Uczenie maszynowe:
wykład 6

Paweł Cichosz

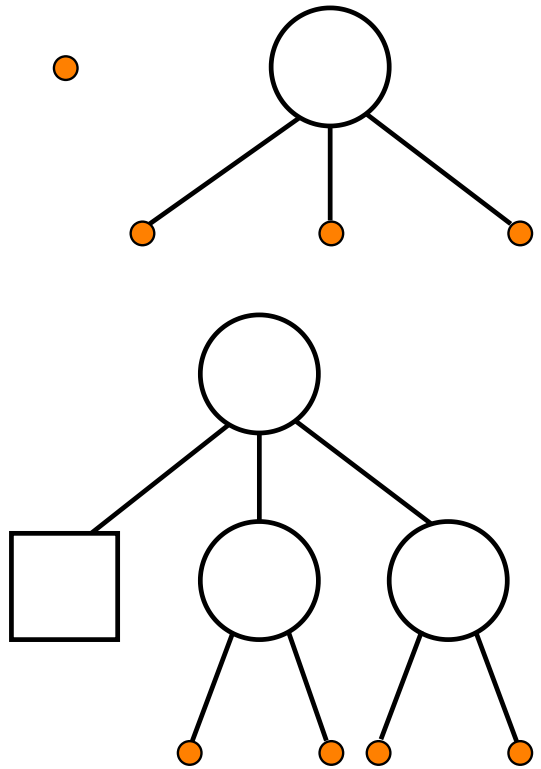
1 Budowa drzewa decyzyjnego

2 Przycinanie drzew decyzyjnych

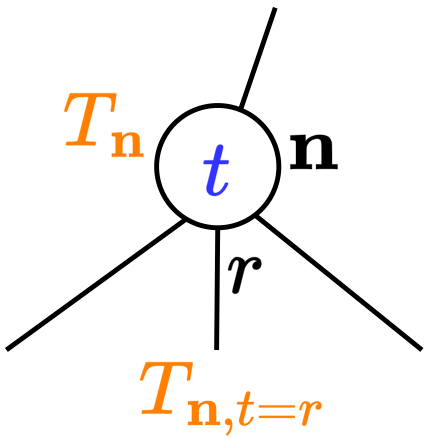
Schemat algorytmu

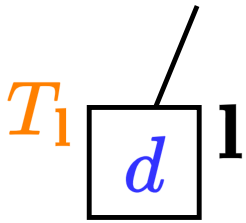
- Sekwencja decyzji:

  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).

Kryterium stopu

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):

$$E_{T_{\textit{outlook=sunny}}}(c) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.971$$
$$E_{T_{\textit{outlook=overcast}}}(c) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$
$$E_{T_{\textit{outlook=rainy}}}(c) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.971$$

- Entropia warunkowa:

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

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

1 Budowa drzewa decyzyjnego

2 Przycinanie drzew decyzyjnych

Przycinanie drzewa

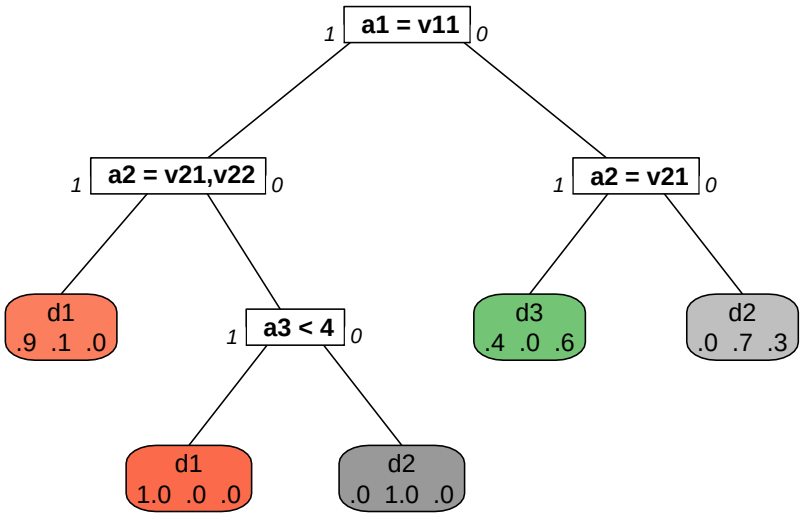
Cel: Zapobieganie nadmiernemu dopasowaniu (bardziej wymagające, ale potencjalnie skuteczniejsze niż kryterium stopu).

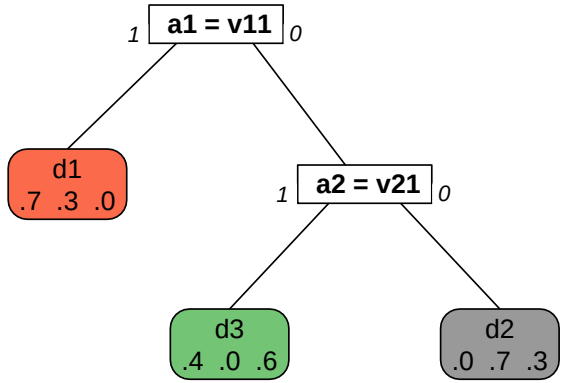
Operator: zastąpienie poddrzewa liściem.

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

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

Przycinanie drzewa





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:

$$e_R(\mathbf{l}) \leq e_R(\mathbf{n})$$

przy czym błąd liścia $e_R(\mathbf{l})$ i błąd węzła $e_R(\mathbf{l})$ wyznacza się na podstawie tych przykładów z R, które „docierają" do miejsca jego położenia w drzewie.

- Dobre jeśli możemy poświęcić osobną pulę przykładów.
- Może czasem lepiej przycinać trochę gorzej, ale budować lepiej (ale inne metody poza zakresem wykładu)?

Konwersja drzewa do zbioru reguł

- Reguła dla każdej ścieżki od korzenia do liścia:

  część warunkowa: koniunkcja warunków odpowiadających wynikom podziałów w węzłach,

  część decyzyjna: klasa z liścia.
- Reprezentacja uważana (czasem) za bardziej czytelną dla człowieka.
- Umożliwia bardziej elastyczne przycinanie (usuwanie bądź pozostawianie warunków niezależnie dla każdej reguły).

Właściwości drzew decyzyjnych

- Zwykle dobra jakość predykcji, chociaż często inne algorytmy dają nieco lepsze modele.
- Reprezentacja modeli czytelna dla człowieka.
- Podatne na nadmierne dopasowanie – konieczne staranne dobranie kryteriów stopu lub zastosowanie przycinania.
- Po niewielkiej modyfikacji mogą służyć jako modele regresji (poza zakresem wykładu).
- Przydatne jako komponenty modeli zespołowych, takich jak las losowy (omawiany na jednym z kolejnych wykładów) i *gradient boosting* (poza zakresem wykładu).