Wstęp do uczenia maszynowego

Metody drzewiaste

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

UNIWERSYTET
WARSZAWSKI

Drzewo regresyjne dla log zarobków w baseballu

Predyktory: liczba lat grania w lidze (Years) i liczba uderzeń, które wykonał w zeszłym roku (Hits).

Etykiety węzłów wewnętrznych: podziały. $X_j < t$ określa lewą gałąź. Prawa spełnia $X_j \geq t$. Np w lewej gałęzi są obserwacje spełniające Years < 4.5, a w prawej Years ≥ 4.5.

Etykiety liści: średnia ze zmiennej objaśnianej (logarytmu pensji rocznej uderzającego) dla obserwacji wpadających do tych liści.

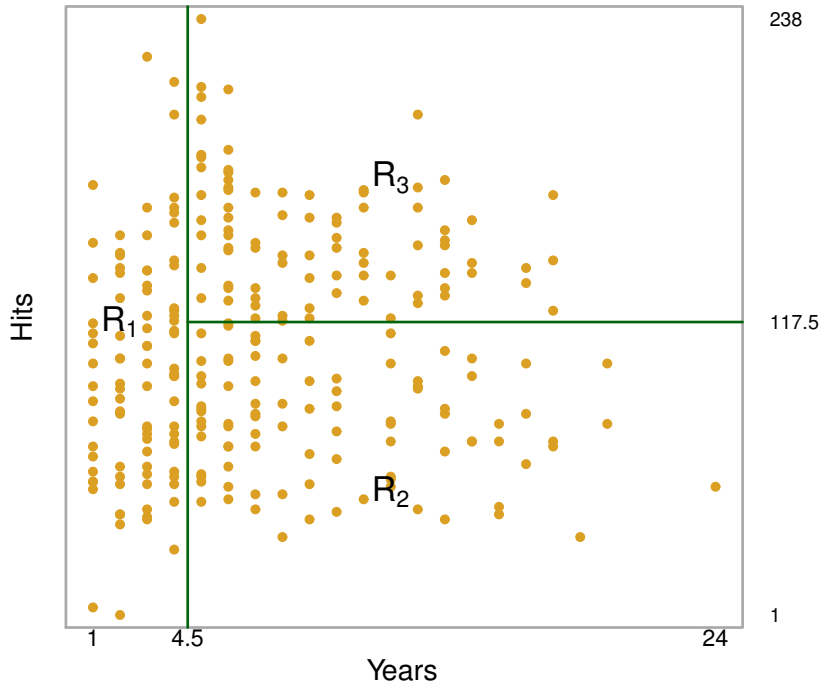
Predykcje na podstawie etykiet. Np dla obserwacji w lewym liściu pensja $e^{5.11}$ i.e. $165,67$ tysięcy dolarów.

Interpretacja:

- Years ważniejszym predyktorem.
- Jeśli gracz gra krótko (niedoświadczony), liczba uderzeń nie wpływa na jego zarobki.
- Ale już dla doświadczonych graczy (Years $\geq$ 4.5) wieksza liczba uderzeń zwieksza zarobki.

Interpretowalność jest główną zaletą drzew decyzyjnych.

Drzewa decyzyjne generują podział przestrzeni wartości predyktorów na rozłączne obszary



$$R_1 = \{X \mid Years < 4.5\},\ R_2 = \{X \mid Years \geq 4.5, Hits < 117.5\},$$
$$R_3 = \{X \mid Years \geq 4.5, Hits \geq 117.5\}.$$

Dwa kroki konstruowania drzew decyzyjnych

- Dzielimy przestrzeń wartości predyktorów na J rozłącznych obszarów $R_1, R_2, \ldots, R_J$. Obszary te są zwykle wielowymiarowymi prostopadłościanami.
- Dla każdej obserwacji, która wpada do obszaru R_j dajemy jako odpowiedź stałą c_j

$$f(x) = \sum_{j}^{J} c_j I(x \in R_j).$$

- Jako ocenę jakości danego wyboru podziału przyjmujemy RSS:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - c_j)^2,$$

gdzie y_i to odpowiedź dla zestawu i wartości predyktorów.

- Przy tej ocenie optymalne c_j to $\hat{y}_{R_j}$, średnia z odpowiedzi dla obszaru R_j.

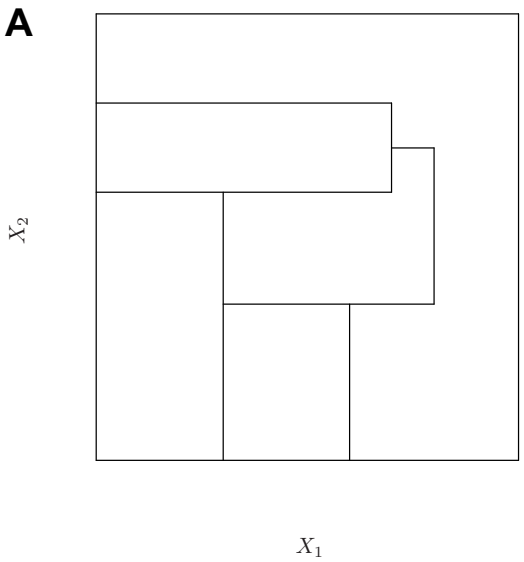
Rekurencyjne binarne dzielenie przestrzeni wartości predyktorów (algorytm zachłanny)

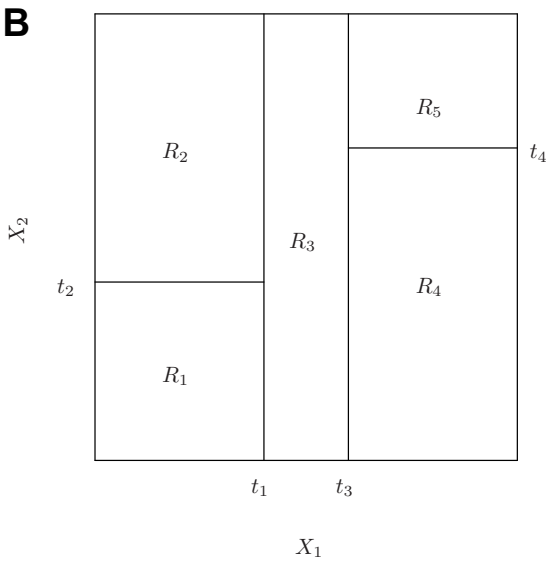
- Wybieramy ten X_j oraz *wartość odcięcia* s, tak aby podział regionu (początkowo całej przestrzeni predyktorów) na obszary $R_1(j,s) = \{X \mid X_j < s\}$ oraz $R_2(j,s) = \{X \mid X_j \geq s\}$ prowadził do maksymalnego spadku RSS. Czyli minimalizujemy wartość

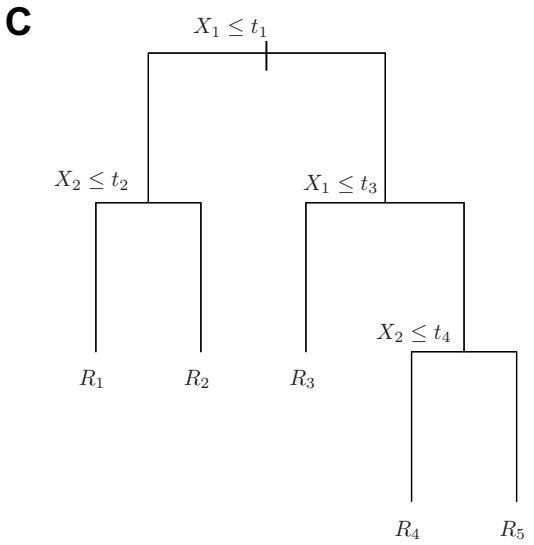
$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2.$$

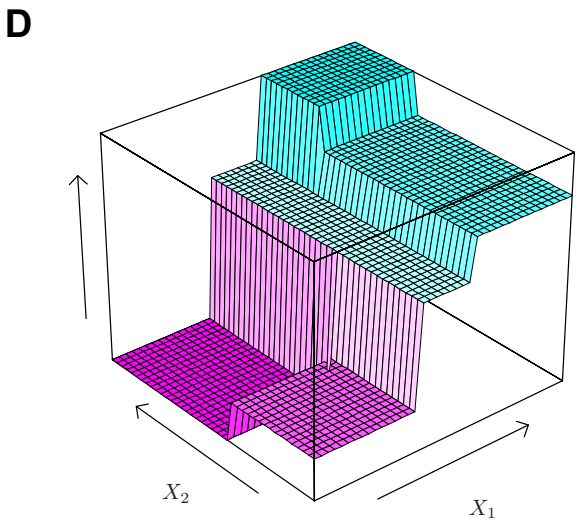
- Na każdym etapie tego procesu mamy pewien zbiór obszarów, które można dalej dzielić. **Kryterium stopu:** nie dzielimy obszarów mających mniej niż, np. 5 obserwacji.
- Po zakończeniu procesu konstruowania podziałów (czyli całego drzewa decyzyjnego), przypisujemy liściom wartość średnią odpowiedzi dla obszaru odpowiadającego temu liściowi.

Przykład podziału na 5 obszarów









A Podział, którego nie można otrzymać z rekurencyjnego binarnego dzielenia.

B Podział, który można otrzymać.

C Drzewo odpowiadające prawemu górnemu podziałowi.

D Wykres wartości predykcji dla tego drzewa.

Przycinanie drzewa

- Drzewa otrzymane metodą rekurencyjnego binarnego dzielenia mogą być zbyt duże, co często prowadzi do przeuczenia. W zawiązku z tym należy je przyciąć.
- **Uwaga:** w podręczniku jest niecodzienna definicja podrzewa – T jest *poddrzewem* drzewa T_0 jeśli T jest otrzymane z T_0 przez zastąpienie pewnej liczby węzłów wewnętrznych liśćmi.
- **Przycinanie najsłabszych gałęzi:** przy ustalonej wartości parametru α wybieramy podrzewo T (w powyższym sensie) drzewa T_0 otrzymanego metodą rekurencyjnego binarnego dzielenia tak aby zminimalizować koszt

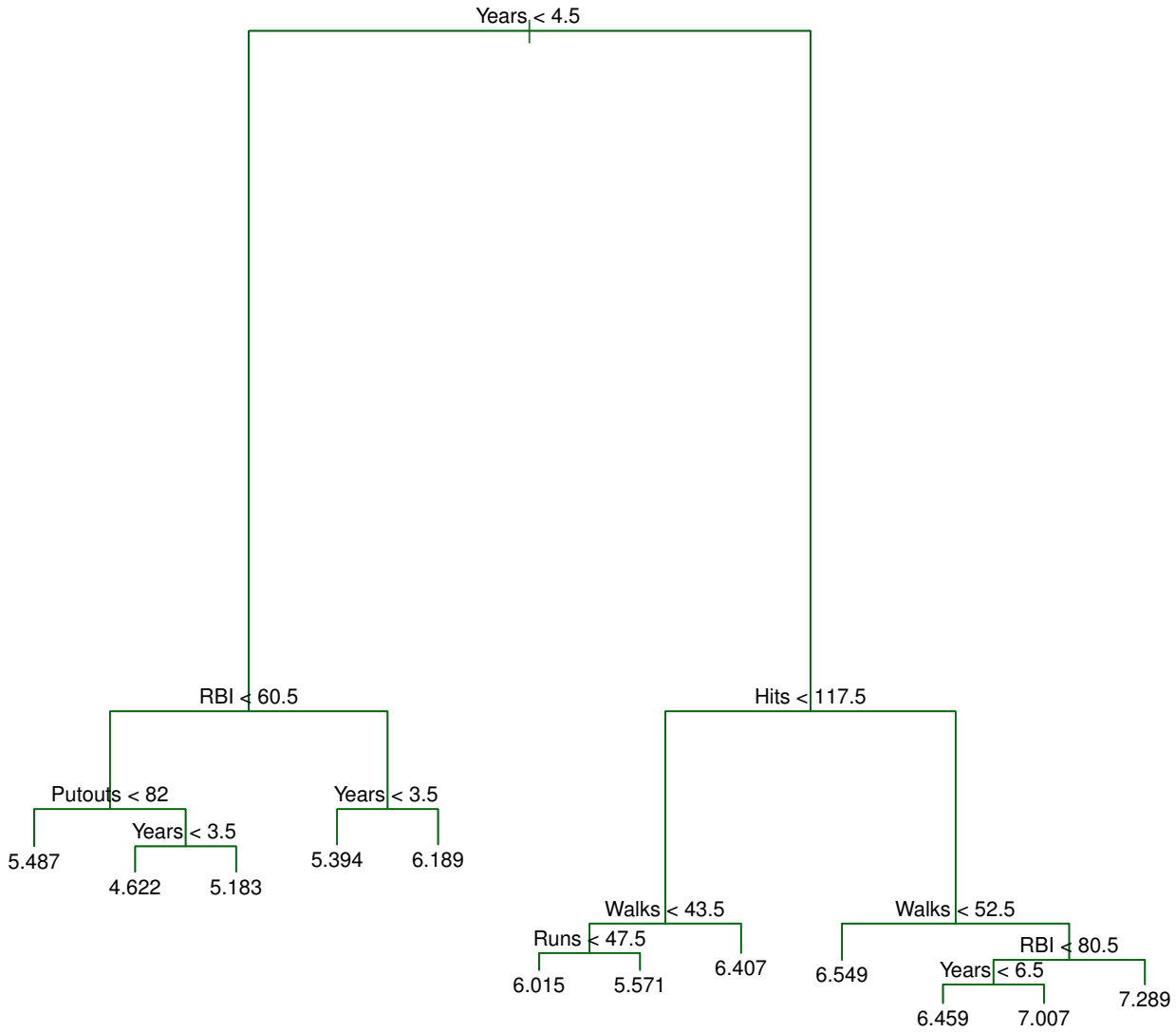
$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{i:\ x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|,$$

gdzie $|T|$ to liczba liści w T.

Przycinanie najsłabszych gałęzi

- Dla zadanego α istnieje dokładnie jedno T_α minimalizujące koszt $C_\alpha(T)$
- Procedura znajdowania T_α
  - Iteracyjnie zastępujemy liśćmi takie kolejne wierzchołki, które powodują najmniejszy przyrost członu $\sum_{m=1}^{|T|} \sum_{i:\ x_i \in R_m} (y_i - \hat{y}_{R_m})^2$ (w tym sensie najsłabsze)
  - postępujemy tak aż dojdziemy do korzenia
  - to generuje sekwencję poddrzew
  - można pokazać, że ta sekwencja zawiera T_α.

Przykład drzewa decyzyjnego przed operacją przycinania



Algorytm konstrukcji drzewa regresyjnego

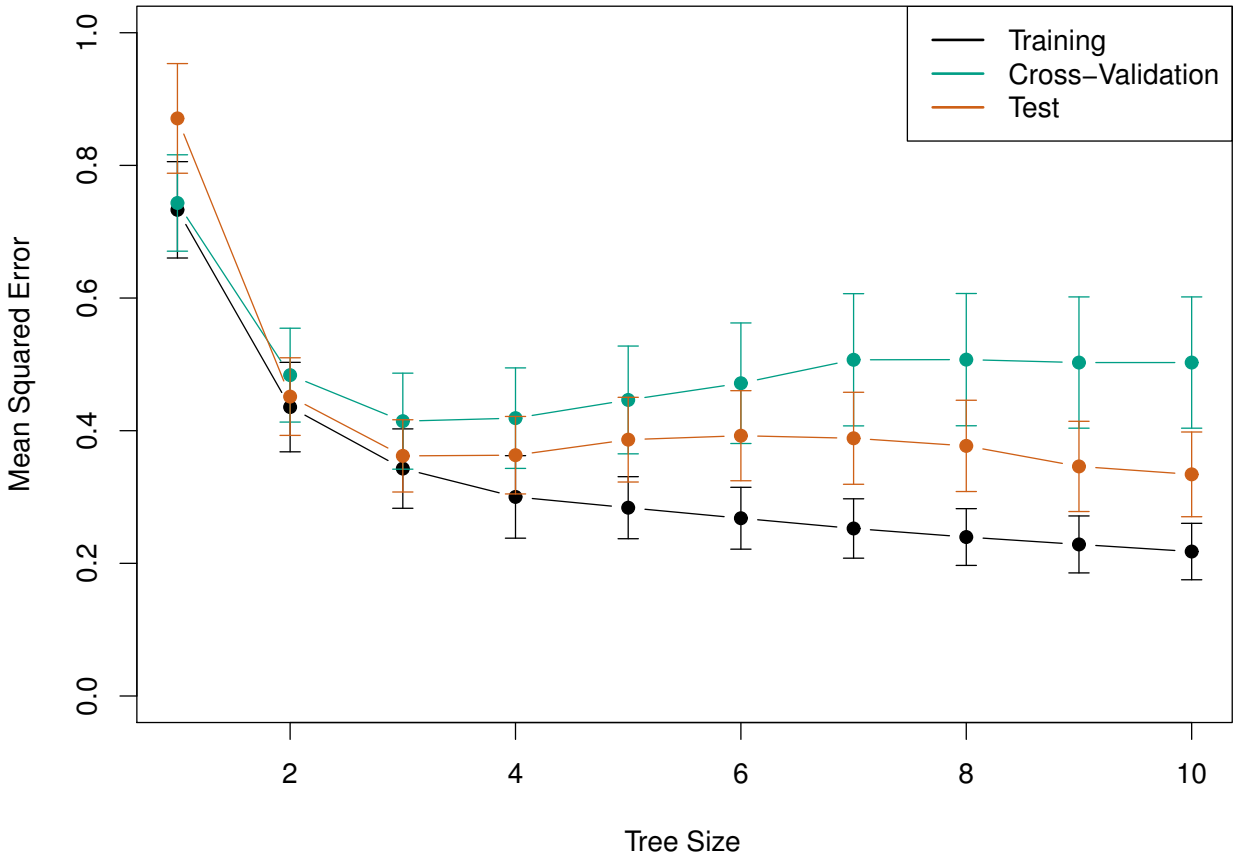
1. Stosując metodę rekurencyjnego binarnego dzielenia zbuduj duże drzewo decyzyjne, stosując jako kryterium stopu ustaloną progową liczbę obserwacji w otrzymanym obszarze.
2. Zastosuj metodę przycinania najsłabszych krawędzi, otrzymując ciąg przyciętych drzew jako funkcję od parametru α.
3. Użyj K-krotnej walidacji krzyżowej do oceny wyboru parametru α: Dla każdego $k = 1, \ldots, K$
   (a) Wykonaj kroki 1 i 2 na wszystkich danych, za wyjątkiem k-tej części.
   (b) Oblicz średni błąd kwadratowy predykcji dla k-tej części jako funkcję od α.

   Wyznacz α_0, przy którym średnia z błędów po wszystkich k jest najmniejsza.
4. Zwróć jako wynik drzewo z kroku 2 odpowiadające znalezionej wartości α_0.

Analiza poziomu błędu dla danych o zarobkach w baseballu

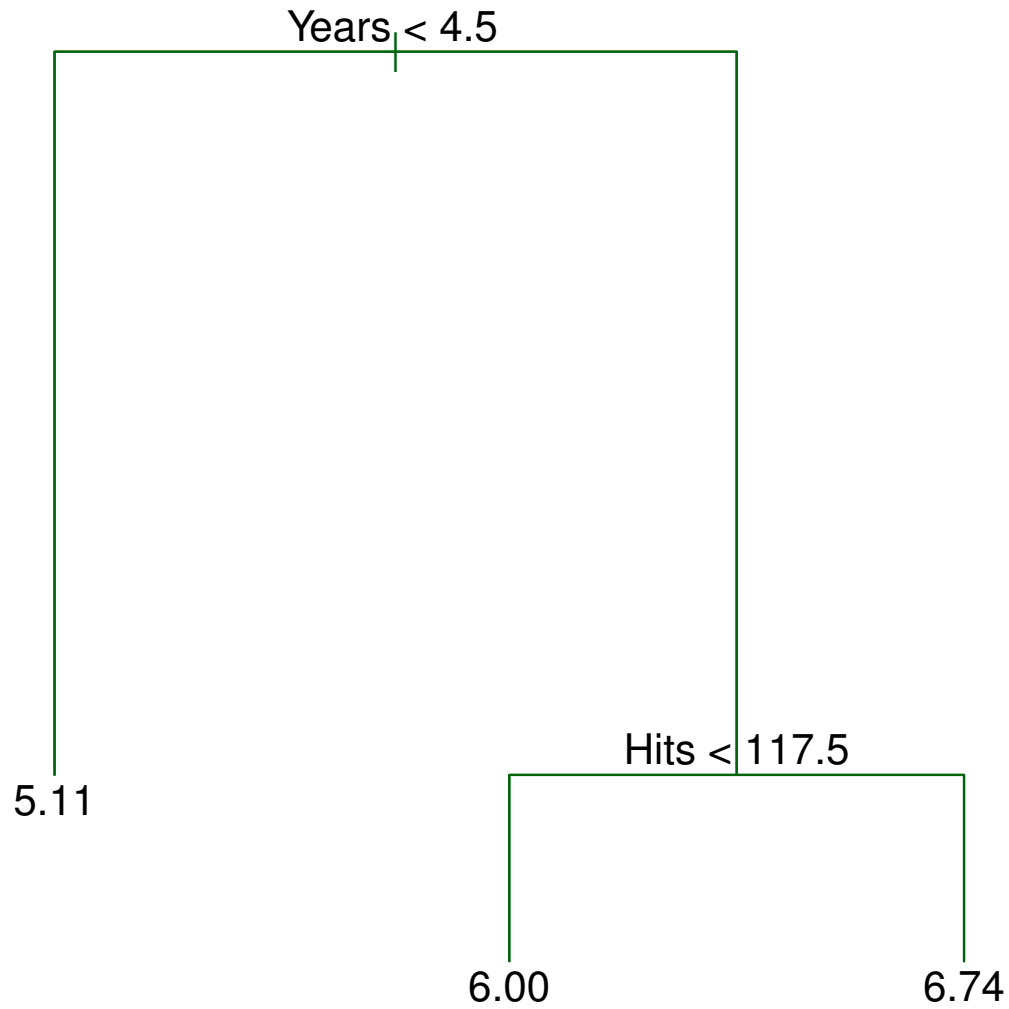
- Podział danych na 132 obserwacje w zbiorze treningowym i 131 w testowym.
- Zbudowanie dużego drzewa regresji na danych treningowych
- Zbudowanie poddrzew dla różnych wartości α
- Wykonanie 6-krokowej walidacji krzyżowej, estymując testowy MSE w zależności od α
- Porównanie wyestymowanego testowego MSE z rzeczywistym (policzonym na danych testowych)

Analiza poziomu błędu dla danych o zarobkach w baseballu



Walidacja krzyżowa wskazuje, że przycięte drzewo o 3 liściach daje najlepszy wynik.

Drzewo o minimalnym MSE



Drzewa klasyfikujące

Konstrukcja drzew klasyfikujących jest podobna do drzew regresyjnych

- Zamiast brać średnią jako predykcję dla danego obszaru (tak jak to było dla drzew regresyjnych) wybieramy tę odpowiedź, która jest najczęstsza wśród odpowiedzi z danego obszaru.
- Miara RSS nie nadaje się do wyboru podziału i jako miara jakości klasyfikacji obserwacji wpadających do regionu dla danego wierzchołka.
- Rozważmy jeden region (wierzchołek) R_m. Niech $\hat{p}_{mk}$ oznacza proporcję obserwacji treningowych z m-tego regionu, które należą do klasy (z klasyfikacji) k.
- $\max_k(\hat{p}_{mk})$: proporcja tych obserwacji w obszarze, które należą do klasy o największej częstości w tym obszarze.

Miary oceny "czystości" klasyfikacji dla obszaru

- **Błąd klasyfikacji**:

$$E_m = 1 - \max_k(\hat{p}_{mk}).$$

- **Indeks Giniego:**

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}).$$

- **Entropia krzyżowa:**

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \ \log(\hat{p}_{mk}).$$

- Miary te przyjmują wartości bliskie 0, gdy $\hat{p}_{mk}$ jest bliskie 1 (można wtedy powiedzieć, że węzeł m jest *czysty*).

Zastosowanie miar

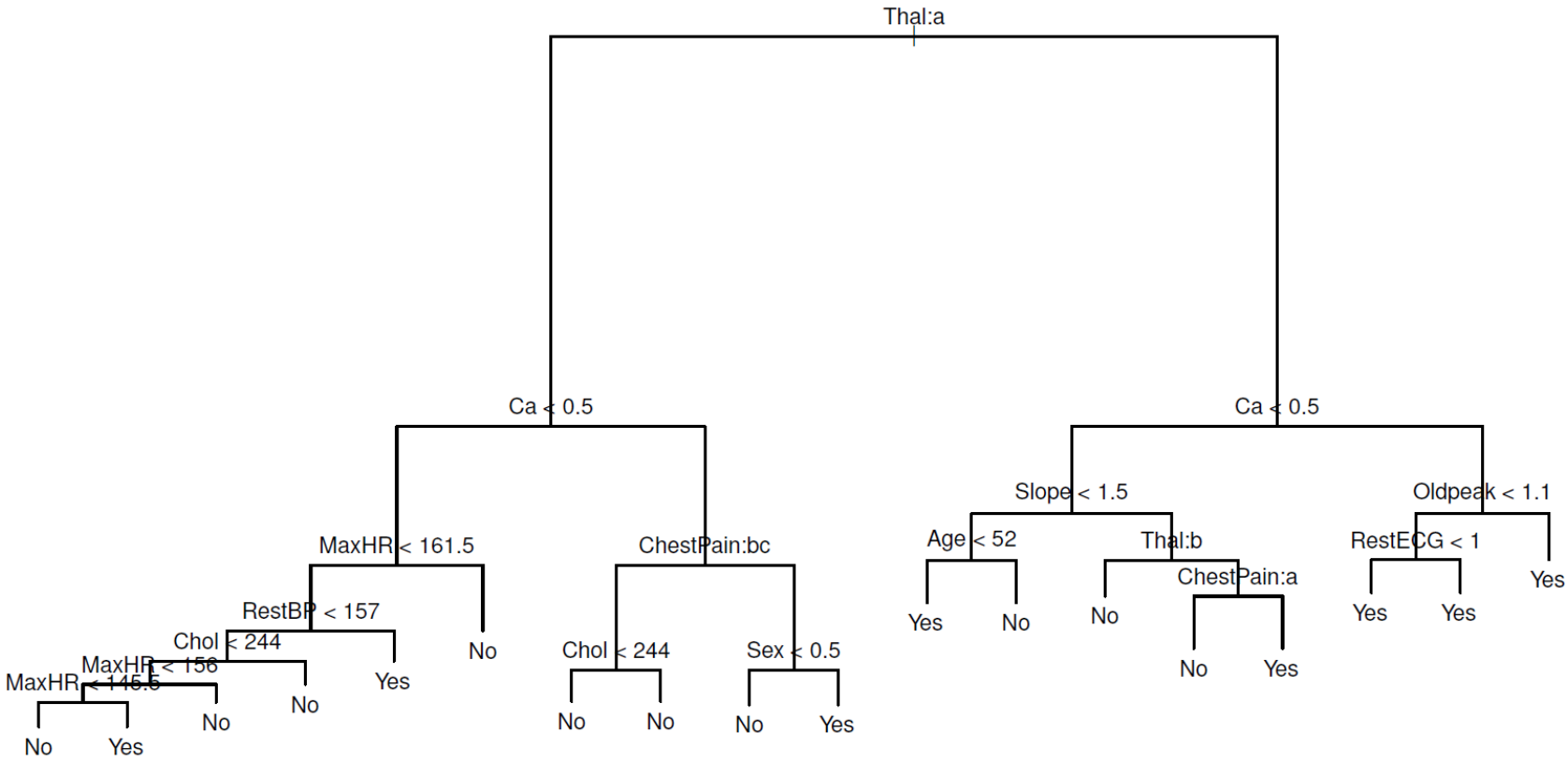
Najczęściej stosuje się

- Przy budowaniu drzewa, gdy oceniamy podział regionu w danej iteracji: Indeks Giniego bądź entropia krzyżowa (są bardziej wrażliwe na czystość wierzchołków).
- Przy ocenie klasyfikacji i przy przycinaniu drzewa: Błąd klasyfikacji.

Przykład: Dane 'heart'

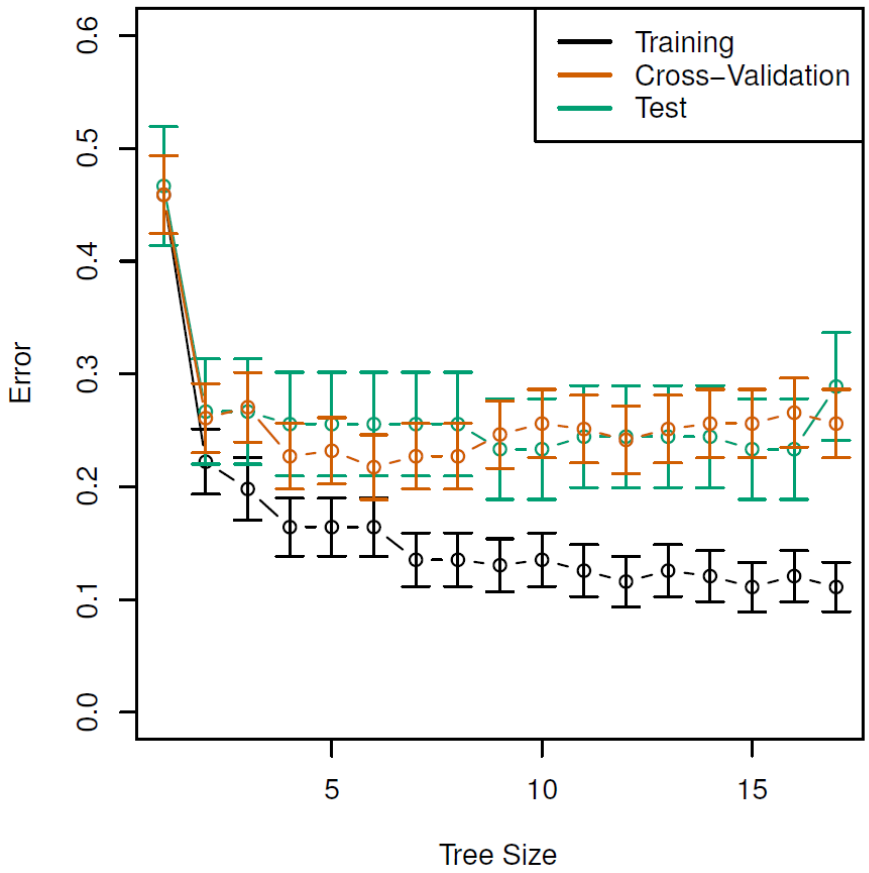
- 303 pacjentów z bólem w klatcie piersiowej.
- Klasy: chory na serce 'Yes', lub nie-chory 'No'.
- Łącznie 13 predyktorów, takich jak Age, Sex, Chol (poziom cholesterolu).

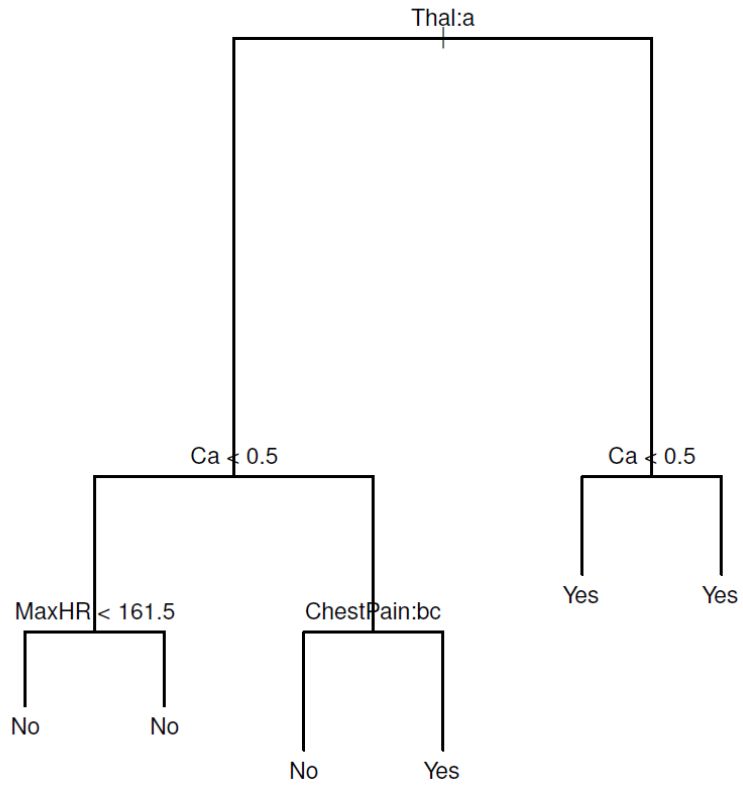
Pełne drzewo



Uwaga: niektóre podziały dają w wyniku liście o identycznych etykietach. To niepotrzebne ze względu na predykcję, ale zwiększa czystość węzłów.

Drzewo klasyfikacyjne po przycięciu



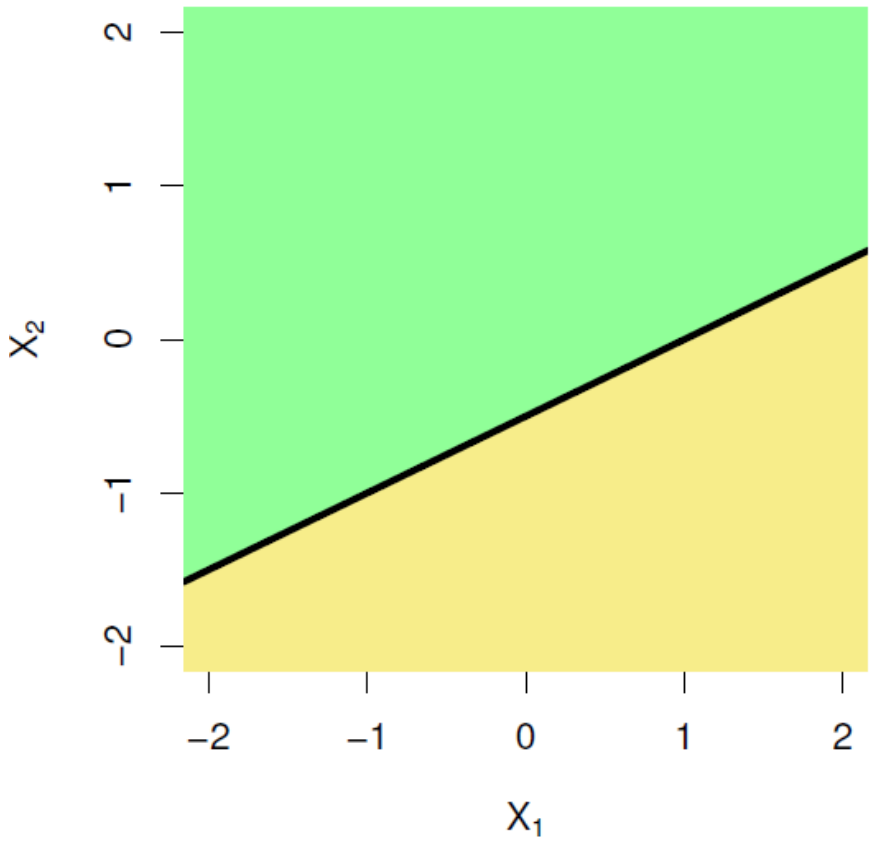


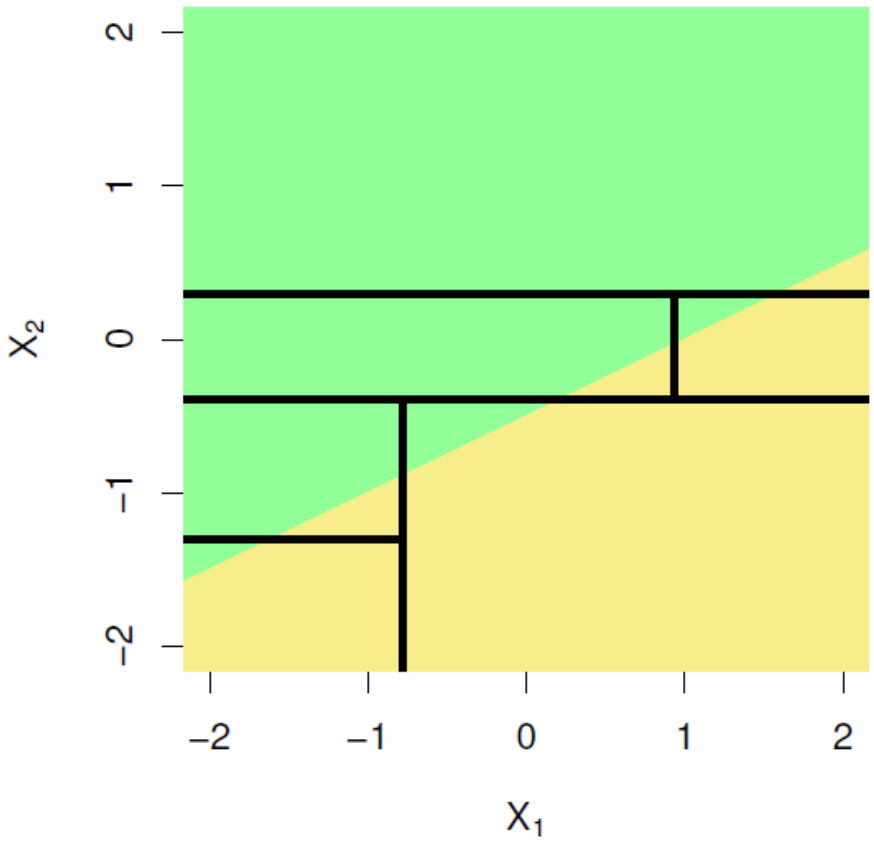
Minimalny błąd osiąga się przy drzewie o 6 liściach.

Podział przestrzeni przy predyktorach jakościowych

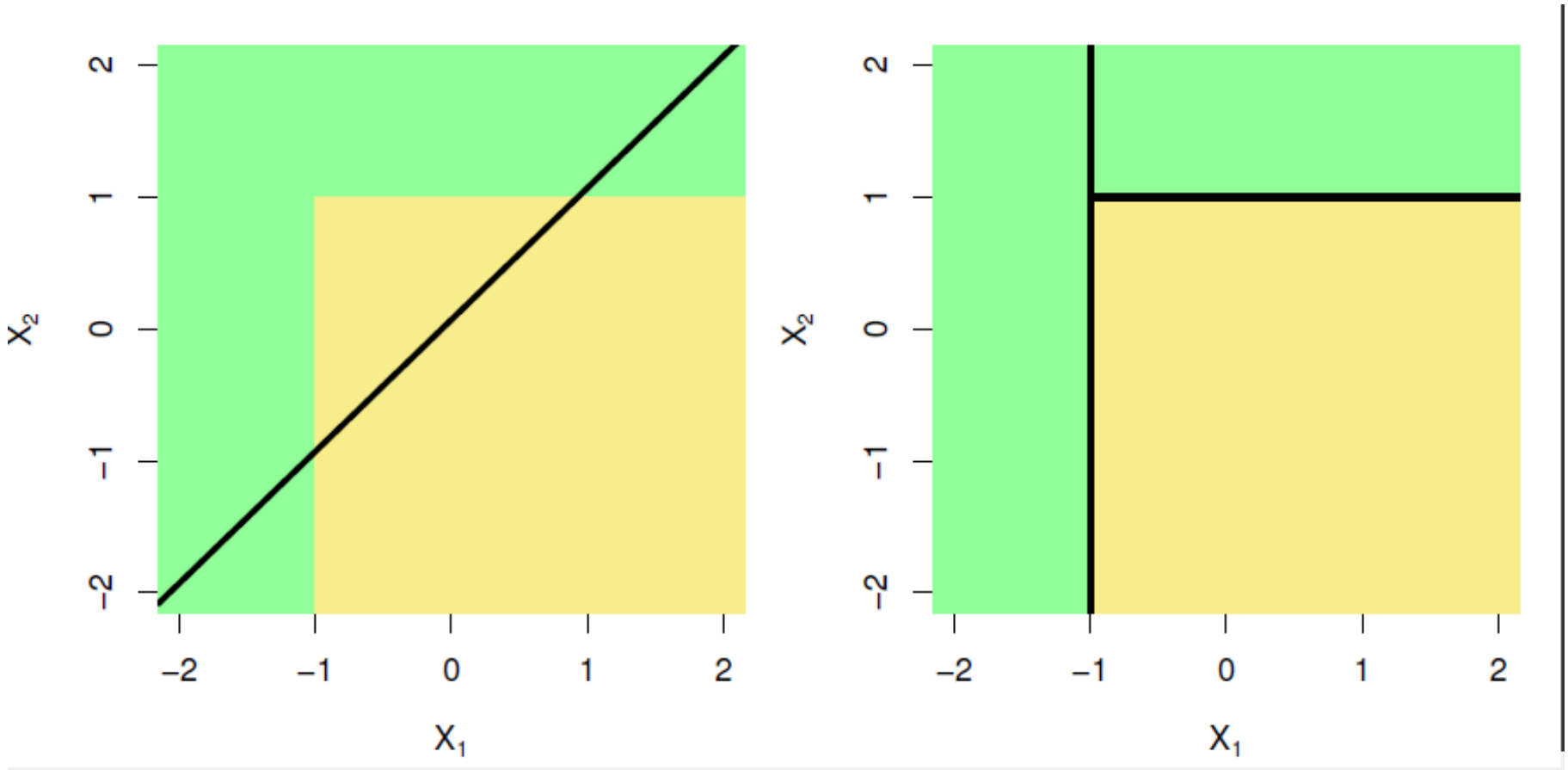
- Predyktor 'Thal' oznacza wynik testu stresowego przy pomocy talu. Możliwe wyniki to 'normal' lub 'stałe' lub 'odwracalne' uszkodzenia.
- Predyktor 'ChestPain' przyjmuje 4 możliwe wartości, między innymi: typowa dusznica, nietypowa dusznica, ból nie-dusznicowy.
- Dla predyktorów X przyjmujących wartości jakościowe (np. 'Thal', 'ChestPain') podział przestrzeni określa sie przez wypisanie które wartości prowadzą do lewego poddrzewa (a oznacza pierwszą wartość predyktora, b drugą, itd.). Niewymienione wartości prowadzą do prawego poddrzewa.

Liniowe modele mogą dawać lepsze wyniki niż drzewa





Ale drzewa mogą też dawać lepsze wyniki niż liniowe modele



Zalety i wady drzew decyzyjnych

Zalety

- Drzewa bardzo łatwo wytłumaczyć
- Drzewa odpowiadają sposobowi podejmowania decyzji przez (niektórych) ludzi
- Mają intuicyjną reprezentację graficzną
- Łatwo buduje się je w oparciu o kategoryczne (nominalne, jakościowe) zmienne, bez potrzeby generowania "dummy variables"

Wady

- Słabe wyniki w zastosowaniu do danych
- Duża wariancja - małe zmiany w danych mogą silnie wpłynąć na model

Metody poprawiania jakości predykcji drzewowych: bootstrap aggregation (bagging)

Bagging jako metoda zmniejszania wariancji

- Oparte na obserwacji, że jeśli mamy zmienne losowe $Z_1, \ldots, Z_n$ o tej samej wariancji σ^2, to wariancja średniej $\bar{Z}$ jest równa σ^2/n.
- Korzystamy z boostrap, czyli udajemy, że mamy B zbiorów treningowych.
- Wykonujemy bootstrap produkując B danych treningowych. Dla b-tych danych bootstrapowych konstruujemy funkcję predykcji $\hat{f}^{*b}(x)$ i następnie obliczamy średnią po wszystkich b

$$f_{bag}(x) = (1/B) \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Bagging jako metoda zmniejszania wariancji

- Tu pokazujemy jako metodę zmniejszania wariancji drzew decyzyjnych
- Jest to podejście ogólne, które można zastosować aby zredukować wariancję różnych metod.
- W przypadku drzew regresyjnych:
  - budujemy B drzew regresji na B prób z bootstrap i uśredniamy ich predykcje.
  - drzewa nie są przycinane - każde z nich ma małe obciążenie, a dużą wariancję, którą zmniejszamy uśredniając po drzewach
- Gdy zmienna objaśniana przyjmuje wartości jakościowe (klasyfikacja), to zamiast brać średnią stosujemy głosowanie większościowe (klasa najczęściej zgłaszana wygrywa).

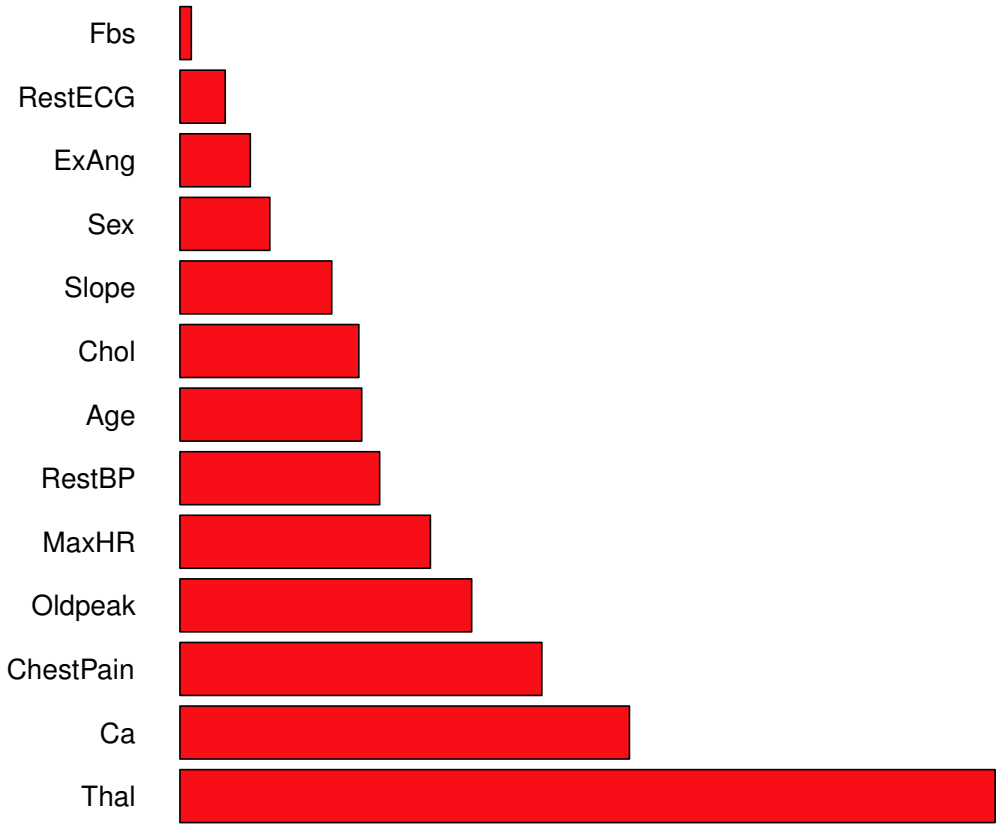
Estymacja błędu testowego dla metody bagging – estymacja out-of-bag (OOB)

- Można pokazać, że średnio około 2/3 obserwacji w procesie boostrapowania jest użyte do konstrukcji drzewa. Obserwacje nie użyte w budowie drzewa nazywane są obserwacjami **out-of-bag** (OOB).
- Zatem jeśli wykonujemy bootstrap B razy to dla każdej obserwacji średnio $B/3$ drzew nie wykorzystywało tej obserwacji. Możemy te drzewa wykorzystać do estymowania błędu predykcji przez wzięcie średniego błędu dla tych drzew.
- Łączny błąd wyestymowany przez OOB (jako średnia błędów po wszystkich obserwacjach) jest dobrym przybliżeniem błędu testowego.

Ustalenie rankingu predyktorów

Dla każdego predyktora wyznaczamy jego "**wagę**" (istotność) przez obliczenie spadku wartości (RSS dla drzew regresyjnych i indeks Giniego dla drzew klasyfikujących), uśrednionego po B drzewach.

Ilustracja: dane "`Heart`"i średni indeks Giniego.



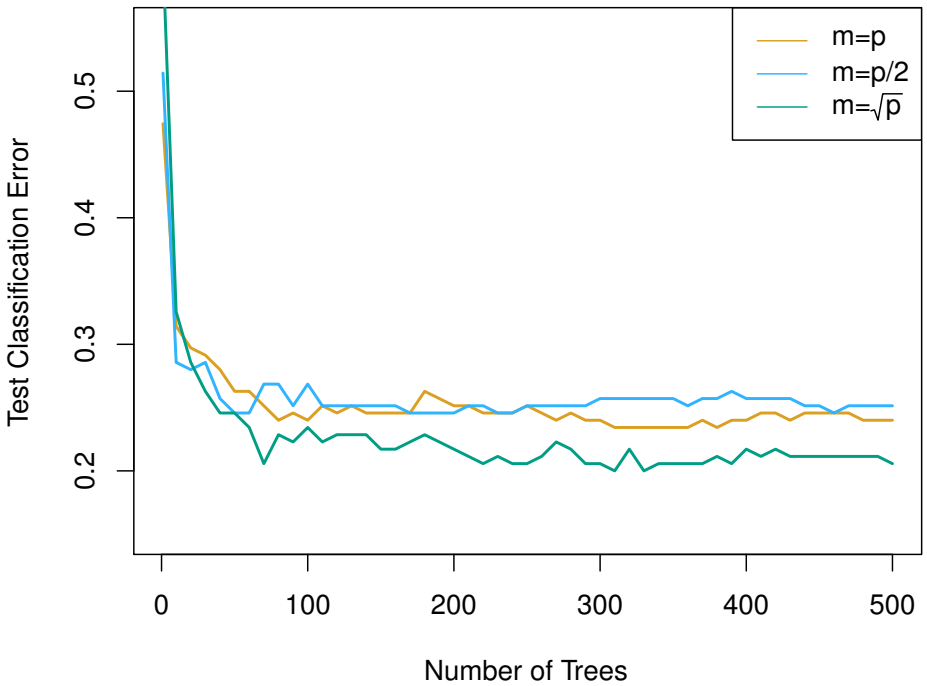
Metody poprawiania jakości predykcji drzewowych: lasy losowe (random forests)

Lasy losowe jako dalsze ulepszenie metody bagging

- Podobnie jak w metodzie bagging dla drzew - budujemy B drzew na bootstrapowanych danych.
- Ale w trakcie budowy drzewa, przy rozważaniu dla którego predyktora zastosować podział, bierzemy pod uwagę tylko m predyktorów wylosowanych spośród wszystkich p predyktorów.
- Tutaj m jest parametrem. Często przyjmuje się $m \approx \sqrt{p}$. Dla $m = p$ metoda sprowadza się do bagging.
- Dzięki temu uwalniamy się od wpływu najsilniejszych predyktorów (czyli takich, które są wybierane na początku do konstrukcji podziału) - mogą one nie wpaść do zbioru m rozważanych.
- Najsilniejsze predyktory często są użyte blisko korzeni drzew, stąd konstruowane drzewa metodą bagging moga być ze sobą silnie skorelowane.
- Restrykcja do m predyktorów może pomóc w redukcji błędu testowego.

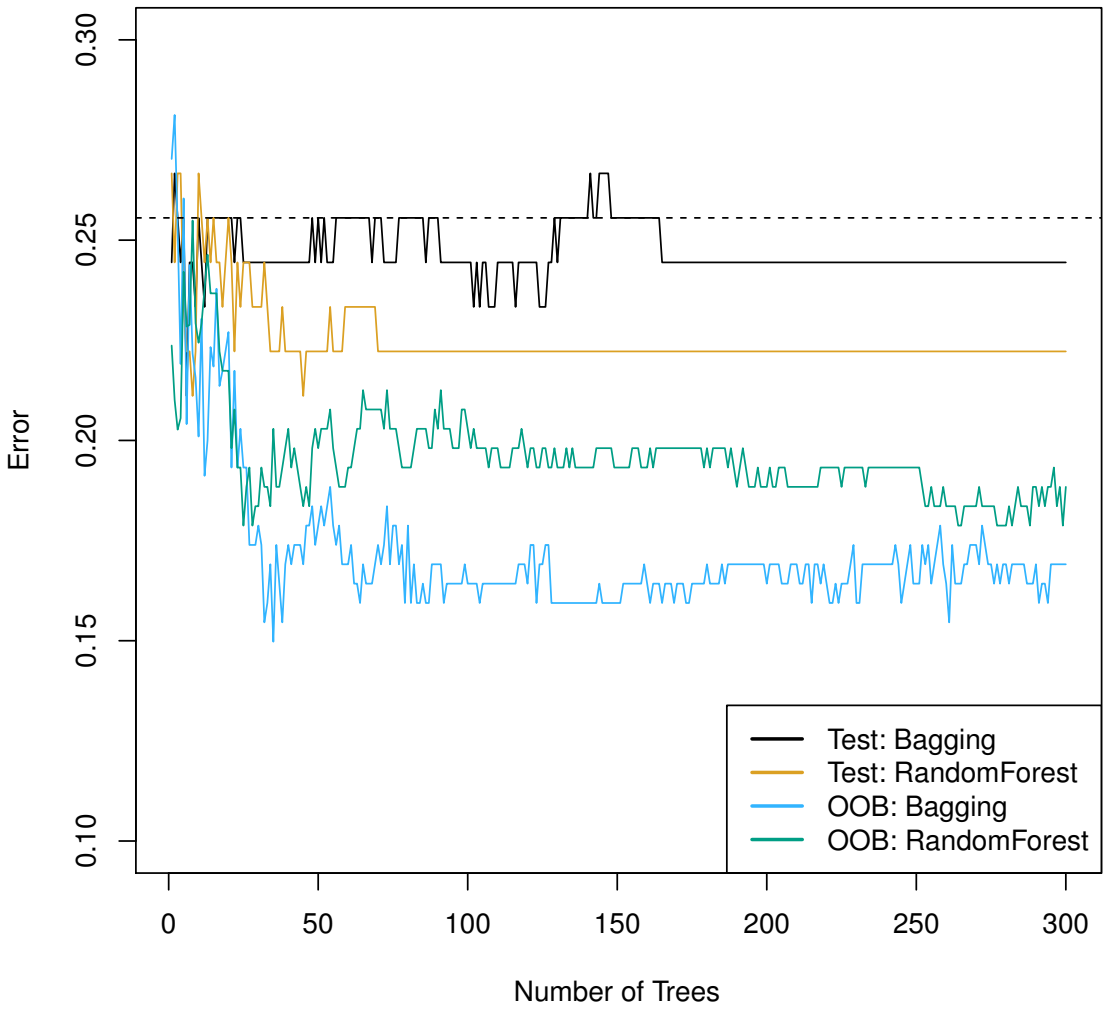
Lasy losowe w predykcji typu choroby nowotworowej

- Trzy różne strategie wyboru parametru m.
- Dane z ekspresji 4718 genów pochodzące od 349 pacjentów.
- Pacjenci są podzieleni na 15 klas (zdrowy oraz 14 typów raka).
- Obserwacje losowo podzielono na treningowe i testowe.
- Cel: użyć drzew losowych na podstawie 500 genów o największej zmienności ekspresji w zbiorze treningowym do predykcji typu raka.



- Pojedyncze drzewo ma błąd 45.7%.
- Model zerowy (przypisujący do dominującej klasy - tutaj 'normal') ma błąd 75.4%.
- las losowy z $m = p$ to bagging

Lasy losowe lepsze niż bagging na danych 'Heart'



- Przerywana linia: błąd testowy pochodzący od jednego drzewa.
- Las losowy dla $m = \sqrt{p}$.

Metody poprawiania jakości predykcji drzewowych: boosting

Boosting

- Ogólne podejście (podobnie jak bagging), można boostować różne metody. Tutaj skupimy się na drzewach.
- Drzewa są konstruowane sekwencyjnie, jedno po drugim.
- Tak otrzymane drzewa są agregowane aby dać uśrednioną predykcję
- Nie korzystamy z bootstrap, każde drzewo budujemy na nieco zmodyfikowanym zbiorze danych
- Uczenie powolne: zamiast jednego dużego modelu zbudowanego naraz, z dużym ryzykiem przeuczenia, postępujemy w kolejnych B krokach, w każdym go nieco go tylko "poduczając"
  - Mając model z danego kroku, w kolejnym dopasowujemy drzewo do *reszt* tego modelu (reszty są zmienną objaśnianą)
  - Dodajemy otrzymane drzewo do modelu i uaktualniamy reszty.

Trzy parametry dla metody boosting

- Liczba drzew: B.
  - Jeśli liczba B będzie za duża, możemy przeuczyć (w odróżnieniu od bagging i random forest, gdzie liczba prób z bootstrap nie ma wpływu na przeuczenie)
- Parametr ściągania (*shrinkage*): $\lambda > 0$.
  - Kontroluje współczynnik "poduczania".
  - Typowe wartości są rzędu 0.01 albo 0.001, zależą od danych.
  - Bardzo małe λ może wymagać bardzo dużego B aby uzyskać dobry model.
- Liczba podziałów w każdym drzewie (czyli wierzchołków wewnętrznych): d.
  - Dla $d = 1$ mamy tylko jeden podział w drzewie (drzewo zamienia się w kikut, ang. *stump*).
  - d określa głębokość interakcji, czyli ile zmiennych jest zaangażowanych w model (drzewo z d podziałami opiera się o wartości co najwyżej d zmiennych)

Algorytm boostingu dla drzew regresyjnych

- Początkowe wartości: $\hat{f}(x) = 0$, reszty $r_i = y_i$ dla $i = 1, \ldots, n$.
- Dla $b = 1, 2, \ldots, B$ powtarzaj:
  (a) Dopasuj drzewo $\hat{f}^b$ o d węzłach wewnętrznych (czyli o $d + 1$ liściach) do danych treningowych (X, r).
  (b) Uaktualnij $\hat{f}$ przez dodanie skurczonej wersji nowego drzewa:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$

  (c) Uaktualnij reszty

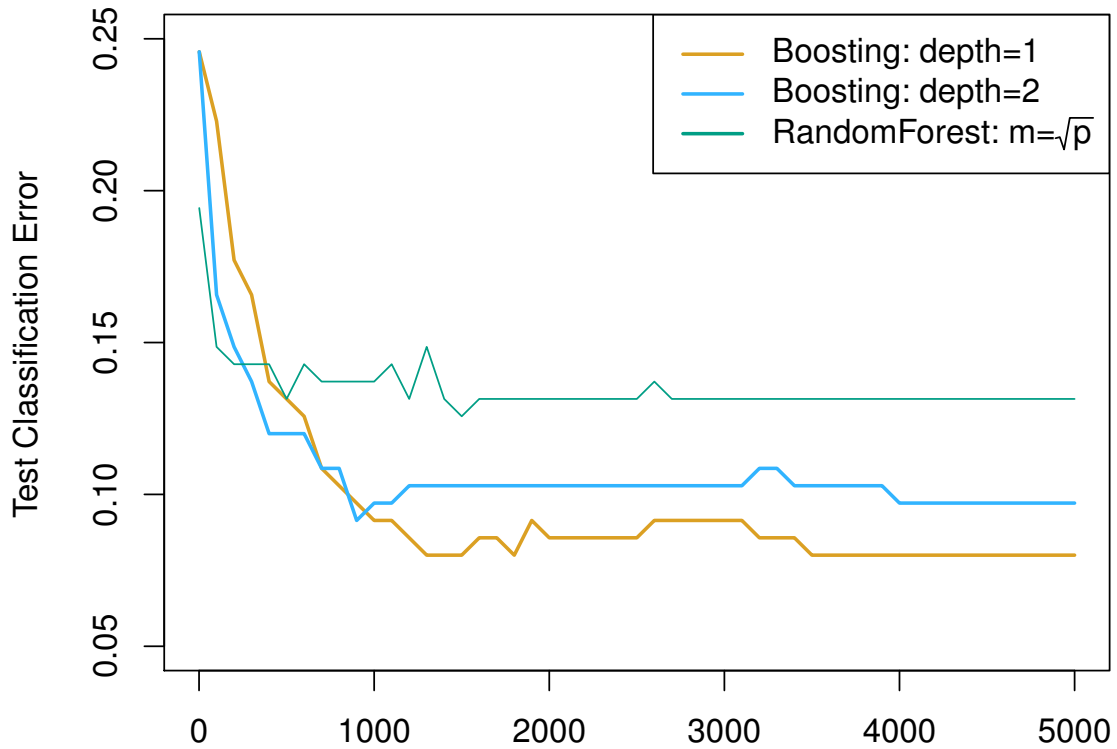
$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

- Wynikiem boostingu jest model

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Zastosowanie boostingu do predykcji raka

- Cel: klasyfikator cancer (którykolwiek z 14 typów) vs normal
- $\lambda = 0.01$
- Boosting wygrywa, ale różnice między metodami nie są statystycznie istotne
- Wszystkie biją pojedyncze drzewo regresji na głowę (błąd 24%)



Kilka metod drzewiastych

- Drzewa decyzyjne
- Bagging
- Random forest
- Boosting