Zaawansowane uczenie maszynowe:
wykład 5

Paweł Cichosz

Przycinanie drzew decyzyjnych

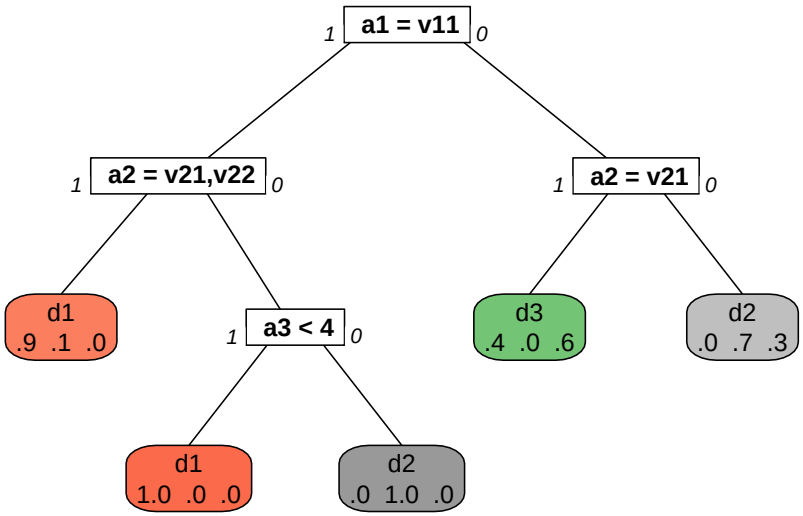
Cel: Zapobieganie nadmiernemu dopasowaniu (bardziej wymagające, ale potencjalnie skuteczniejsze niż kryterium stopu).

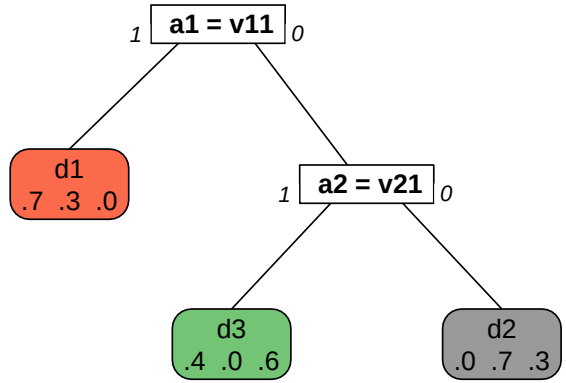
Operator: zastąpienie poddrzewa liściem.

Kryterium: oczekiwana redukcja błędu rzeczywistego (wiele różnych szczegółowych wariantów).

Strategia: najczęściej wstępująca (czasem zstępująca, najpierw najlepszy).

Przycinanie drzew decyzyjnych





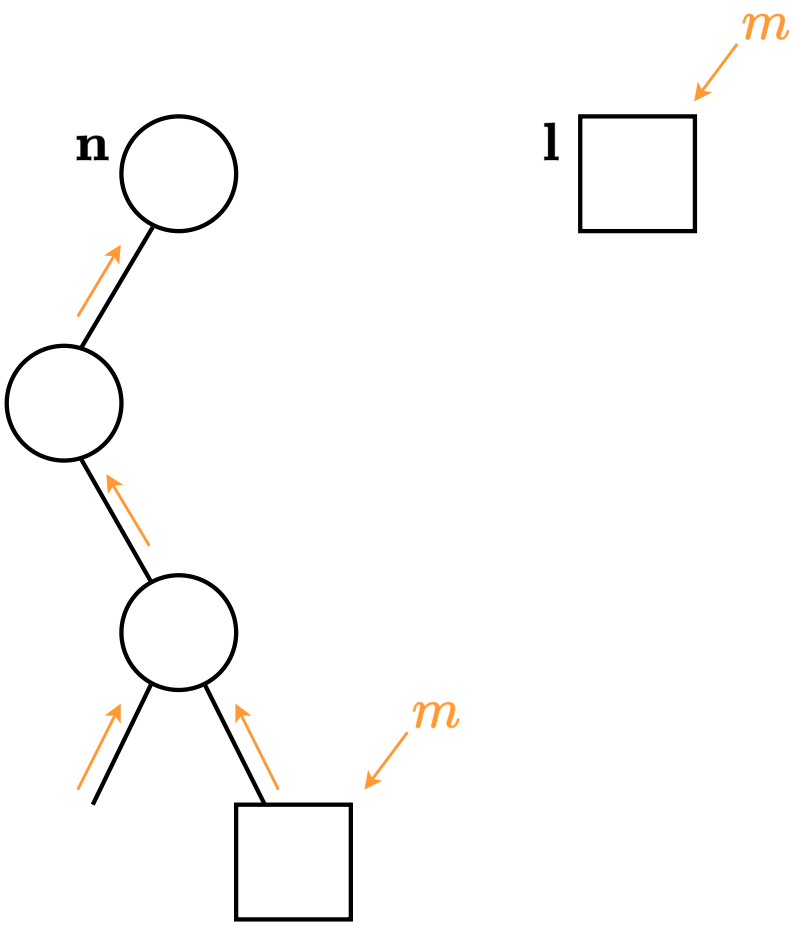
REP (Reduced Error Pruning)

- Kryterium przycinania oparte na estymacji błędu rzeczywistego z wykorzystaniem osobnego podzbioru przykładów R nieużywanego w trakcie budowy drzewa.
- Węzeł $\mathbf{n}$ zastępowany liściem $\mathbf{l}$ jeśli:

$$e_{R_{\mathbf{n}}}(\mathbf{l}) \leq e_{R_{\mathbf{n}}}(\mathbf{n})$$

- Strategia wstępująca.
- Dobre jeśli możemy poświęcić osobną pulę przykładów.
- Może czasem lepiej przycinać trochę gorzej, ale budować lepiej?

MEP (Minimum Error Pruning)



- Kryterium przycinania oparte na m-estymacji (wygładzaniu Cestnika) dokładności na zbiorze trenującym.
- Wygładzanie na poziomie liści – większy wpływ na liście niżej położone w drzewie o mniejszej liczbie przykładów).

MEP (Minimum Error Pruning)

- Węzeł $\mathbf{n}$ zastępowany liściem $\mathbf{l}$ jeśli:

$$\hat{e}_{T_\mathbf{n}}(\mathbf{l}) \leq \hat{e}_{T_\mathbf{n}}(\mathbf{n})$$

- Błąd dla dowolnego liścia jest dopełnieniem do 1 jego m-estymowanej dokładności:

$$\hat{e}_{T_\mathbf{l}}(\mathbf{l}) = 1 - \frac{|T_{\mathbf{l},c=d_\mathbf{l}}| + mp}{|T_\mathbf{l}| + m}$$

- Błąd dla dowolnego węzła jest średnią ważoną błędów węzłów lub liści potomnych:

$$\hat{e}_{T_\mathbf{n}}(\mathbf{n}) = \sum_{\mathbf{n}' \in \mathbf{N}(\mathbf{n})} \frac{|T_{\mathbf{n}'}|}{|T_\mathbf{n}|} \hat{e}_{T_{\mathbf{n}'}}(\mathbf{n}')$$

  gdzie $\mathbf{N}(\mathbf{n})$ oznacza zbiór węzłów lub liści potomnych węzła $\mathbf{n}$.
- Zwykle $p = \frac{1}{|C|}$, a m kontroluje agresywność przycinania.
- Strategia wstępująca.

CCP (Cost-Complexity Pruning)

- Kryterium w postaci nierównościowej:

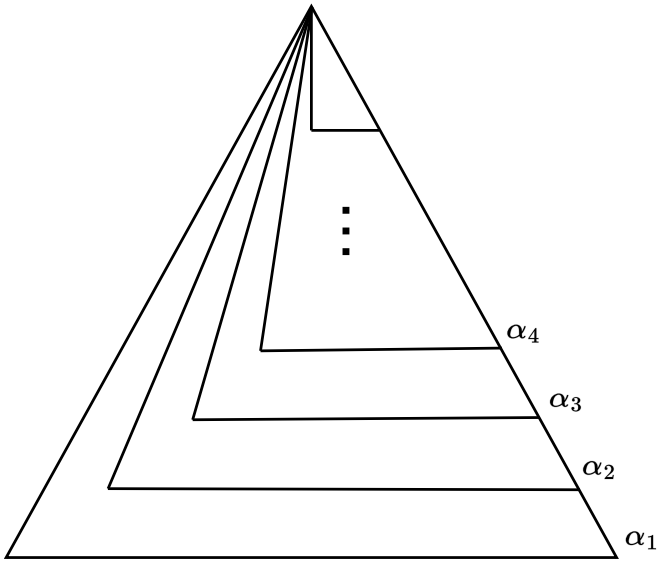
$$e_{T_{\mathbf{n}}}(\mathbf{l}) \leq e_{T_{\mathbf{n}}}(\mathbf{n}) + \alpha \mathcal{C}(\mathbf{n})$$

  gdzie $\mathcal{C}(\mathbf{n})$ oznacza złożoność poddrzewa o korzeniu w węźle $\mathbf{n}$, mierzoną liczbą liści, a α jest współczynnikiem złożoności.
- Przycinanie realizowane faktycznie nie przez stosowanie kryterium nierównościowego, lecz przez wybór drzewa, które minimalizuje:

$$e_T(\mathbf{n}_1) + \alpha \mathcal{C}(\mathbf{n}_1)$$

  gdzie $\mathbf{n}_1$ oznacza węzeł-korzeń drzewa.
- Dla każdego węzła istnieje progowa wartość α, powyżej której zostanie on przycięty, aby uzyskać minimalny koszt.

CCP (Cost-Complexity Pruning)



- Można wyznaczyć ciąg zagnieżdżonych drzew (tzn. takich, że każde kolejne ma jeden lub więcej przyciętych węzłów od poprzedniego), od pełnego pierwotnego drzewa do pojedynczego liścia, odpowiadających przedziałom wartości α.

CCP (Cost-Complexity Pruning)

- Wybór wartości α na podstawie błędu estymowanego za pomocą procedury k-krotnej walidacji krzyżowej, np.:
  - α odpowiadające minimalnemu estymowanemu błędowi,
  - największe α, dla którego odpowiadające estymowany błąd nie przekracza minimalnego estymowanego błędu o więcej niż jedno odchylenie standardowe.
- Można określić kryterium stopu dla budowy drzewa przez podanie wartości α (minimalny niezbędny spadek błędu wymagany do utworzenia węzła).

Obsługa brakujących wartości atrybutów

Techniki ogólnego przeznaczenia: pomijanie przykładów, pomijanie atrybutów, imputacja (średnią, medianą, modą, wartością przewidywaną przez model klasyfikacji/regresji, wielokrotna), rozszerzenie przeciwdziedziny (słownika) dla atrybutów dyskretnych.

Techniki specyficzne dla drzew:

- podziały zastępcze (dla drzew binarnych),
- przykłady ułamkowe.

Podziały zastępcze

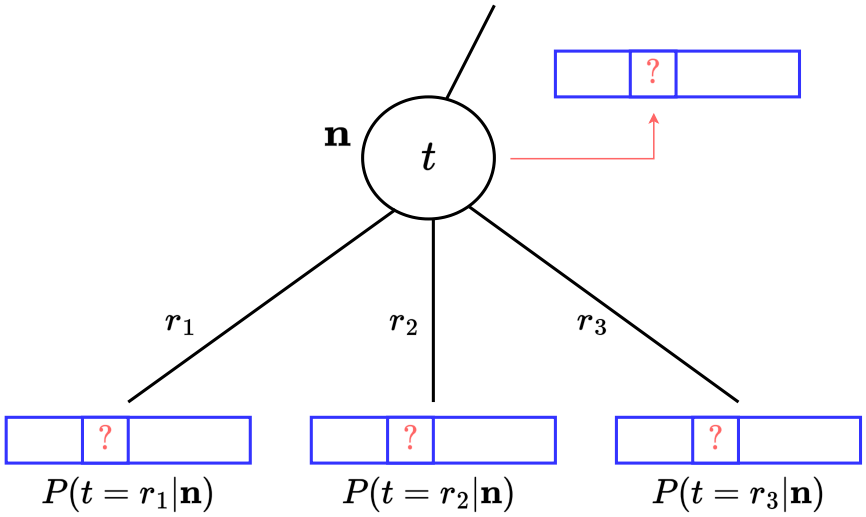
Ocena podziałów: brakujące wartości pomijane przy wyznaczaniu prawdopodobieństw (wartość oceny może być redukowana proporcjonalnie do liczby braków w celu zniechęcenia do używania atrybutów z wieloma wartościami brakującymi).

Stosowanie podziałów: w każdym węźle oprócz podziału podstawowego t jest uporządkowana lista podziałów zastępczych $t_1, t_2, \ldots$:

- każdy podział zastępczy używa innego atrybutu,
- jeśli brakuje wartości atrybutu używanego przez podział podstawowy, stosowany jest pierwszy podział zastępczy, dla którego wartość używanego przez niego atrybutu jest znana.

Wybór podziałów zastępczych: na podstawie zgodności z podziałem podstawowym (oceny, jak dobrze potencjalny podział zastępczy przewiduje wynik podziału podstawowego).

Przykłady ułamkowe



Ocena i stosowanie podziałów:

w przypadku brakującej wartości zakłada się, że możliwy jest każdy wynik podziału, co można interpretować jako rozdzielenie przykładu na przykłady ułamkowe odpowiadające poszczególnym wynikom.

Prawdopodobieństwo wyniku: może być interpretowane jako ułamkowa liczba egzemplarzy przykładu x dla wyniku r, jeśli wartość atrybutu sprawdzanego przez podział t jest nieznana:

$$P(t = r|\mathbf{n}) = \frac{|T_{\mathbf{n},t=r}|}{|T_{\mathbf{n},t\neq?}|}$$

Przykłady ułamkowe

Powtórny podział z brakującą wartością: prawdopodobieństwa (ułamki) się mnożą.

Liczba egzemplarzy/prawdopodobieństwo przykładu w korzeniu: $w_x = 1$ (o ile nie określono innych początkowych wag).

Liczba egzemplarzy/prawdopodobieństwo przykładu po podziale:

$$w_{x,\mathbf{n}_r} = P(t = r|\mathbf{n}) w_{x,\mathbf{n}}$$

gdzie $\mathbf{n}_r$ jest potomkiem węzła $\mathbf{n}$ odpowiadającym wynikowi r stosowanego w węźle $\mathbf{n}$ podziału.

Przykłady ułamkowe

Używanie przykładów ułamkowych: zliczanie przykładów (np. w celu estymacji rozkładu klas) zastąpione sumowaniem liczb egzemplarzy, np.:

$$|T_{\mathbf{n},c=d}| = \sum_{x \in T_{\mathbf{n},c=d}} w_{x,\mathbf{n}}$$

Predykcja: agregacja prawdopodobieństw klas w liściach, do których docierają przykłady ułamkowe:

$$P(\mathbf{l}|x) = w_{x,\mathbf{l}}$$

$$P(d|x) = \sum_{\mathbf{l}} P(\mathbf{l}|x) P(d|\mathbf{l})$$

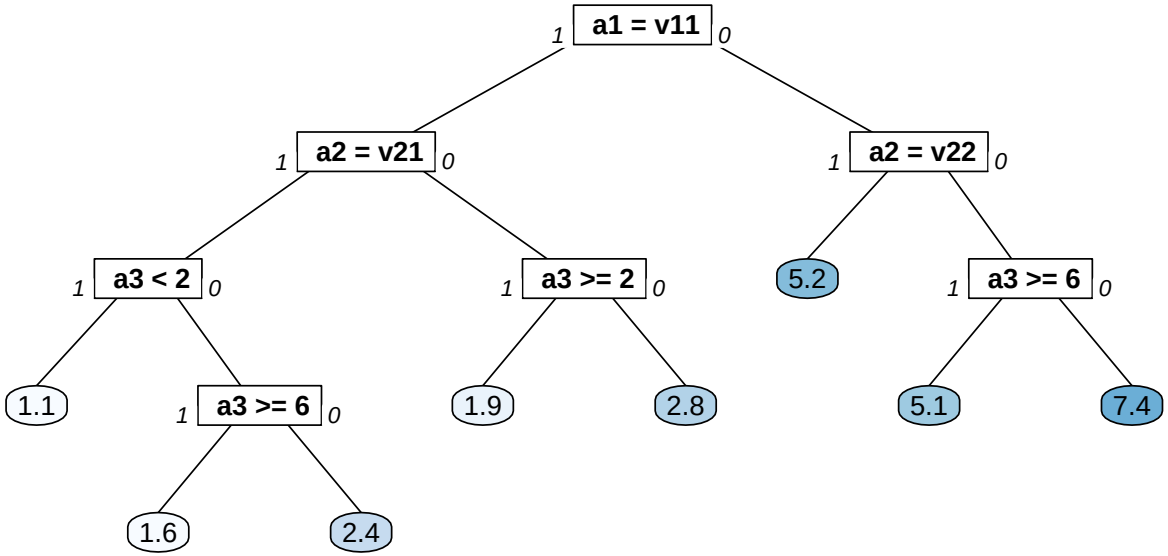
$$h(x) = \arg\max_{d \in C} P(d|x)$$

Drzewa regresji

Węzły, gałęzie: podziały, wyniki podziałów jak w drzewie decyzyjnym.

Liście: liczbowe przewidywane wartości funkcji docelowej.

Proces predykcji: przykład x propagowany od korzenia drzewa do liścia ścieżką wyznaczaną przez wyniki podziałów w kolejnych odwiedzonych węzłach $t_1(x), t_2(x), \ldots$, liczba z osiągniętego liścia jest wartością $h(x)$.



Drzewa regresji

Budowa: ten sam schemat jak dla drzewa decyzyjnego, lecz inne:

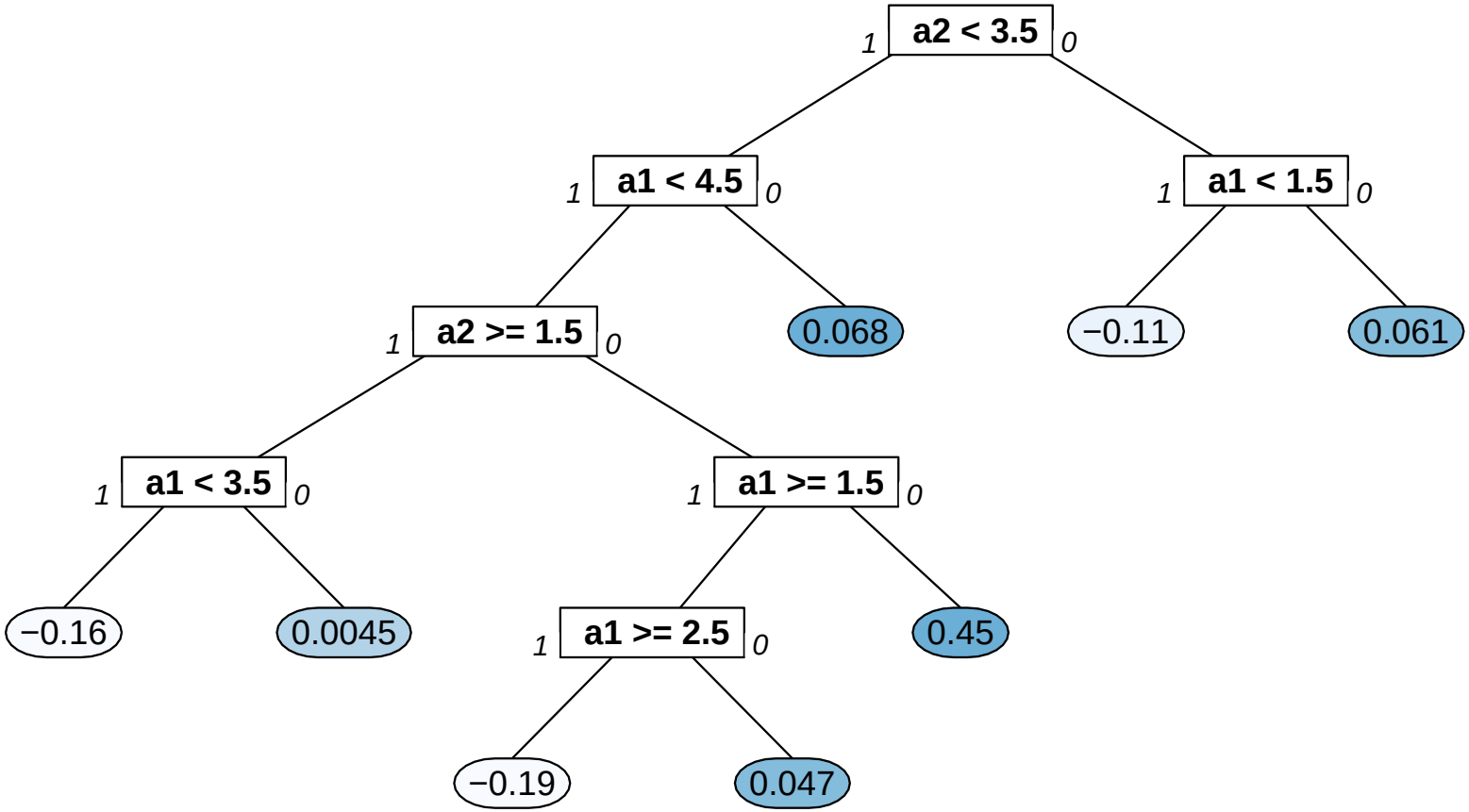
- **kryterium stopu:** na podstawie rozrzutu wartości funkcji docelowej – np. wystarczająco małe odchylenie standardowe,
- **wybór wartości dla liścia:** wartość minimalizująca stratę przy predykcji za pomocą stałej (średnia w przypadku straty kwadratowej),
- **kryterium wyboru podziału:** na podstawie rozrzutu wartości funkcji docelowej – np. średnie ważone odchylenie standardowe po podziale.

Drzewa regresji

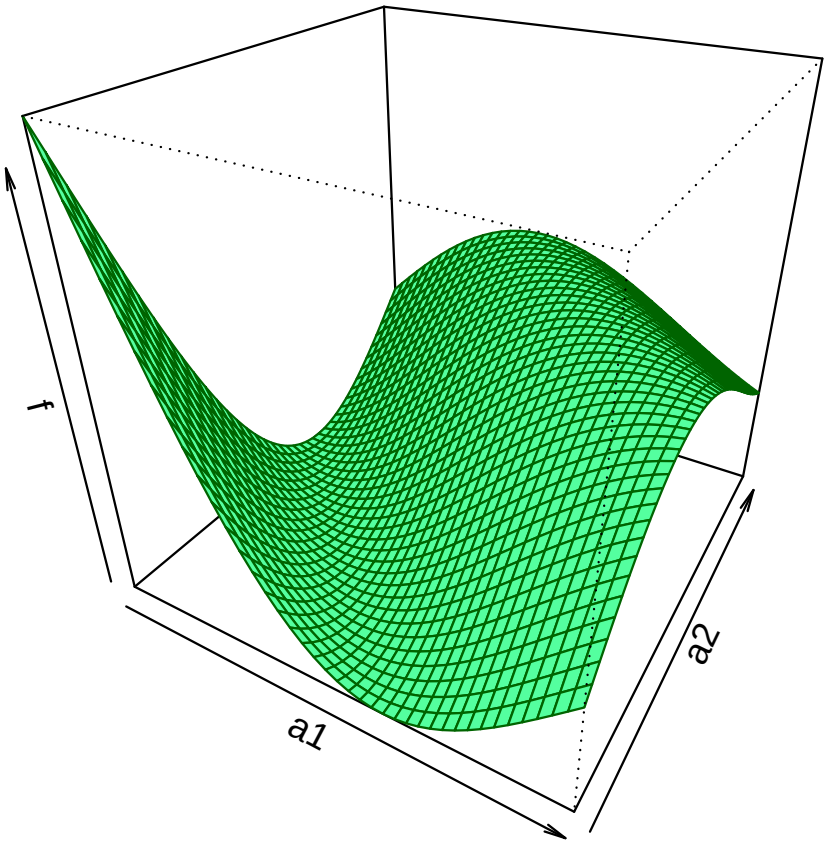
Przycinanie: analogiczne jak dla drzew decyzyjnych, lecz z kryteriami uwzględniającymi błąd regresji (np. błąd średniokwadratowy) zamiast błędu klasyfikacji.

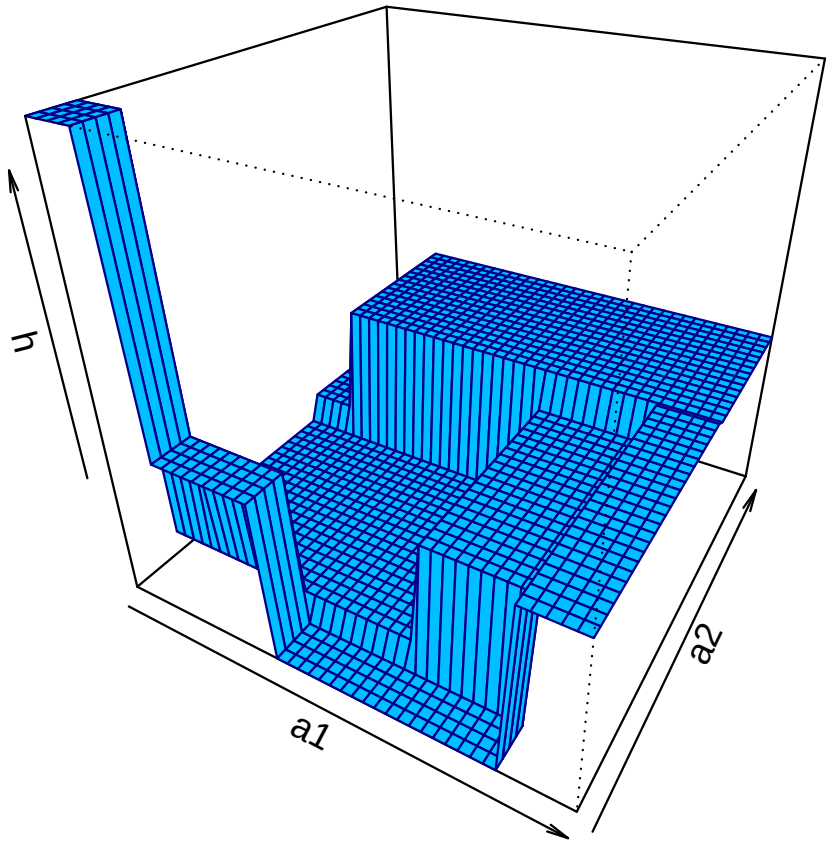
Ograniczenia: reprezentacja kawałkami stała („schodkowa"), trudno o zadowalającą równowagę między niedopasowaniem a nadmiernym dopasowaniem.

Drzewa regresji



Drzewa regresji



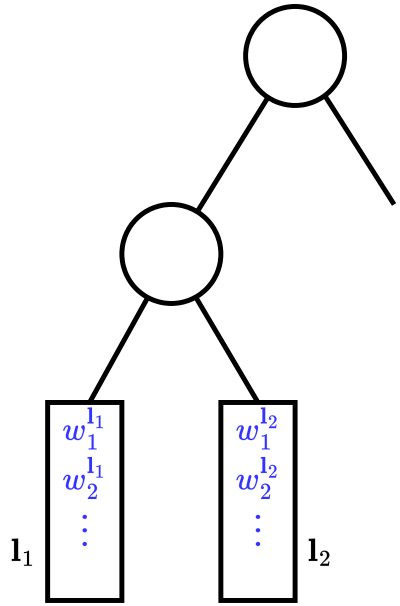


Drzewa modeli

Dekompozycja dziedziny: jak w drzewie decyzyjnym i drzewie regresji.

Węzły, gałęzie: jak w drzewie decyzyjnym i drzewie regresji.

Liście: liniowe modele regresji z parametrami wyznaczanymi na podstawie podzbiorów przykładów trenujących odpowiadających liściom, zwykle z pominięciem atrybutów wykorzystanych wyżej do podziału (na ścieżce prowadzącej od korzenia do liścia).



Drzewa modeli

Budowa: jak dla drzew regresji, z tworzeniem modeli linowych w liściach zamiast stałej wartości.

Ograniczenie: reprezentacja kawałkami liniowa.

Wariant wygładzony: modele liniowe także w węzłach, przy predykcji wyznaczana średnia ważona predykcji modelu z liścia i wszystkich węzłów na ścieżce od korzenia do liścia.