WYKŁAD IV: DRZEWA KLASYFIKACYJNE I REGRESYJNE. Metoda CART

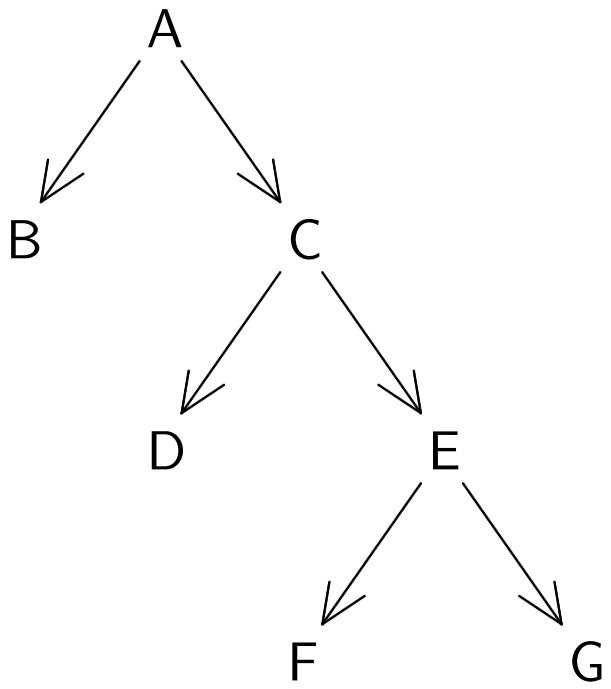
MiNI PW, semestr letni 2013/2014

Drzewa służą do konstrukcji klasyfikatorów prognozujących $Y \in \{1, 2, \ldots, g\}$ na podstawie p-wymiarowego wektora atrybutów (dowolne atrybuty: nominalne, nominalne na skali porządkowej, ciągłe) lub do konstrukcji estymatorów funkcji regresji

$$Y = f(\mathbf{X}) + \varepsilon$$

gdzie ε błąd losowy, taki, że $E\varepsilon = 0$.
Omówimy metodologię wprowadzoną przez Breimana i in. (1984): Classification and Regression Trees (CART). Inne podejście Quinlan (1993,2004) C4.5, C.5 (www.rulequest.com)

Drzewo – graf skierowany, acykliczny, spójny, wierzchołek wyróżniony - korzeń drzewa.
Drzewa binarne – z każdego wierzchołka wychodzą 2 krawędzie (lub 0)(dla liści)



B, D, F, G – liście

D i E są dziećmi węzła C, F i G jego potomkami
Konwencja: drzewa rosną od góry do dołu – korzeń na górze rysunku, liście na dole

W każdym węźle warunek logiczny $\{X_i \leqslant c\}$ (lub $\{X_i < c\}$, $\{X_i > c\}$)
– spełniony: ścieżka lewa
– niespełniony: scieżka prawa

Zmienna X_i jest jedną ze zmiennych objaśniających i z reguły zmienia się przy przejściu z węzła do węzła.

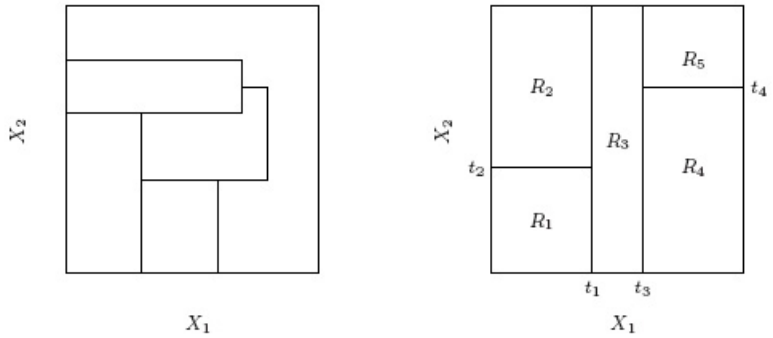
W rezultacie z każdym liściem związana jest hiperkostka określona przez warunki na drodze łączącej liść z korzeniem, hiperkostki tworzą rozbicie $\mathcal{X} = R^p$.

Dla drzewa klasyfikacyjnego:

decyzja związana z liściem: wybierz tę klasę, do której należy większość elementów próby uczącej, które trafiły do danego liścia po przepuszczeniu przez drzewo;

dla drzewa regresyjnego:

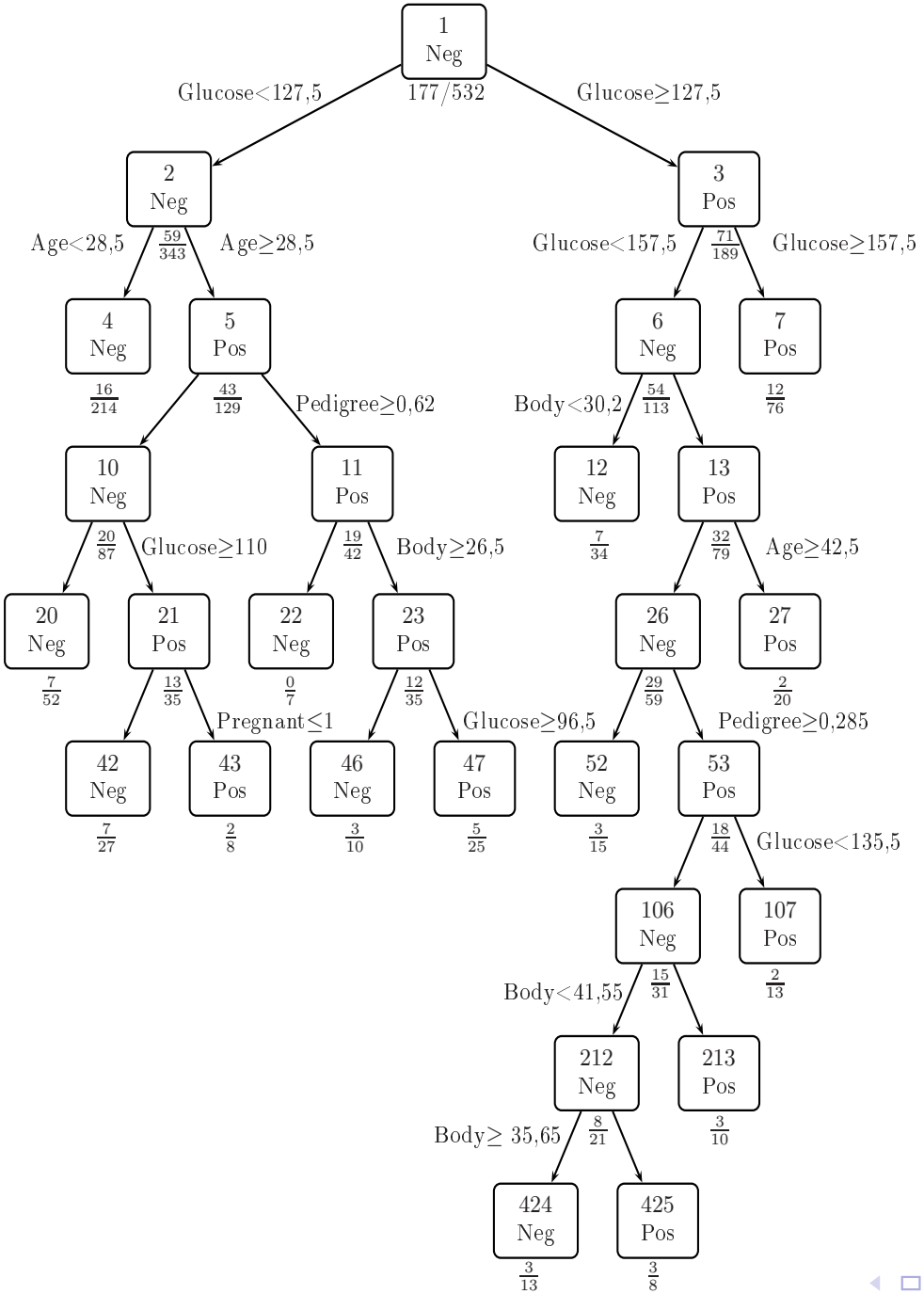
$x \in R_m$ – kostka w przestrzeni $\mathcal{X}$ wyznaczona przez liść
$\hat{E}(Y|X = x) =$ średnia z wartości y dla elementów znajdujących się w liściu



Przykład. Występowanie cukrzycy wśród Indianek Pima (Arizona, USA).
$y \longrightarrow$ pos (przypadki dodatnie)
$y \longrightarrow$ neg (przypadki ujemne)
X : liczba ciąż, wynik testu glukozowego ($\in (56, 199)$), ciśnienie rozkurczowe, grubość fałdy skórnej na tricepsie (mm), indeks masy ciała BMI (ciężar / (wzrost w m)^2), współczynnik podatności na cukrzycę ($\in (0.085, 2.42)$), wiek

$n = 532$, z tego 33% – cukrzyca

Liść nr 4: test glukozowy < 127.5, wiek < 28.5,
osoby w tym liściu zaklasyfikowane jako zdrowe, 214 elementów, w tym 16 błędnie sklasyfikowanych.


1
Neg
177/532
Glucose<127,5
Glucose≥127,5
2
Neg
3
Pos
Age<28,5
59/343
Age≥28,5
Glucose<157,5
71/189
Glucose≥157,5
4
Neg
16/214
5
Pos
43/129
Pedigree≥0,62
6
Neg
54/113
7
Pos
12/76
Body<30,2
10
Neg
20/87
Glucose≥110
11
Pos
19/42
Body≥26,5
12
Neg
7/34
13
Pos
32/79
Age≥42,5
20
Neg
7/52
21
Pos
13/35
22
Neg
0/7
23
Pos
12/35
26
Neg
29/59
27
Pos
2/20
Pregnant≤1
Glucose≥96,5
Pedigree≥0,285
42
Neg
7/27
43
Pos
2/8
46
Neg
3/10
47
Pos
5/25
52
Neg
3/15
53
Pos
18/44
Glucose<135,5
106
Neg
15/31
107
Pos
2/13
Body<41,55
212
Neg
8/21
213
Pos
3/10
Body≥ 35,65
424
Neg
3/13
425
Pos
3/8

Reguły podziału - funkcje różnorodności

Reguły podziału w węzłach drzewa klasyfikacyjnego.
Podpróba znajdująca się w węźle charakteryzuje się pewną różnorodnością klas.
Dążymy do tego, żeby różnorodność klas dla dzieci węzła była jak najmniejsza.

węzeł: 80 – klasa 1, 20 – klasa 2

idealny podział
(zmniejszył różnorodność klas w dzieciach do 0)

potomek lewy: 80 (klasa 1) potomek prawy: 20 (klasa 2)

Potrzebujemy:

– miary różnorodności klas w węźle;
– oceny zmiany różnorodności klas po przejściu o poziom wyżej;
– algorytmu maksymalizacji zmiany różnorodności.

$$(x_i, y_i), i = 1, 2, \ldots, n \quad - \quad \text{próba ucząca}$$

węzeł m wyznaczony przez warunek $x \in R_m \subset \mathcal{X}$

frakcja elementów z klasy k w węźle m

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i = k) = \frac{n_{mk}}{n_m}$$

n_m – liczba obserwacji w węźle m
n_{mk} – liczba obserwacji z klasy k w węźle m

Rozsądna miara różnorodności klas powinna być
$= 0$, gdy elementy tylko z jednej klasy
$=$ max, gdy $\hat{p}_{m1} = \hat{p}_{m2} = \cdots = \hat{p}_{mg} = 1/g$

$$k(m) = \arg\max_k \hat{p}_{mk} \tag{1}$$

Miary różnorodności klas w węźle m drzewa T

$$Q_m(T) = \begin{cases} 1 - \hat{p}_{mk(m)} \\ \sum_{k=1}^{g} \hat{p}_{mk}(1 - \hat{p}_{mk}) & \text{indeks Giniego} \\ -\sum_{k=1}^{g} \hat{p}_{mk} \log \hat{p}_{mk} & \text{entropia} \end{cases} \tag{2}$$

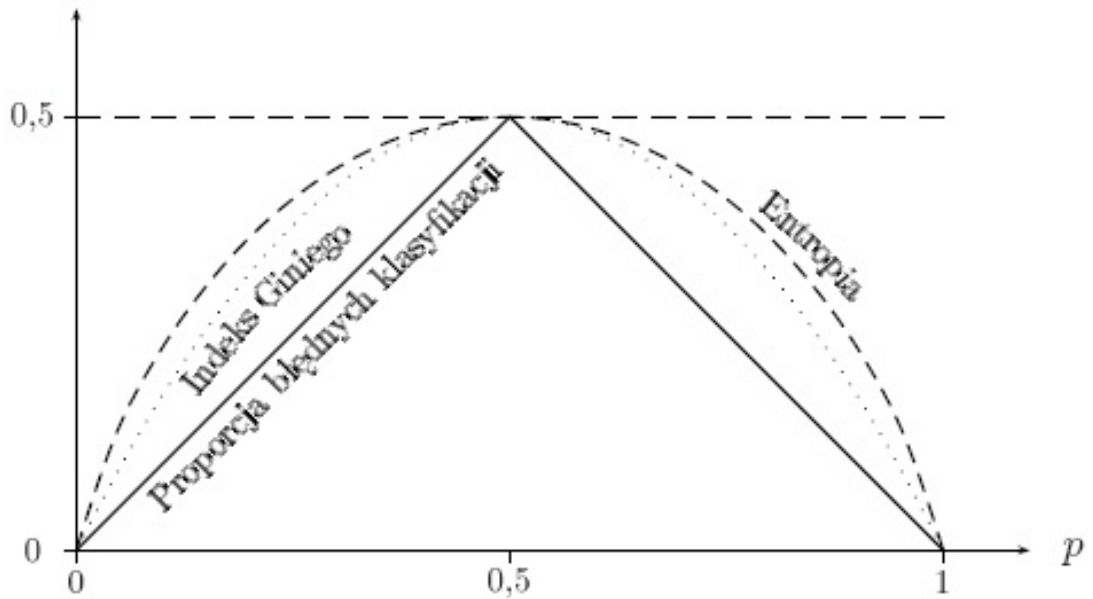
$\mathbf{p} = (p_1, p_2 \ldots, p_k)$. Z, Z' dwie niezależne zmienne losowe przyjmujące wartość i z prawdopodobieństwem p_i. Indeks Giniego dla $\mathbf{p}$ $= \sum_{i=1}^{k} p_i(1 - p_i) = P(Z \neq Z')$.

Interpretacja indeksu Giniego w drzewie klasyfikacyjnym: oszacowanie pr. błędnej decyzji, gdy obserwacje klasyfikowane są do klasy k z pr. $\hat{p}_{mk}$.

W przypadku dwóch klas, $g = 2$, podane trzy miary przyjmują postać:

$$Q_m(T) = \begin{cases} 1 - \max(p, 1-p) \\ 2p(1-p) \\ -p \log p - (1-p) \log (1-p), \end{cases} \tag{3}$$

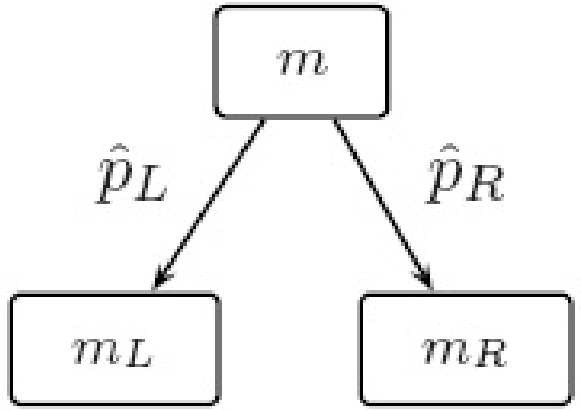
p: jest ułamkiem przynależności do klasy 2. (Na rysunku entropia przemnożona przez 0.5)



Oznaczmy dzieci węzła-rodzica m symbolami m_L i m_R.

$$\hat{p}_L = \frac{n_{m_L}}{n_m} \qquad \hat{p}_R = \frac{n_{m_R}}{n_m} = 1 - \hat{p}_L$$

$\hat{p}_L$ ($\hat{p}_R$) jest ułamkiem elementów próby uczącej, które z węzła m przeszły do m_L (m_R), a n_{m_L} (n_{m_R}) oznacza liczbę obserwacji w m_L (m_R).



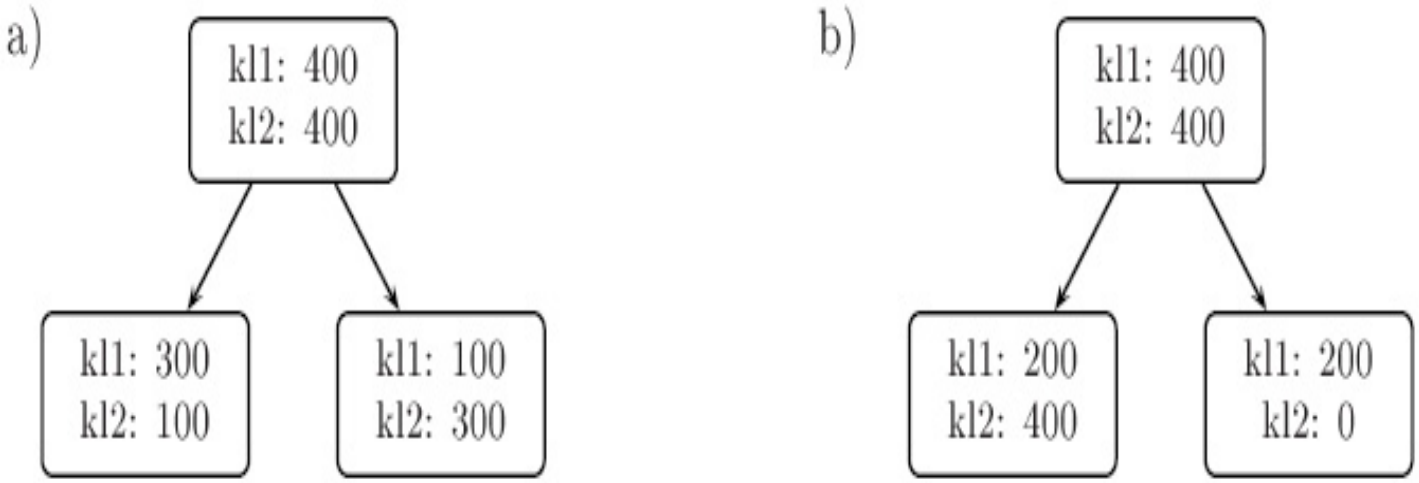
Węzeł-rodzic m i jego dzieci m_L i m_R

Łączną miara różnorodności klas w dzieciach węzła m

$$Q_{m_L,m_R}(T) = \hat{p}_L Q_{mL}(T) + \hat{p}_R Q_{mR}(T),$$

uśredniona miara różnorodności w dzieciach. Uśrednienie uwzględnia frakcje obserwacji w lewym i prawym potomku.
Zmiana różnorodności klas przy przejściu od rodzica do dzieci

$$\Delta Q_{m_L,m_R}(T) = Q_m(T) - Q_{m_L,m_R}(T).$$



Uwaga: Indeks Giniego i entropia **bardziej czułe** na zmiany rozkładów klas niż proporcja błędnych klasyfikacji.

Rysunek: dla frakcji błędnych klasyfikacji $\Delta Q_{m_L,m_R}(T) =$ w obu przypadkach taka sama, gdy drugi daje (intuicyjnie!) większe zmniejszenie różnorodności klas.
Cel: maksymalizacja $\Delta Q_{m_L,m_R}(T)$ ze względu na zmienną objaśniającą i próg c.

Atrybuty nominalne

Dla atrybutu nominalnego przyjmującego L wartości: gdyby przyjąć podział na podstawie dowolnego podzbioru wartości, to mielibyśmy

$$\frac{1}{2}2^L - 1 = 2^{L-1} - 1 \quad \text{podziałów}$$

Duży koszt obliczeniowy. Ograniczamy się do podziałów:

$$\{x_i \leqslant c_k\}, \quad \text{zakładamy, że zmienna } x_i \text{ na skali porządkowej.}$$

Dla nominalnej cechy x_i przyjmującej L wartości i $g = 2$ porządkujemy jej wartości $x_i^{(k)}$ według $p(1|x_i^{(k)})$ i traktujemy ją jako cechę na skali porządkowej tzn $x_i^{(k)} \prec x_i^{(l)}$ jeśli $p(1|x_i^{(k)}) < p(1|x_i^{(l)})$.
Twierdzenie (por. tw. 4.1 w KC (2005)). W przypadku miary różnorodności Giniego i entropii powyższa procedura prowadzi do wyboru optymalnego podziału spośród wszystkich $2^{L-1} - 1$ podziałów.

Drzewa regresyjne

Szukamy podziału m na m_L i m_R, aby

$$SSE(m_L) + SSE(m_R) \quad (\star)$$

było minimalne.

$SSE(m_L)$ – suma kwadratów rezyduów, gdy regresja dla m_L estymowana jest przez średnią próbkową wartości zmiennej objaśnianej dla tego węzła itd.

Minimalizacja $(\star)$ równoważna maksymalizacji różnicy zmiany SSE przy przejściu od rodzica do dzieci.

Strategia wyboru najlepszego drzewa

- Utwórz pełne drzewo T_0 zatrzymując podziały kiedy pewna minimalna wielkość węzła została osiągnięta;
- przy różnych parametrach określających koszt złożoności drzew przytnij drzewo T_0 do mniejszego drzewa;
- spośród tak utworzonej skończonej rodziny drzew wybierz drzewo dające najmniejszy błąd w oparciu o kroswalidację.

Reguły przycinania drzew

Kontynuując metodę optymalnych podziałów dojdziemy do drzewa z (najczęsciej) jednoelementowymi liśćmi (przeuczenie - przetrenowanie drzewa)

$R(T)$ – miara niedoskonałości drzewa
– dla drzewa klasyfikacyjnego: frakcja błędnych klasyfikacji
– dla drzewa regresyjnego: $\sum_{\text{liście}} SSE$

Wprowadzamy karę za złożoność drzewa

$$R_\alpha(T) = R(T) + \alpha|T| \qquad (*)$$

$|T|$ - liczba liści w drzewie T, $\alpha > 0$.

Przy ustalonym α minimalizujemy $R_\alpha(T)$.

Dla $\alpha = 0$: drzewo pełne T_0,
duże α: sam korzeń.

Zwiększając α od 0 dostaniemy dyskretną rodzinę poddrzew T_j drzewa T_0 takich że

$$T_j \text{ minimalizuje } (*) \text{ dla } \alpha \in [\alpha_j, \alpha_{j+1}),\ j = 1, 2, \ldots, k.$$

Wybieramy „dobre" drzewo spośród drzew $T_1, T_2, \ldots, T_k$ (kandydatów na dobre drzewa).

Metodą kroswalidacji liczymy $R^{CV}(T_i)$. Później wyznaczamy j_0:

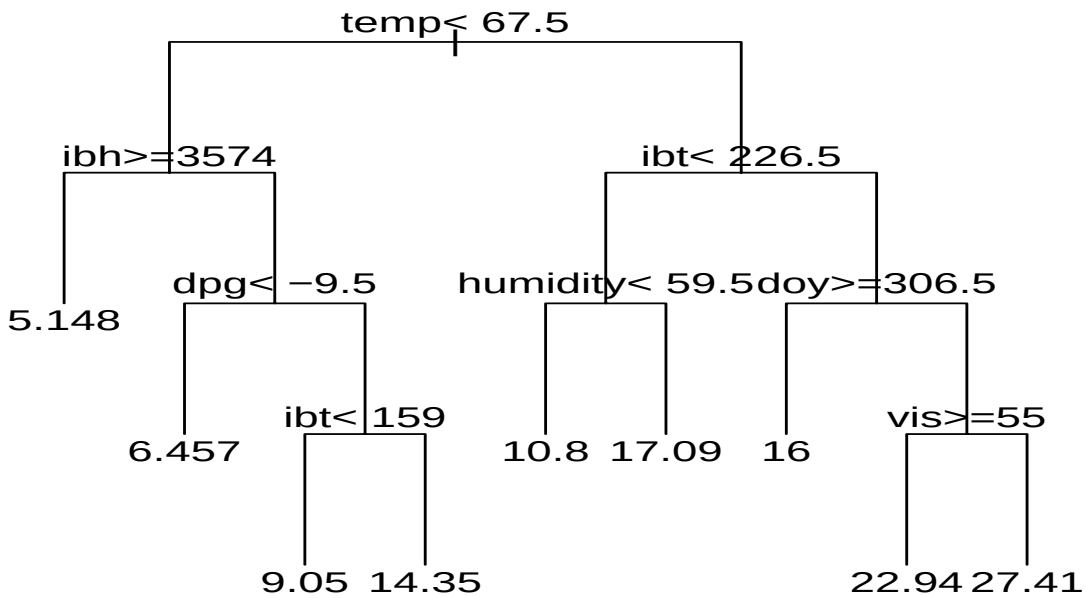
$$R^{CV}(T_{j_0}) = \min_j R^{CV}(T_j) \quad (!)$$

lub **reguła 1SE** Wybieramy najmniejsze drzewo T_j dla którego

$$R^{CV}(T_j) \leqslant R^{CV}(T_{j_0}) + SE(R^{CV}(T_{j_0})),$$

gdzie $SE(R^{CV}(T_{j_0})) = (R^{CV}(T_{j_0})(1 - R^{CV}(T_{j_0}))/V)^{1/2}$, błąd standardowy dla kroswalidacji V-krotnej. Reguła 1SE uwzględnia wypłaszczanie się funkcji $R^{CV}(T_j)$ w okolicach minimum.

Przykład. Zależność między stężeniem ozonu (O3), a warunkami meteo:
sbtp – temperatura
hmdt – wilgotność powiewtrza
vsty – visibility
ibtp – inversion base temperature
dgpg – gradient ciśnienia, ibht – inversion base height

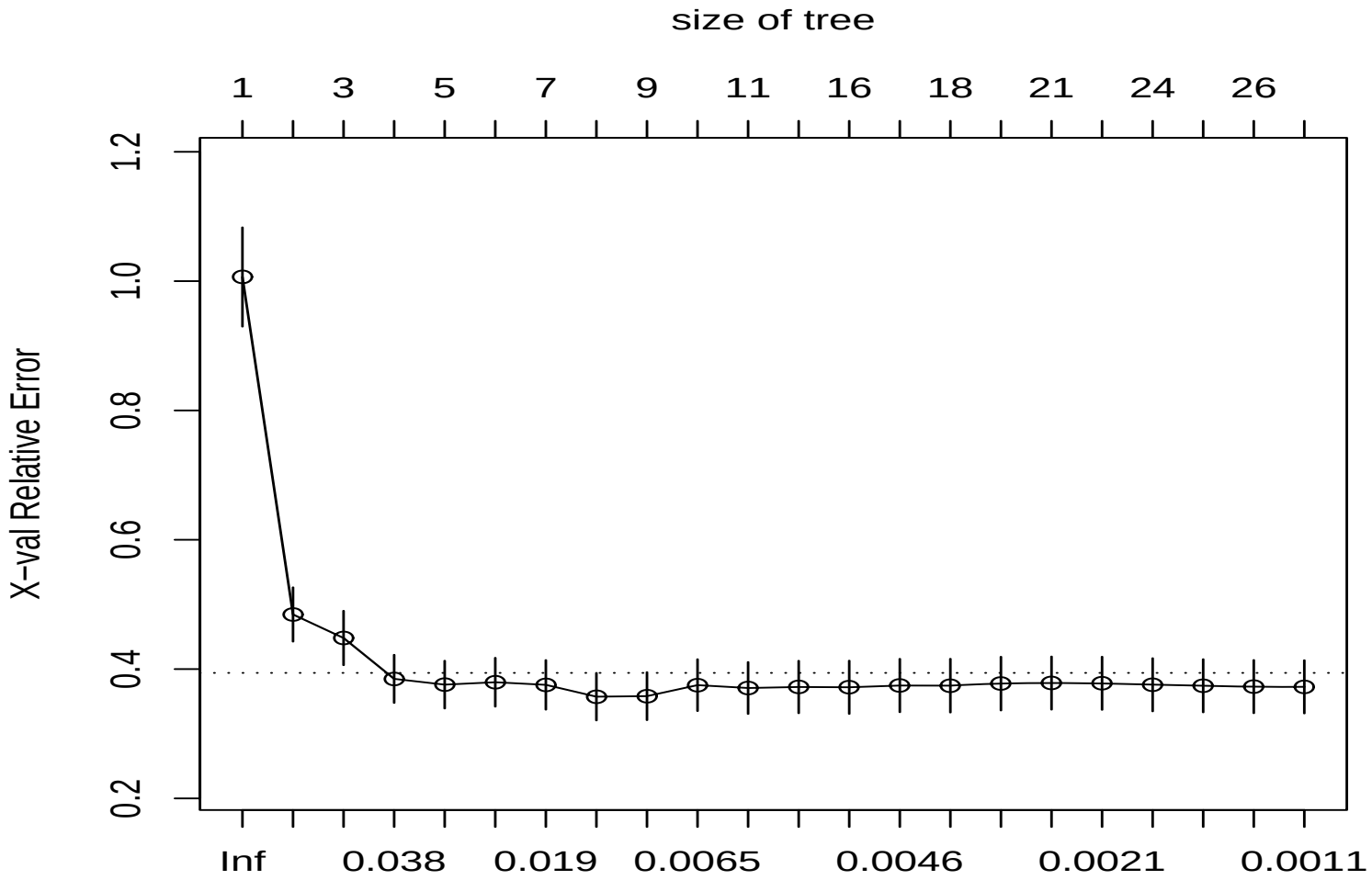


```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 330 21115.4100 11.775760
   2) temp< 67.5 214  4114.3040  7.425234
     4) ibh>=3573.5 108   689.6296  5.148148 *
     5) ibh< 3573.5 106  2294.1230  9.745283
       ................
   3) temp>=67.5 116  5478.4400 19.801720
     6) ibt< 226.5 55  1276.8360 15.945450

       30) vis>=55 36  1149.8890 22.944440 *
       31) vis< 55 17   380.1176 27.411760 *

          CP nsplit rel error  xerror     xstd
1  0.5456993      0   1.00000 1.00902 0.076836
2  0.0736591      1   0.45430 0.48649 0.041507
3  0.0535415      2   0.38064 0.42977 0.038669
4  0.0267557      3   0.32710 0.38608 0.035752
5  0.0232760      4   0.30034 0.37523 0.036056
6  0.0231021      5   0.27707 0.36050 0.035629
7  0.0153249      6   0.25397 0.35120 0.035700
8  0.0109137      7   0.23864 0.34461 0.034955
9  0.0070746      8   0.22773 0.35944 0.038598
.................................
22 0.0010000     26   0.16016 0.34697 0.037049
```



Drzewo dające minimalny xerror =0.3446 (stosunek $R_{CV}(T)$ i SSE dla korzenia) oparte na siedmiu podziałach. Odpowiadający SE=0.035. Drzewo wybrane metodą 1SE ma xerror=0.3752 ($<$ 0.3446+0.035) i jest oparte na 4 podziałach. cp (complexity) odpowiada α.

Konstrukcja drzew regresyjnych i klasyfikacyjnych w R: pakiet `rpart`, funkcja rpart.

```
data.rpart<-rpart(03 ~., cp=0.001, minsplit=2,data=..)
```

cp odpowiada wartości α (współczynnik w karze za złożoność drzewa), minsplit -minimalna liczba elementów w węźle, przy której dokonuje się jeszcze podziału elementów węzła. Wykres przedstawiający drzewo:

```
plot(data.rpart, uniform=TRUE,margin=0.1)
text(data.rpart)
```

Wykres i wydruk xerror i jego błędu standardowego:

```
plotcp(data.rpart)
printcp(data.rpart)
```

Uwagi

- Wartości odstające: przy konstrukcji podziału węzła rozpatrujemy tylko zmienne nie mające braków w zbiorze uczacym i poza najlepszym podziałem wyznaczamy tzw. podziały zastepcze (*surrogate splits*) tzn. podział drugi, trzeci w kolejności itd. 'Przepuszczając' obserwacje przez drzewo znajdujemy w każdym węźle najlepszy realizowalny dla tej obserwacji podział.
- Niestabilność drzew i duża wariancja rozwiązania: nieduże zmiany w danych mogą spowodować istotne zmiany w strukturze podziałów (związane z hierarchiczną strukturą drzewa: zmiana w podziale na górze propaguje się w dół). Również wartość optymalnego C_p i optymalne drzewo może zmieniać się od wykonania do wykonania. Komitety drzew (bagging) zmniejszają wariancję.
- Drzewa regresyjne nie dają ciągłego estymatora regresji;
- Drzewa regresyjne nieprzydatne w przypadku zależności addytywnych od predyktorów $E(Y|X) = f(X_1) + \ldots + f(X_p)$.