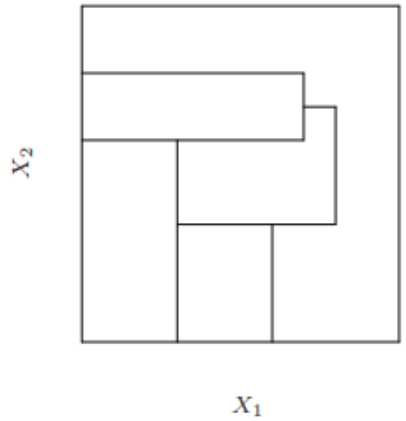
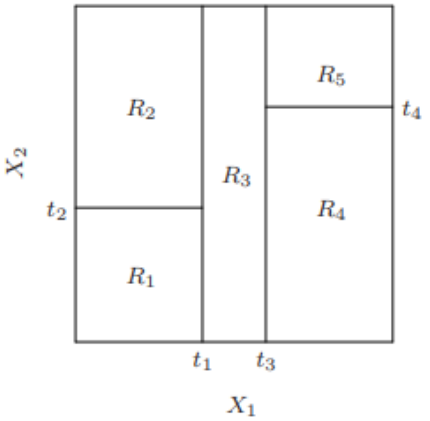
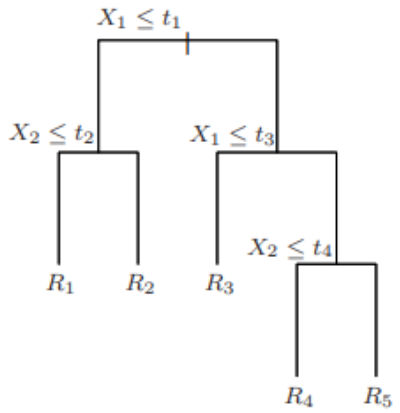
Metody oparte na drzewach

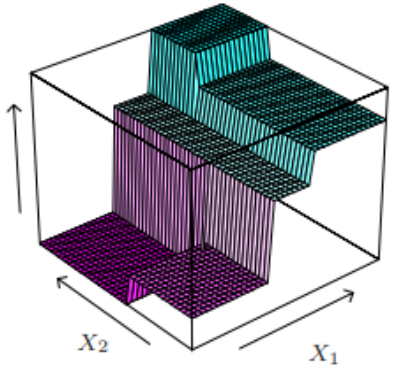
Wprowadzenie

- Metody oparte na drzewach mają zastosowanie zarówno w problemach regresyjnych jak i klasyfikacyjnych
- Stosowane są szczególnie często , gdy funkcyjna postać związku pomiędzy predyktorami a zmienną zależną jest nieznana lub ciężka do ustalenia.
- Kluczową zaletą rekursywnego drzewa binarnego jest jego interpretowalność.









Drzewa regresyjne

- Załóżmy, że mamy *N* obserwacji par (x_i, y_i), dla i = 1,…,*N*, gdzie $x_i = (x_{i1}, \ldots, x_{ip})$.
- Załóżmy też, że mamy podział na M regionów $R_1, R_2, \ldots, R_M$ i modelujemy odpowiedź jako stałą c_m w każdym regionie:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m).$$

- Przyjmując za kryterium minimalizację MSE estymator c_m jest średnią z odpowiedzi w danym regionie:

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m)$$

Proces budowania drzewa regresyjnego

1. Dzielimy przestrzeń predyktorów na J rozłącznych regionów $R_1,\ldots,R_J$.
2. Dla każdej obserwacji, która wpada do regionu R_J dokonujemy tej samej predykcji, biorąc średnią z wartości zmiennej odpowiedzi dla obserwacji z regionu R_J.

- Podziału na J regionów szukamy tak, aby zminimalizować RSS. Ponieważ jest to komputerowo niemożliwe do obliczenia wszystkich możliwych podziałów, stosujemy podejście rekursywnego podziału binarnego. Najpierw wybieramy predyktor X_j, następnie punkt odcięcia s, tak aby podział przestrzeni prektorów na dwa regiony $\{X|X_j <s\}$ i $\{X|X_j>s\}$ skutkował najmniejszym RSS.
- Tj. Szukamy takich j i s, które minimalizuja wyrażenie

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

- Następnie powtarzamy ten proces, tylko tym razem zamiast dzielić całą przestrzeń, dzielimy jeden z wcześniej zidentyfikowanych regionów.
- Kiedy wszystkie J regionów zostanie zidentyfikowanych, dokonujemy predykcji biorąc średnią wartość zmiennej objaśnianej dla każdego z regionów osobno.

Rozmiar drzewa

- Rozmiar drzewa jest parametrem odpowiedzialnym za złożoność modelu, a optymalny rozmiar drzewa powinien być dobierany na podstawie danych.
- Preferowaną strategią jest zbudowanie dużego drzewa T_0, a następnie przycięcie go za pomocą przycinania o złożoności kosztowej (cost-complexity pruning).
- Zdefiniujmy poddrzewo T, uzyskane z przycięcia drzewa T_0. Końcowe węzły indeksujemy poprzez m, gdzie węzeł m odpowiada regionowi R_m.
- Przez |T| oznaczmy liczbę węzłów końcowych w T.
- Przyjmijmy:

$$N_m = \#\{x_i \in R_m\}, \qquad \hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i, \qquad Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

- Wtedy możemy zdefiniować kryterium złożoności kosztów jako:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

- Naszym ostatecznym drzewem jest $T_{\hat{\alpha}}$.

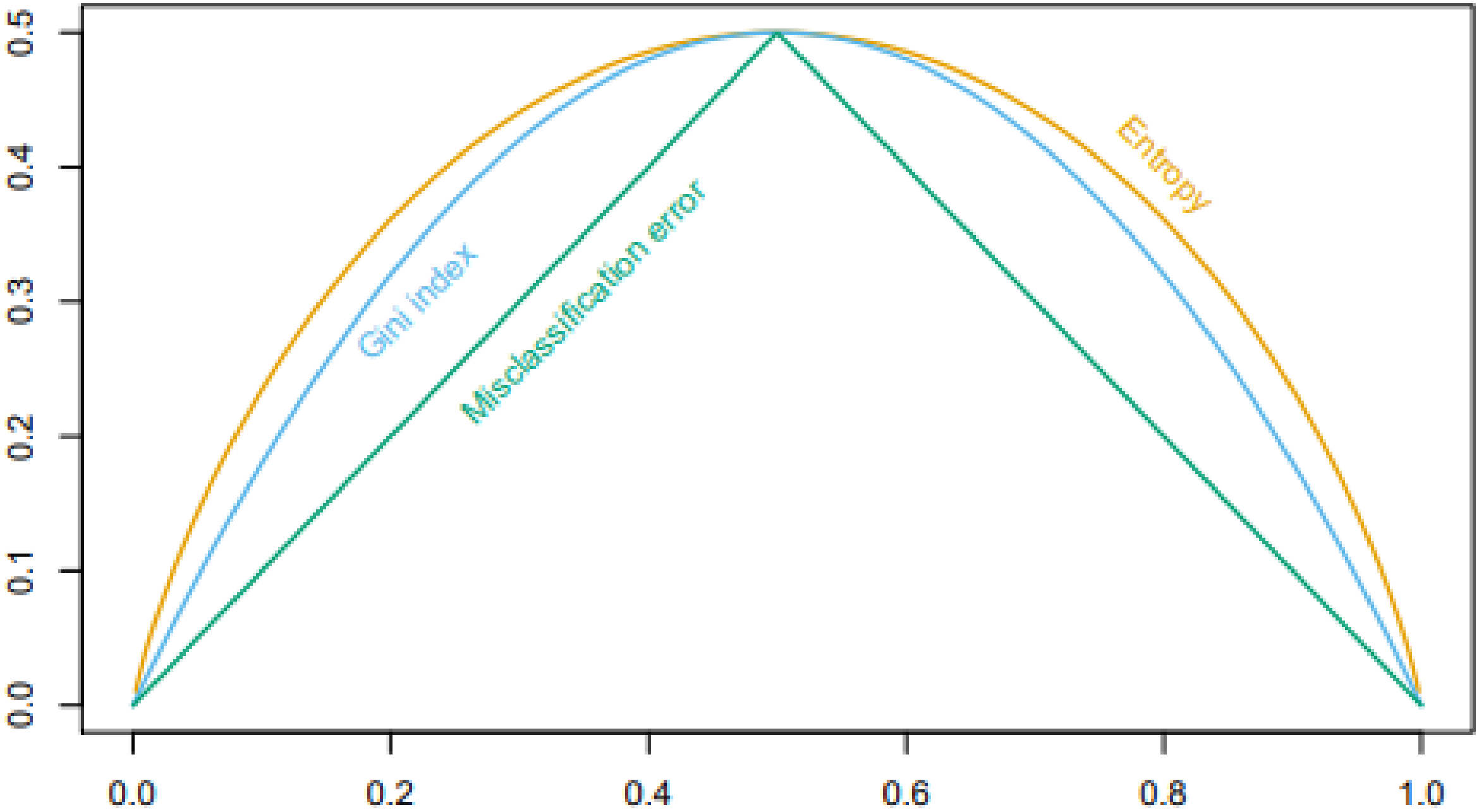
Drzewa klasyfikacyjne

- Kryteria do podziałów węzłów korzystają z proporcji obserwacji każdej klasy w węzłach

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

- Miary zanieczyszczenia węzła:

Misclassification error:	$\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}.$
Gini index:	$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}).$
Cross-entropy or deviance:	$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$


0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Entropy
Gini index
Misclassification error

Ograniczenia i problemy metody

- Zmienne kategoryczne z dużą liczbą możliwych, nieuporządkowanych wartości.
- Niestabilność drzew.
- Brak gładkości przestrzeni predykcyjnej.
- Trudność w modelowaniu struktury addytywnej.

Dziękuje za uwagę

Zuzanna Różak