Zaawansowane uczenie maszynowe:
wykład 2

Paweł Cichosz

Charakterystyka rozkładu atrybutów dyskretnych

Wygładzone prawdopodobieństwo:

$$P_{S,m,p_0}(a = v) = \frac{|S_{a=v}| + mp_0}{|S| + m}$$

wygładzanie Cestnika (m-estymacja)

$$P_{S,l}(a = v) = \frac{|S_{a=v}| + l}{|S| + l|A|}$$

wygładzanie Laplace'a

Entropia:

$$E_S(a) = \sum_{v \in A} -P_S(a = v) \log P_S(a = v)$$

Indeks Giniego:

$$\mathrm{GI}_S(a) = 1 - \sum_{v \in A} P_S^2(a = v)$$

Miary zależności atrybutów dyskretnych

Entropia warunkowa:

$$E_S(a_1|a_2) = \sum_{v \in A_2} P_S(a_2 = v) E_{S_{a_2=v}}(a_1)$$

Informacja wzajemna:

$$\begin{aligned} I_S(a_1, a_2) &= \sum_{v_1 \in A_1} \sum_{v_2 \in A_2} P_S(a_1 = v_1, a_2 = v_2) \log \frac{P_S(a_1 = v_1, a_2 = v_2)}{P_S(a_1 = v_1) \cdot P_S(a_2 = v_2)} \\ &= E_S(a_1) - E_S(a_1|a_2) = E_S(a_2) - E_S(a_2|a_1) \end{aligned}$$

Symetryczna niepewność:

$$U_S(a_1, a_2) = \frac{2 I_S(a_1, a_2)}{E_S(a_1) + E_S(a_2)}$$

– znormalizowana postać informacji wzajemnej kompensująca wpływ „arności” atrybutów.

PAC-uczenie się

PAC (*probably approximately correct*): $P(e_{\Omega,c}(h) \leq \epsilon) \geq 1 - \delta$

Klasa pojęć $\mathbb{C}$ dla dziedzinie X jest PAC-nauczalna za pomocą przestrzeni modeli $\mathbb{H}$, jeśli

- istnieje algorytm uczenia się używający $\mathbb{H}$,
- którego uruchomienie z dostępem do źródła przykładów $\mathrm{EX}(\Omega, c)$ oraz z parametrami ϵ i δ,
- daje w wyniku z prawdopodobieństwem co najmniej $1 - \delta$ model $h \in \mathbb{H}$, dla którego $e_{\Omega,c}(h) \leq \epsilon$,
- dla dowolnego pojęcia $c \in \mathbb{C}$, dowolnego rozkładu prawdopodobieństwa Ω na X oraz dowolnych $0 < \epsilon < 1$ i $0 < \delta < 1$.

PAC-ograniczenia

Ograniczenia na liczbę przykładów: ile przykładów trenujących wystarczy, aby z prawdopodobieństwem $1 - \delta$ uzyskać model o błędzie rzeczywistym nieprzekraczającym ϵ?

Ograniczenia na błąd: jaki poziom błędu rzeczywistego modelu można zagwarantować z prawdopodobieństwem $1 - \delta$, jeśli wykorzystano m przykładów trenujących?

Spójne uczenie się

Algorytm zwraca spójny model (o zerowym błędzie na zbiorze trenującym) albo zawodzi, jeśli takiego modelu nie ma w przestrzeni modeli $\mathbb{H}$.

$$m \geq \frac{1}{\epsilon}(\ln |\mathbb{H}| + \ln \frac{1}{\delta})$$
$$\epsilon \geq \frac{1}{m}(\ln |\mathbb{H}| + \ln \frac{1}{\delta})$$

- Algorytmy spójne używające skończonej przestrzeni modeli mogą osiągnąć dowolnie mały błąd, mimo podatności na nadmierne dopasowanie (wystarczy odpowiednio wiele przykładów).
- Dla ustalonej liczby przykładów można określić, na jak niski poziom błędu rzeczywistego można liczyć.

Agnostyczne uczenie się

Brak pewności, czy $c \in \mathbb{H}$ – nie można ograniczyć poziomu błędu modelu, ale można ograniczyć poziom różnicy między błędem rzeczywistym a błędem na zbiorze trenującym:

$$P(e_{\Omega,c}(h) \leq e_{T,c}(h) + \epsilon) \geq 1 - \delta$$

$$m \geq \frac{1}{2\epsilon^2}(\ln |\mathbb{H}| + \ln \frac{1}{\delta})$$

$$\epsilon \geq \sqrt{\frac{1}{2m}(\ln |\mathbb{H}| + \ln \frac{1}{\delta})}$$

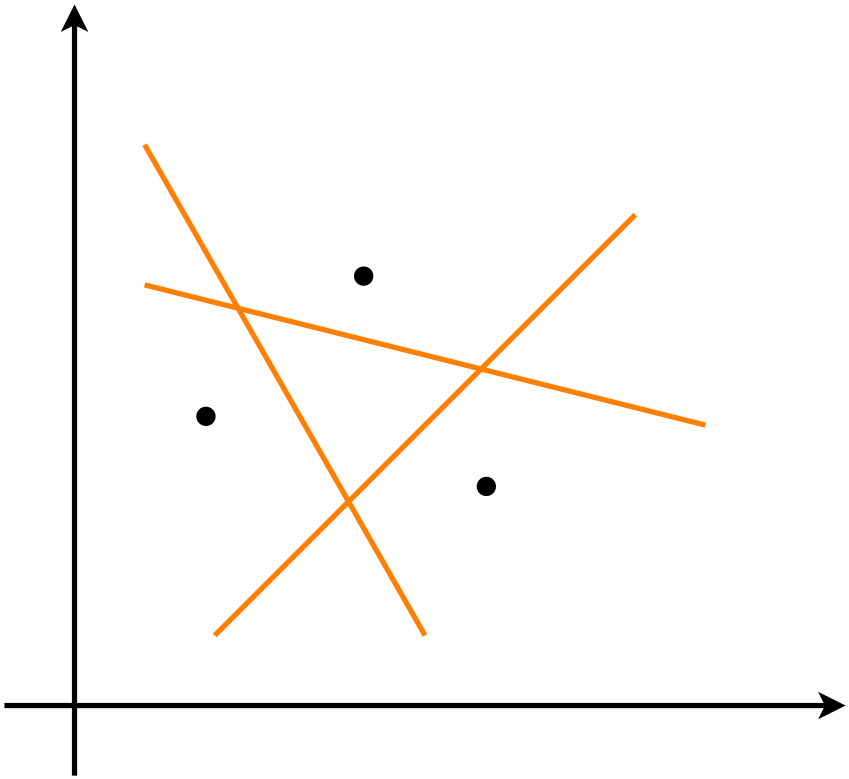
$$e_{\Omega,c}(h) \leq e_{T,c}(h) + \sqrt{\frac{1}{2m}(\ln |\mathbb{H}| + \ln \frac{1}{\delta})}$$

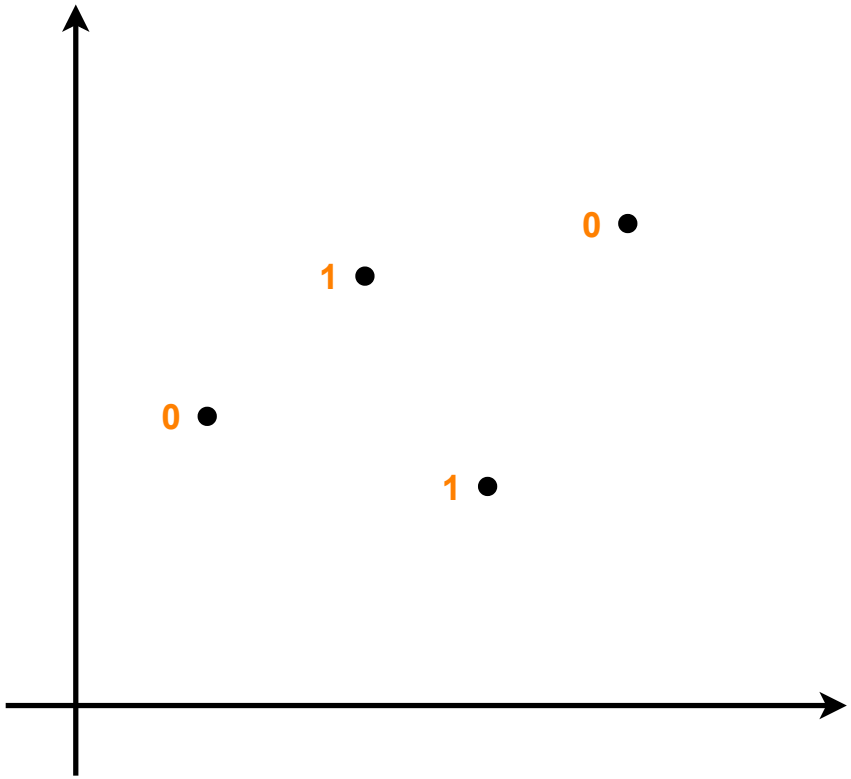
- Ograniczenie dotyczy to wszystkich modeli.
- Nie ma gwarancji, że algorytm znajdzie najlepszy model.

Wymiar VC

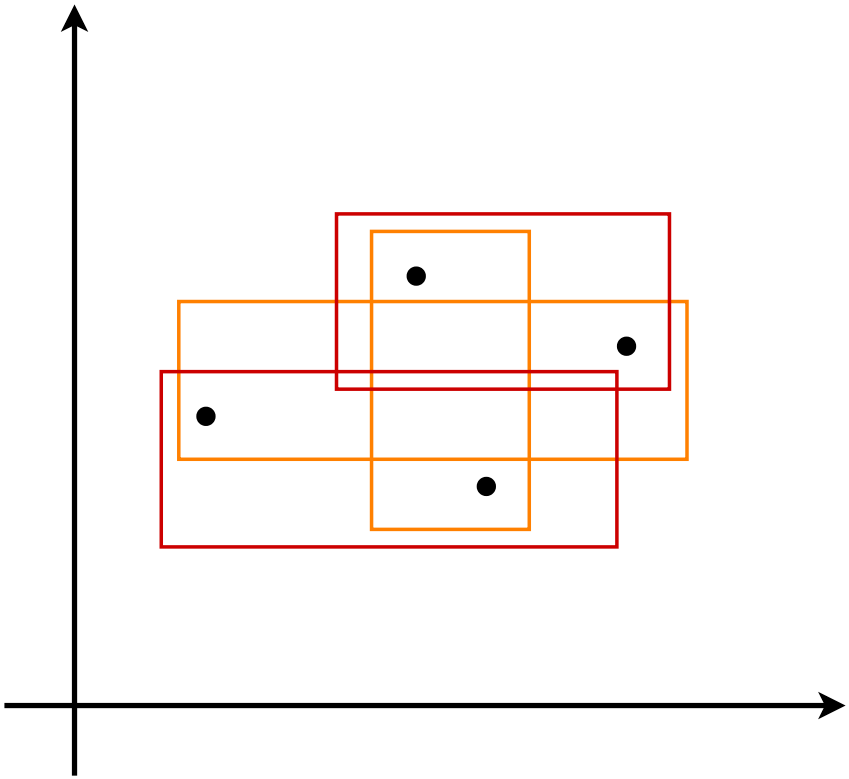
- *Maksymalna liczba przykładów, dla których przestrzeń modeli zapewnia realizację wszystkich etykietowań.*
- Dla k przykładów i $C = \{0, 1\}$ istnieje 2^k możliwych etykietowań.
- $\mathrm{VC}(\mathbb{H})$ – maksymalna wartość k taka, że istnieje k przykładów z X, dla których każde spośród 2^k możliwych etykietowań jest realizowane przez pewien model z $\mathbb{H}$.
- $\mathrm{VC}(\mathbb{H}) = \infty$ jeśli dla dowolnego k istnieje k przykładów z X, dla których każde spośród 2^k możliwych etykietowań jest realizowane przez pewien model z $\mathbb{H}$.

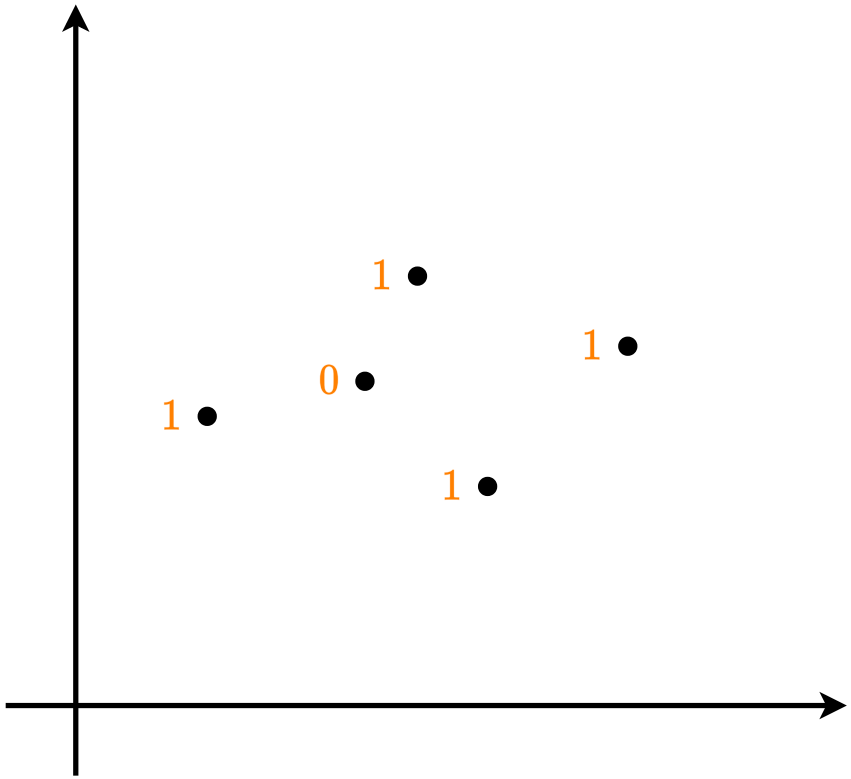
Wymiar VC





Wymiar VC





Ograniczenia oparte na wymiarze VC

Spójne uczenie się: do uzyskania przez spójny algorytm uczenia się z prawdopodobieństwem co najmniej $1-\delta$ modelu o błędzie rzeczywistym nieprzekraczającym ϵ wystarczy:

$$m \geq \frac{1}{\epsilon}\left(4\log_2\frac{2}{\delta} + 8\mathrm{VC}(\mathbb{H})\log_2\frac{13}{\epsilon}\right)$$

przykładów trenujących.

Agnostyczne uczenie się: z prawdopodobieństwem $1-\delta$:

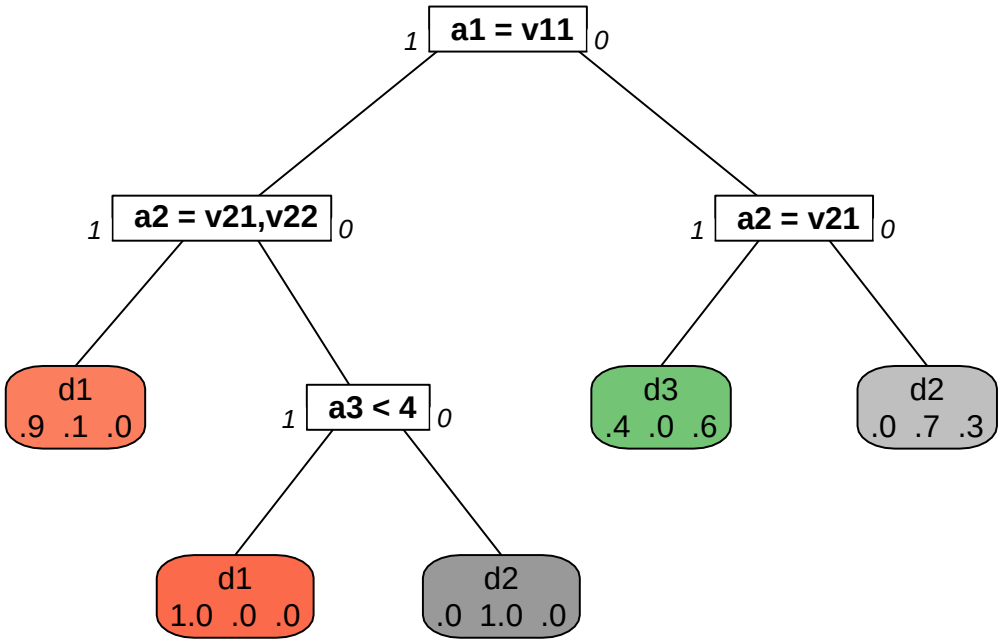
$$e_{\Omega,c}(h) \leq e_{T,c}(h) + \sqrt{\frac{1}{m}\left(\mathrm{VC}(\mathbb{H})\left(\ln\frac{2m}{\mathrm{VC}(\mathbb{H})}+1\right)+\ln\frac{4}{\delta}\right)}$$

1. Uzupełnienie wiedzy ze statystyki i teorii informacji

2. Podstawowe wyniki obliczeniowej teorii uczenia się

3. Podstawowe algorytmy uczenia się (część 1)
   - Drzewa decyzyjne

Drzewa decyzyjne



Drzewa decyzyjne: reprezentacja modelu

Węzły: podziały (testy) na podstawie warunków dotyczących wartości atrybutów $t : X \to R_t$.

Gałęzie: dla każdego wyniku $r \in R_t$ podziału t w węźle prowadzą z tego węzła do jego węzłów potomnych.

Liście: klasy lub prawdopodobieństwa klas.

Proces predykcji: przykład x propagowany od korzenia drzewa do liścia ścieżką wyznaczaną przez wyniki podziałów w kolejnych odwiedzonych węzłach:

$h(x)$: klasa z osiągniętego liścia,
$P(d|x)$: prawdopodobieństwo klasy d z osiągniętego liścia.

Drzewa decyzyjne: typy podziałów

Wielowartościowe na podstawie wartości atrybutu: $t(x) = a(x)$

Binarne na podstawie równości:

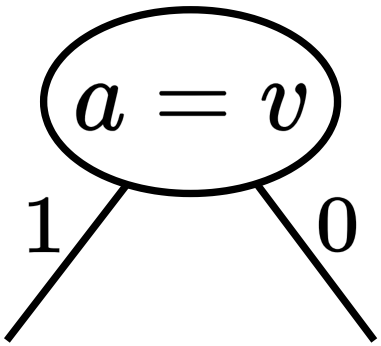
$$t(x) = \begin{cases} 1 & \text{jeżeli } a(x) = v \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

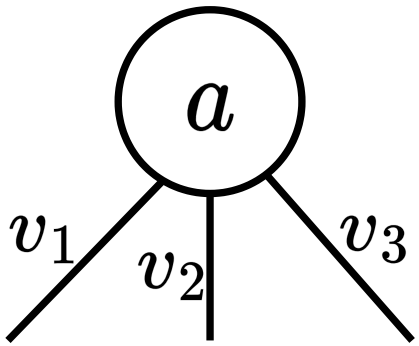
Binarne na podstawie nierówności:

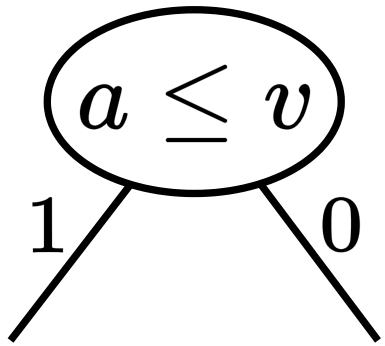
$$t(x) = \begin{cases} 1 & \text{jeżeli } a(x) \leq v \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

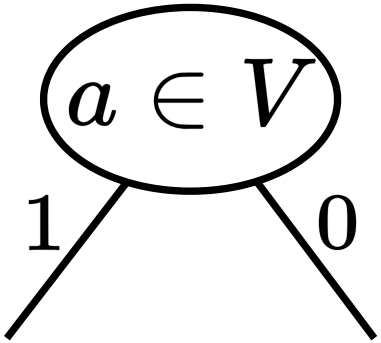
Warianty uogólnione: podzbiory wartości atrybutów zamiast pojedynczych wartości.

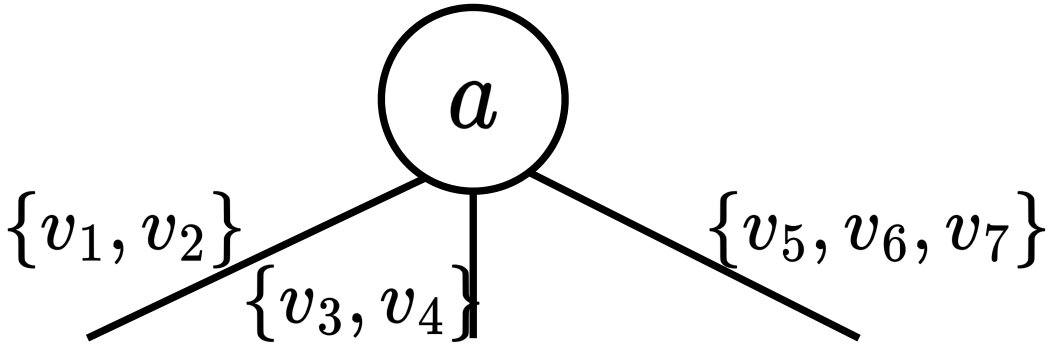
Drzewa decyzyjne: typy podziałów



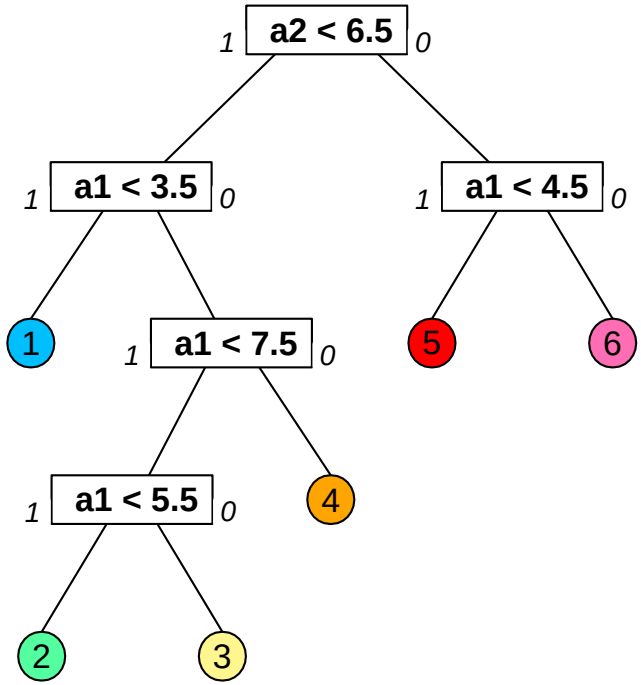


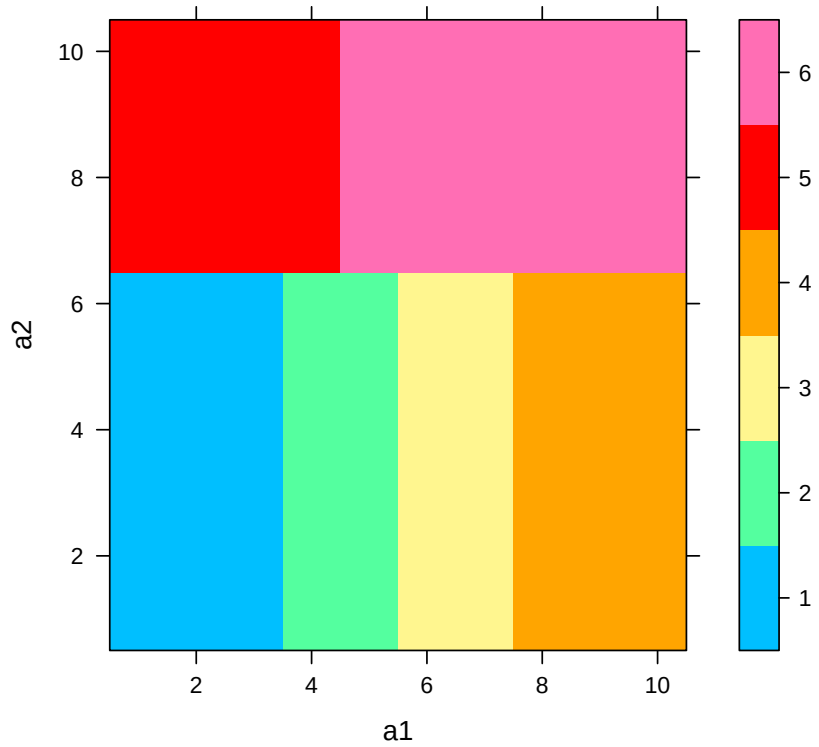






Drzewa decyzyjne: podział dziedziny





Drzewa decyzyjne: wymiar VC

Teoretyczny, bez atrybutów ciągłych: N, gdzie N – liczba możliwych kombinacji wartości atrybutów.

Teoretyczny, z atrybutami ciągłymi: ∞.

Efektywny: redukowany przez zastosowanie kryteriów stopu i przycinania.

Drzewa decyzyjne: budowa

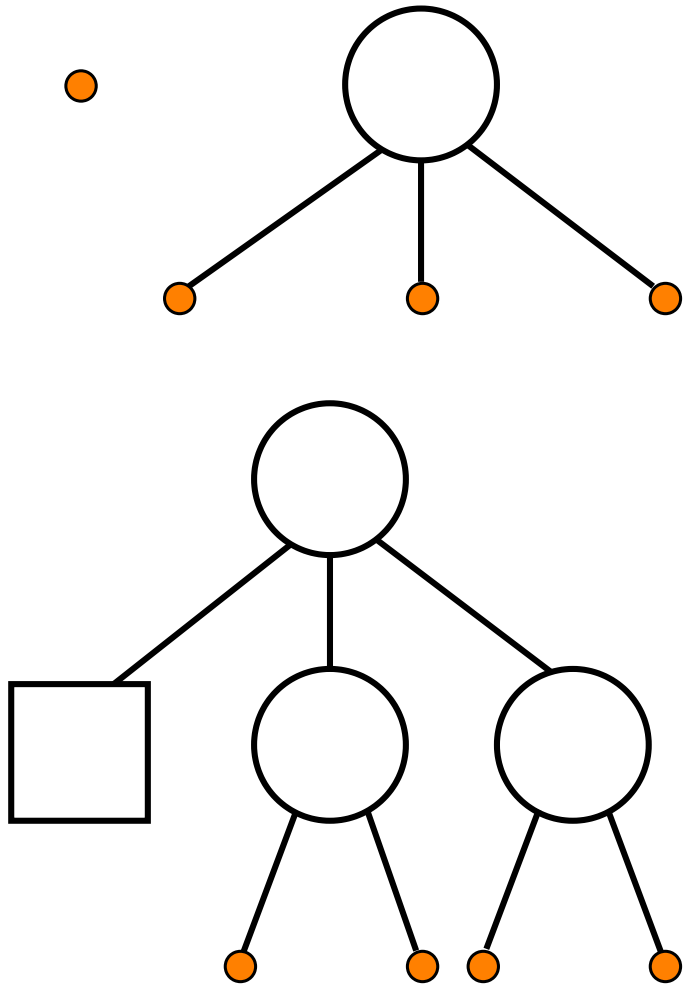
- Sekwencja decyzji:

  kryterium stopu: węzeł czy liść?

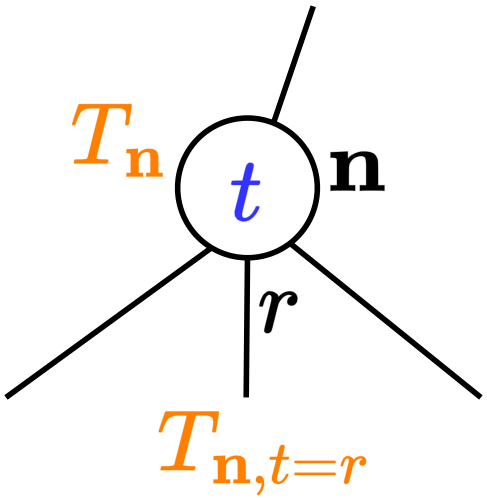
  wybór klasy dla liścia: jaka predykcja najlepsza w liściu?

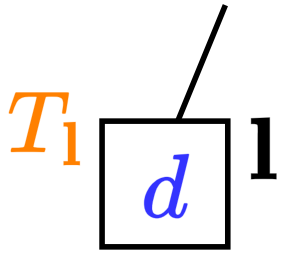
  wybór podziału dla węzła: jaki podział najlepszy w węźle?

- Ten sam schemat powtarzany dla korzenia drzewa, jego węzłów potomnych, ich węzłów potomnych itd.



Drzewa decyzyjne: węzeł i liść





Drzewa decyzyjne: kryterium stopu

Jednolita klasa: pozostały przykłady dokładnie jednej klasy.

Brak przykładów: brak przykładów.

Brak możliwości podziału: każdy możliwy podział osiąga tylko jeden wynik (nie dzieli).

Warianty rozluźnione: dominacja jednej klasy, mała liczba przykładów, najlepszy podział zbyt słaby.

Wpływ na właściwości drzewa: kryteria stopu (w nierozluźnionej postaci) gwarantują minimalizację błędu na zbiorze trenującym oraz (w rozluźnionej postaci) kontrolują poziom dopasowania do danych.

Uzupełniające kryterium: maksymalna liczba poziomów.

Drzewa decyzyjne: kryterium wyboru podziału

Motywacja: preferencja dla prostych drzew ograniczająca ryzyko nadmiernego dopasowania.

Realizacja: ocena jakości podziałów na podstawie nieczystości rozkładu klas po podziale, wybór najlepszego podziału w każdym węźle (ze skończonego zbioru kandydatów).

Ocena jakości podziału:

- węzeł $\mathbf{n}$,
- zbiór przykładów trenujących $T_{\mathbf{n}}$,
- rozważany podział $t : X \to R_t$,
- dla każdego wyniku podziału $r \in R_t$ odpowiedni podzbiór przykładów trenujących $T_{\mathbf{n},t=r} = \{x \in T_{\mathbf{n}} \mid t(x) = r\}$.

Drzewa decyzyjne: kryterium wyboru podziału

Entropia warunkowa:

$$E_{T_{\mathbf{n}}}(c|t) = \sum_{r \in R_t} \frac{|T_{\mathbf{n},t=r}|}{|T_{\mathbf{n}}|} E_{T_{\mathbf{n},t=r}}(c)$$

$$E_{T_{\mathbf{n},t=r}}(c) = \sum_{d \in C} -P_{T_{\mathbf{n},t=r}}(c=d) \log P_{T_{\mathbf{n},t=r}}(c=d)$$

Warunkowy indeks Giniego:

$$\mathrm{GI}_{T_{\mathbf{n}}}(c|t) = \sum_{r \in R_t} \frac{|T_{\mathbf{n},t=r}|}{|T_{\mathbf{n}}|} \mathrm{GI}_{T_{\mathbf{n},t=r}}(c)$$

$$\mathrm{GI}_{T_{\mathbf{n},t=r}}(c) = 1 - \sum_{d \in C} P^2_{T_{\mathbf{n},t=r}}(c=d)$$

Drzewa decyzyjne: zapobieganie nadmiernemu dopasowaniu

Wybór podziałów: preferencja dla prostszych drzew.

Rozluźnianie kryteriów stopu: wcześniejsze tworzenie liści.

Przycinanie: zastępowanie „słabych” węzłów liśćmi – będzie omawiane dalej.

Drzewa decyzyjne: właściwości

- Zwykle dobra jakość predykcji (chociaż często inne algorytmy dają nieco lepsze modele).
- Reprezentacja modelu czytelna dla człowieka.
- Możliwa konwersja do postaci reguł i połączenie z wiedzą ekspercką.
- Podatne na nadmierne dopasowanie – konieczne staranne dobranie kryteriów stopu lub zastosowanie przycinania.
- Znakomite jako komponenty modeli zespołowych (o których będzie mowa za kilka tygodni).