingu tracimy interpretowalność.

Jednakże wciąż możemy:

- Zanotować o ile zmniejsza się **RSS** dla danego predyktora na każdym podziale
- Zanotować o ile mniejsza się **Gini Index** dla danego predyktora na każdym podziale

VARIABLE IMPORTANCE MEASURE



LAS LOSOWY

Algorytm lasu losowego jest pewną wariacją baggingu.

W algorytmie lasu losowego, gdy budujemy drzewo to na każdym podziale wykorzystujemy **m losowo wybranych** zmiennych objaśniających.

Zazwyczaj: m $\approx \sqrt{p}$.

LAS LOSOWY

Algorytm lasu losowego jest pewną wariacją baggingu.

W algorytmie lasu losowego, gdy budujemy drzewo to na każdym podziale wykorzystujemy **m losowo wybranych** zmiennych objaśniających.

Zazwyczaj: m $\approx \sqrt{p}$.

LAS LOSOWY

Dlaczego jest to pomocne?

Wyobraźmy sobie, że mamy jeden bardzo silny predyktor w zbiorze danych.

Większość drzew w baggingu użyje tego predyktora do pierwszego podziału.

W konsekwencji większość drzew będzie do siebie podobna (będą ze sobą skorelowane).

Uśrednianie wartości silnie ze sobą skorelowanych nie wpływa na redukcję wariancji.

LAS LOSOWY

Dlaczego jest to pomocne?

Wyobraźmy sobie, że mamy jeden bardzo silny predyktor w zbiorze danych.

Większość drzew w baggingu użyje tego predyktora do pierwszego podziału.

W konsekwencji większość drzew będzie do siebie podobna (będą ze sobą skorelowane).

Uśrednianie wartości silnie ze sobą skorelowanych nie wpływa na redukcję wariancji.

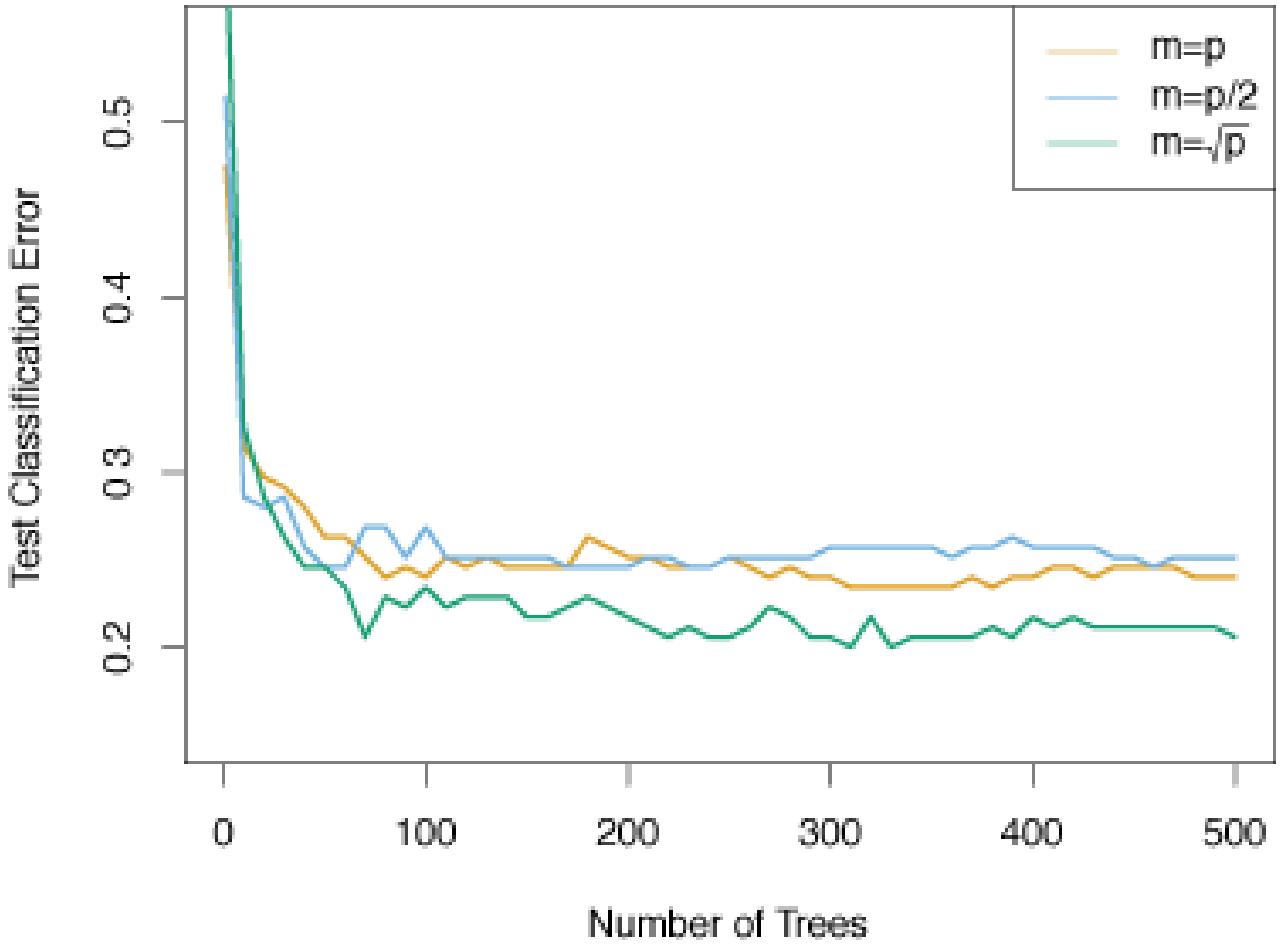
BAGGING VS LAS LOSOWY

Różni się tylko wyborem m.

Bagging: m=p

Las losowy: m=$\sqrt{p}$ (zazwyczaj)

BAGGING VS LAS LOSOWY



BOOSTING

Tak jak w bagginu, w boostingu buduje się dużą liczbę drzew decyzyjnych:

$$\hat{f}^1, \cdots, \hat{f}^B$$

Jednakże w boostingu drzewa dopasowywane są sekwencyjnie.

Każde drzewo używa informacji z poprzedniego drzewa.

However in boosting, trees are fitted sequentially. Every tree uses information from previous fitting.

W boostingu dopasowujemy drzewa nie do zmiennej objaśnianej Y, ale do reszt modelu.

BOOSTING

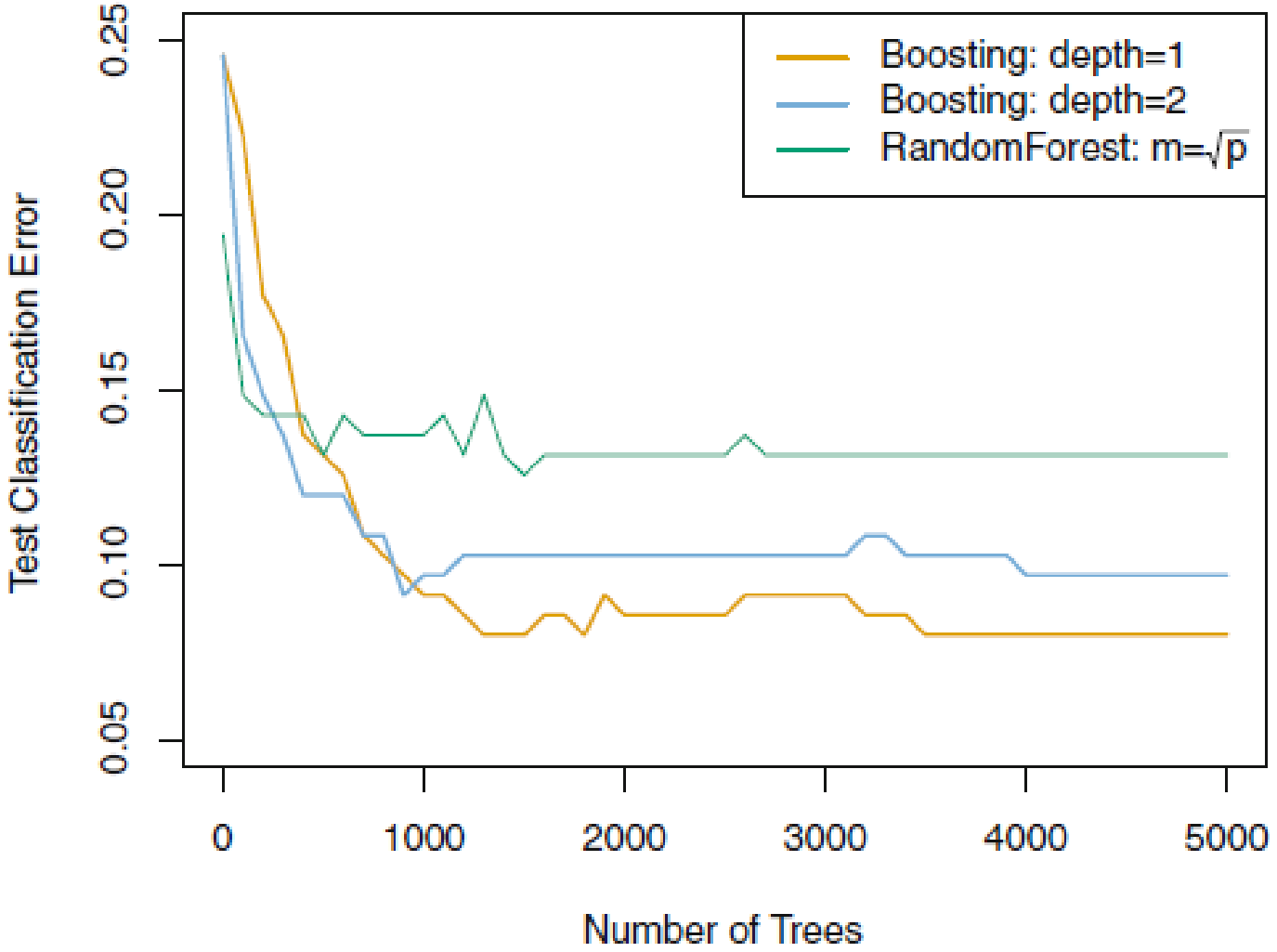
W Bostingu mamy 3 parametry:

- B - liczba drzew
  - w przeciwieństwie do baggingu zbyt duża liczba drzew może sprawić, że model będzie zbyt bardzo dopasowany do danych,
  - wybieramy używając kroswalidacji,

- λ -parametr ściągający
  - kontroluje learning rate,
  - zazwyczaj sa to wartości 0.01, 0.001, - bardzo mała wartość może wymagać dużej liczby drzew do osiągnięcia dobrych wyników,

- d - liczba podziałów w każdym drzewie
  - kontroluje złożonośc modelu,
  - nawet $d = 1$ zazwyczaj działa dobrze.

RANDOM FOREST VS BOOSTING



DZIĘKUJĘ ZA UWAGĘ!

Pola Lubińska