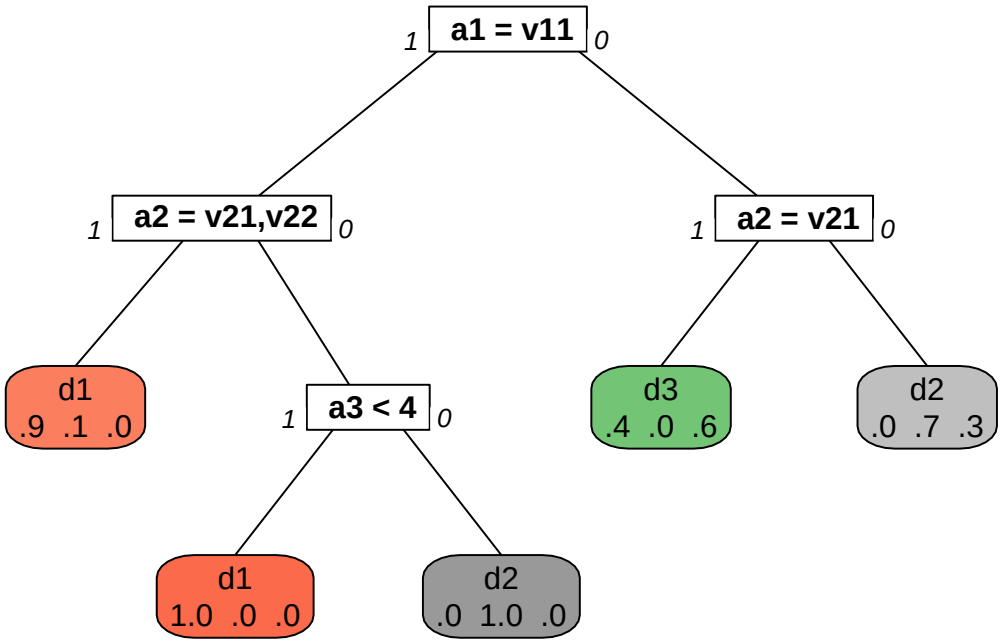
# Uczenie maszynowe:
# *wykład 8*

Paweł Cichosz

# Struktura drzewa



# Struktura drzewa

Węzły: podziały (testy) na podstawie warunków dotyczących wartości atrybutów $t : X \to R_t$ (pełnią analogiczną rolę jak selektory w regułach).

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

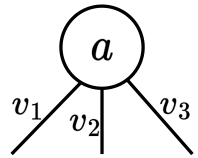
Liście: klasy i prawdopodobieństwa klas.

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

# Podziały dla atrybutów dyskretnych

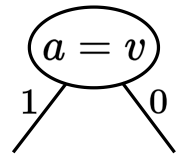
Wielowartościowe na podstawie wartości atrybutu:

$$t(x) = a(x)$$



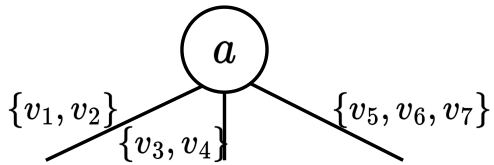

Binarne na podstawie równości:

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



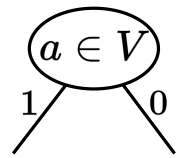

Wielowartościowe na podstawie podzbiorów wartości atrybutu:

$$t(x) = \begin{cases} 1 & \text{jeżeli } a(x) \in V_1 \\ 2 & \text{jeżeli } a(x) \in V_2 \\ \dots & \\ k & \text{jeżeli } a(x) \in V_k \end{cases}$$




Binarne na podstawie przynależności do zbioru:

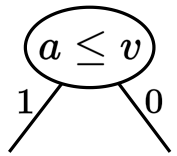
$$t(x) = \begin{cases} 1 & \text{jeżeli } a(x) \in V \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$



# Podziały dla atrybutów ciągłych

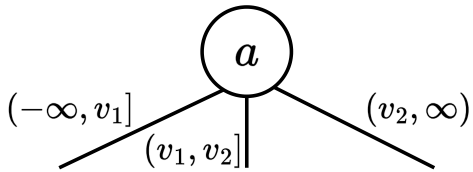
## Binarne na podstawie nierówności:

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

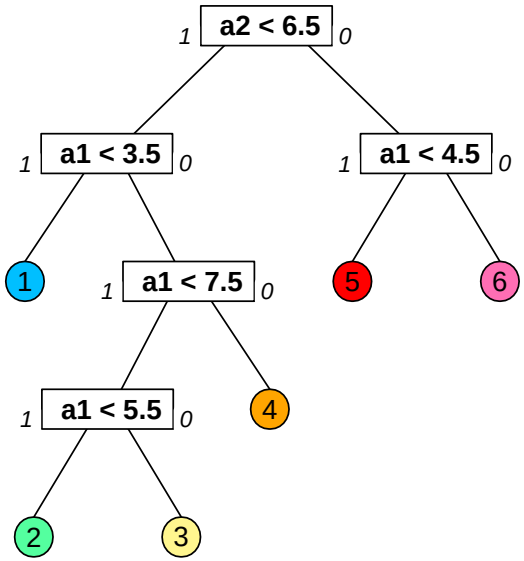


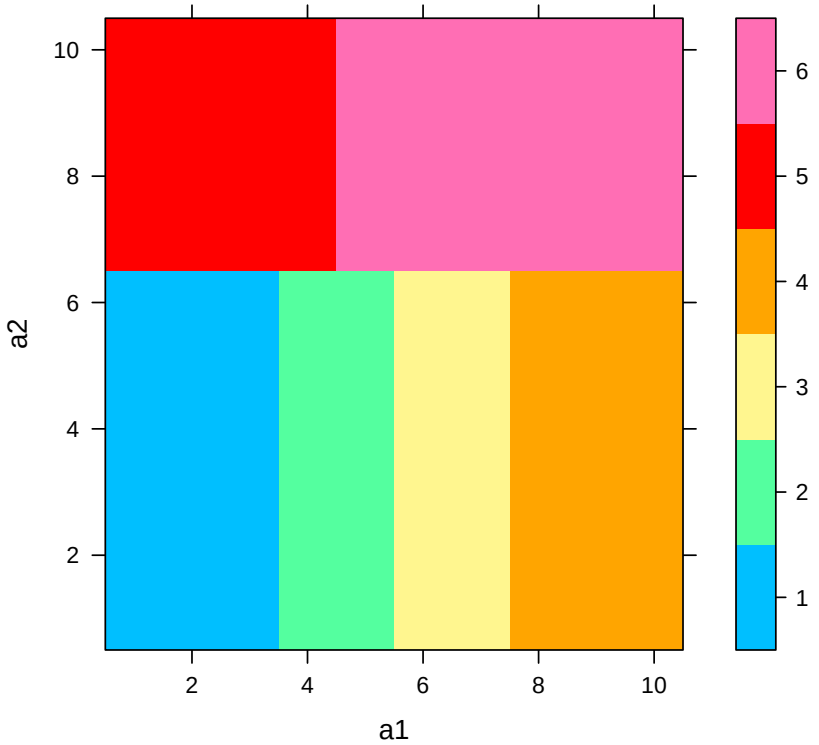

## Wielowartościowe na podstawie przedziałów wartości atrybutu:

$$t(x) = \begin{cases} 1 & \text{jeżeli } a(x) \in (-\infty, v_1] \\ 2 & \text{jeżeli } a(x) \in (v_1, v_2] \\ \dots & \\ k & \text{jeżeli } a(x) \in (v_{k-1}, \infty) \end{cases}$$



# Dekompozycja dziedziny




Z każdym węzłem lub liściem $\mathbf{n}$ jest związany odpowiedni podzbiór $X_{\mathbf{n}}$ dziedziny $X$ i odpowiedni podzbiór $S_{\mathbf{n}}$ dowolnego zbioru danych $S$, zawierający przykłady, które docierają do tego węzła lub liścia, jeśli są propagowane z korzenia drzewa ścieżkami wyznaczanymi przez wyniki kolejnych podziałów.

# Wymiar VC

Teoretyczny, bez atrybutów ciągłych: liczba możliwych kombinacji wartości atrybutów.

Teoretyczny, z atrybutami ciągłymi: $\infty$.

Efektywny: redukowany przez zastosowanie kryteriów stopu i przycinania (jak zobaczymy dalej).

# Schemat algorytmu

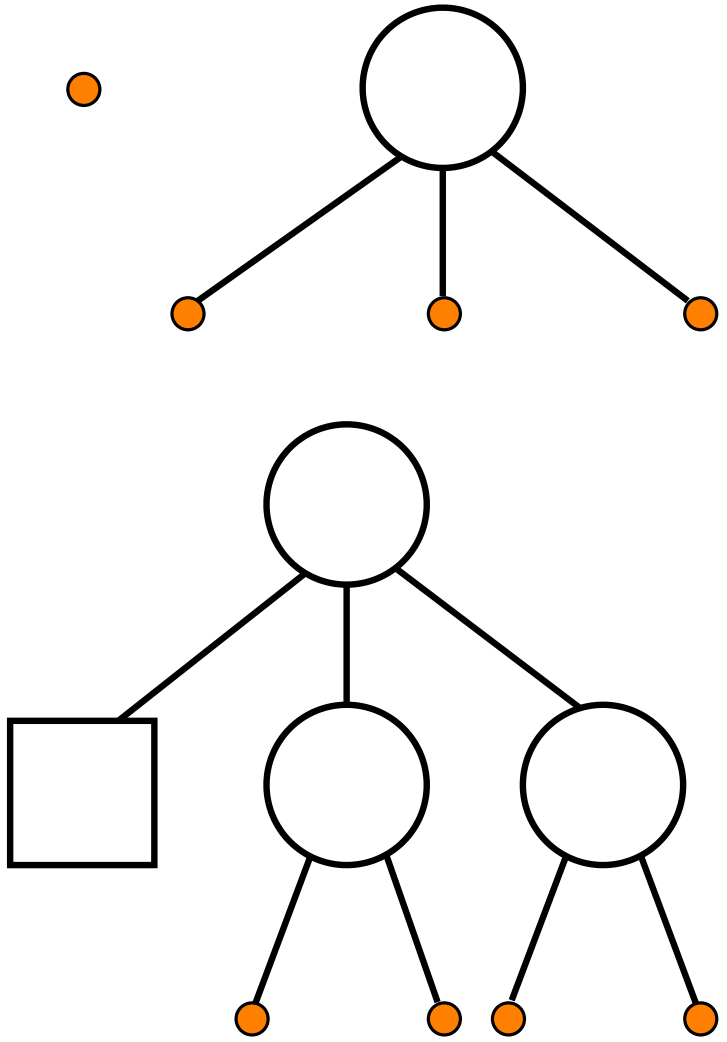
- Sekwencja decyzji:

  kryterium stopu: węzeł czy liść?

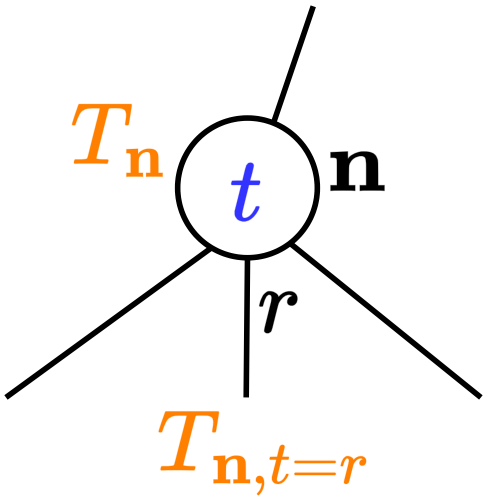
  wybór klasy dla liścia: jaka predykcja najlepsza w liściu?

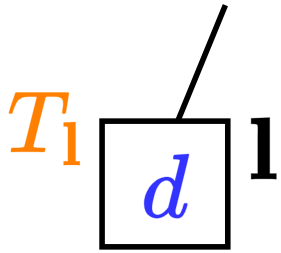
  wybór podziału dla węzła: jaki podział najlepszy w węźle?

- Ten sam schemat powtarzany dla korzenia drzewa, jego węzłów potomnych, ich węzłów potomnych itd.
- Kolejność zazwyczaj nieistotna (zwykle: w głąb albo wszerz).



# Drzewa decyzyjne: notacja



# Kryterium stopu

Podejmowana decyzja: czy węzeł $\mathbf{n}$, któremu odpowiada podzbiór przykładów trenujących $T_{\mathbf{n}}$, powinien być liściem?

Jednolita klasa: $T_{\mathbf{n}}$ zawiera przykłady dokładnie jednej klasy (wybieranej jako klasa dla liścia).

Brak przykładów: $T_{\mathbf{n}} = \emptyset$ (klasa dla liścia z węzła macierzystego).

Brak możliwości podziału: każdy możliwy podział osiąga dla $T_{\mathbf{n}}$ 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.

Uzupełniające kryterium: maksymalna liczba poziomów.

Prawdpodobieństwa klas: na podstawie rozkładu przykładów trenujących w liściu.

# 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 $T_{\mathbf{n},t=r} = \{x \in T_{\mathbf{n}} \mid t(x) = r\}$.

# Kryterium wyboru podziału

Entropia warunkowa:

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

gdzie:

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

$$P_{T_{\mathbf{n},t=r}}(c=d) = \frac{|T_{\mathbf{n},t=r,c=d}|}{|T_{\mathbf{n},t=r}|}$$

Redukcja nieczystości w wyniku podziału (przyrost informacji):

$$\Delta E_{T_{\mathbf{n}}}(c|t) = E_{T_{\mathbf{n}}}(c) - E_{T_{\mathbf{n}}}(c|t)$$

Wybór podziału: minimalizacja nieczystości po podziale (równoważnie: maksymalizacja redukcji nieczystości w wyniku podziału).

# Kryterium wyboru podziału

Warunkowy indeks Giniego:

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

gdzie:

$$G_{T_{\mathbf{n},t=r}}(c) = 1 - \sum_{d \in C} P_{T_{\mathbf{n},t=r}}(c = d)^2$$

Redukcja nieczystości w wyniku podziału:

$$\Delta G_{T_{\mathbf{n}}}(c|t) = G_{T_{\mathbf{n}}}(c) - G_{T_{\mathbf{n}}}(c|t)$$

Wybór podziału: minimalizacja nieczystości po podziale (równoważnie: maksymalizacja redukcji nieczystości w wyniku podziału).

# Przykład: pogoda

| $x$ | *outlook* | *temperature* | *humidity* | *wind* | *play* |
|---|---|---|---|---|---|
| 1 | *sunny* | *hot* | *high* | *normal* | *no* |
| 2 | *sunny* | *hot* | *high* | *high* | *no* |
| 3 | *overcast* | *hot* | *high* | *normal* | *yes* |
| 4 | *rainy* | *mild* | *high* | *normal* | *yes* |
| 5 | *rainy* | *cold* | *normal* | *normal* | *yes* |
| 6 | *rainy* | *cold* | *normal* | *high* | *no* |
| 7 | *overcast* | *cold* | *normal* | *high* | *yes* |
| 8 | *sunny* | *mild* | *high* | *normal* | *no* |
| 9 | *sunny* | *cold* | *normal* | *normal* | *yes* |
| 10 | *rainy* | *mild* | *normal* | *normal* | *yes* |
| 11 | *sunny* | *mild* | *normal* | *high* | *yes* |
| 12 | *overcast* | *mild* | *high* | *high* | *yes* |
| 13 | *overcast* | *hot* | *normal* | *normal* | *yes* |
| 14 | *rainy* | *mild* | *high* | *high* | *no* |

# Przykład: pogoda

| $x$ | *outlook* | *temperature* | *humidity* | *wind* | *play* |
|---|---|---|---|---|---|
| 1 | *sunny* | *hot* | *high* | *normal* | *no* |
| 2 | *sunny* | *hot* | *high* | *high* | *no* |
| 3 | *overcast* | *hot* | *high* | *normal* | *yes* |
| 4 | *rainy* | *mild* | *high* | *normal* | *yes* |
| 5 | *rainy* | *cold* | *normal* | *normal* | *yes* |
| 6 | *rainy* | *cold* | *normal* | *high* | *no* |
| 7 | *overcast* | *cold* | *normal* | *high* | *yes* |
| 8 | *sunny* | *mild* | *high* | *normal* | *no* |
| 9 | *sunny* | *cold* | *normal* | *normal* | *yes* |
| 10 | *rainy* | *mild* | *normal* | *normal* | *yes* |
| 11 | *sunny* | *mild* | *normal* | *high* | *yes* |
| 12 | *overcast* | *mild* | *high* | *high* | *yes* |
| 13 | *overcast* | *hot* | *normal* | *normal* | *yes* |
| 14 | *rainy* | *mild* | *high* | *high* | *no* |

Wybór podziału dla korzenia drzewa, w którym $T_{\mathbf{n}} = T$.

- Dla podziału *outlook* (stosując logarytmy dwójkowe):

$$E_{T_{outlook=sunny}}(c) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.971$$

$$E_{T_{outlook=overcast}}(c) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$E_{T_{outlook=rainy}}(c) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.971$$

- Entropia warunkowa:

$$E_T(c|outlook) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 \approx 0.694$$

- Analogicznie dla pozostałych podziałów (kontynuacja pozostaje jako ćwiczenie).