Uczenie maszynowe:
wykład 9

Paweł Cichosz

Przycinanie drzewa

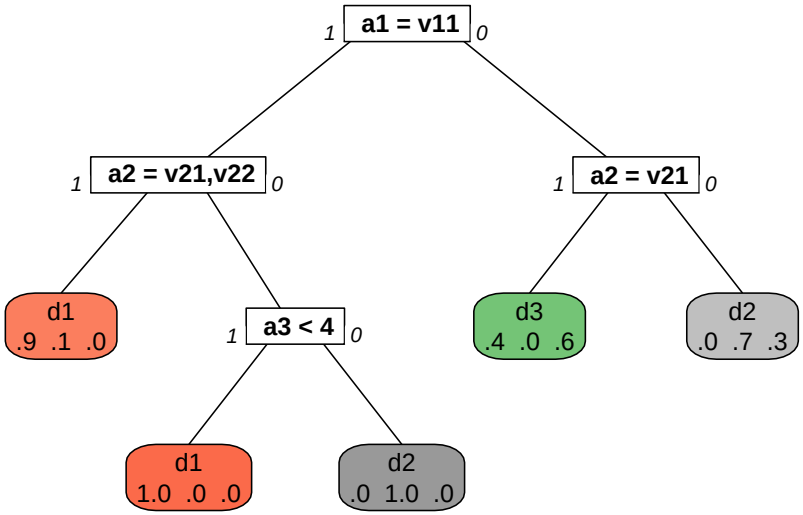
Cel: Zapobieganie nadmiernemu dopasowaniu (bardziej wymagające, ale potencjalnie skuteczniejsze niż kryterium stopu).

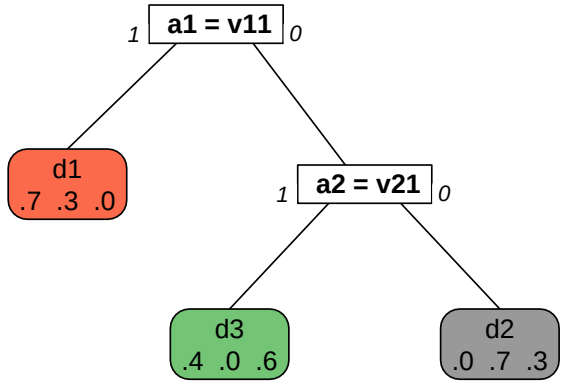
Operator: zastąpienie poddrzewa liściem.

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).

Naiwny klasyfikator bayesowski

Wzór Bayesa:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Zastosowanie do predykcji prawdopodobieństw klas:

$$\pi_d(x) = P(d|x) = P(c = d \mid a_1 = a_1(x), a_2 = a_2(x), \ldots, a_n = a_n(x))$$

Prawdopodobieństwo klasy na podstawie wartości atrybutów:

$$\begin{aligned} &P(c = d \mid a_1 = v_1, \ldots, a_n = v_n) \\ &\quad = \frac{P(c = d)P(a_1 = v_1, \ldots, a_n = v_n \mid c = d)}{P(a_1 = v_1, \ldots, a_n = v_n)} \end{aligned}$$

Założenie o niezależności:

$$P(a_1 = v_1, \ldots, a_n = v_n \mid c = d) = \prod_{i=1}^{n} P(a_i = v_i|c = d)$$

Estymacja prawdopodobieństw

Prawdopodobieństwo *a priori* klasy:

$$P(c = d) = P_T(c = d) = \frac{|T_{c=d}|}{|T|}$$

Prawdopodobieństwo warunkowe wartości atrybutu:

$$P(a_i = v_i | c = d) = P_{T_{c=d}}(a_i = v_i) = \frac{|T_{c=d,a_i=v_i}|}{|T_{c=d}|}$$

Mianownik: stała normalizująca:

$$P(a_1 = v_1, \ldots, a_n = v_n) = \sum_{d \in C} P(c = d) P(a_1 = v_1, \ldots, a_n = v_n \mid c = d)$$

Prawdopodobieństwa zerowe i prawie zerowe

Stała wartość dodatnia: $P(a_i = v_i|c = d) = \epsilon$ jeśli $T_{c=d,a_i=v_i} = \emptyset$.

Wygładzanie Laplace'a:

$$P(a_i = v_i|c = d) = \frac{|T_{c=d,a_i=v_i}| + l}{|T_{c=d}| + l|A_i|}$$

$(l \geq 0)$

Logarytm prawdopodobieństwa: mniejsze ryzyko błędów numerycznych przy mnożeniu małych prawdopodobieństw:

$$\log \left(P(c = d) \prod_{i=1}^{n} P(a_i = v_i|c = d) \right)$$

$$= \log P(c = d) + \sum_{i=1}^{n} \log P(a_i = v_i|c = d)$$

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*

Klasyfikacja przykładu:

15 | *sunny* *mild* *normal* *normal* | *no*

$$P(\textit{no}) = \frac{5}{14} \qquad P(\textit{yes}) = \frac{9}{14}$$

$$P(\textit{outlook} = \textit{sunny}|\textit{no}) = \frac{3}{5} \qquad P(\textit{outlook} = \textit{sunny}|\textit{yes}) = \frac{2}{9}$$

$$P(\textit{temperature} = \textit{mild}|\textit{no}) = \frac{2}{5} \qquad P(\textit{temperature} = \textit{mild}|\textit{yes}) = \frac{4}{9}$$

$$P(\textit{humidity} = \textit{normal}|\textit{no}) = \frac{1}{5} \qquad P(\textit{humidity} = \textit{normal}|\textit{yes}) = \frac{6}{9}$$

$$P(\textit{wind} = \textit{normal}|\textit{no}) = \frac{2}{5} \qquad P(\textit{wind} = \textit{normal}|\textit{yes}) = \frac{6}{9}$$

$$P(\textit{no}|\textit{sunny}, \textit{mild}, \textit{normal}, \textit{normal}) = \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} / P(\textit{sunny}, \textit{mild}, \textit{normal}, \textit{normal})$$

$$P(\textit{yes}|\textit{sunny}, \textit{mild}, \textit{normal}, \textit{normal}) = \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} / P(\textit{sunny}, \textit{mild}, \textit{normal}, \textit{normal})$$

$$P(\textit{sunny}, \textit{mild}, \textit{normal}, \textit{normal}) = \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} + \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{6}{9}$$

Atrybuty ciągłe

Funkcja gęstości: zastąpienie $P(a_i = v_i|c = d)$ przez $g_{d,i}(v_i)$, gdzie $g_{d,i}$ jest funkcją gęstości atrybutu a_i w klasie d – zakładany rozkład normalny z parametrami estymowanymi jako:

$$m_{T_{c=d}}(a_i) = \frac{1}{|T_{c=d}|} \sum_{x \in T_{c=d}} a_i(x)$$

$$s_{T_{c=d}}(a_i) = \sqrt{\frac{1}{|T_{c=d}| - 1} \sum_{x \in T_{c=d}} \left(a_i(x) - m_{T_{c=d}}(a_i)\right)}$$

Dyskretyzacja: często lepsze, chociaż bardziej pracochłonne podejście.

Obsługa brakujących wartości

Tworzenie modelu: pomijanie brakujących wartości przy estymacji prawdopodobieństw $P(a_i = v_i|c = d)$,

Predykcja: pomijanie prawdopodobieństw $P(a_i = v_i|c = d)$ jeśli wartość a_i nie jest znana dla klasyfikowanego przykładu.

Właściwości naiwnego klasyfikatora bayesowskiego

- Prosty koncepcyjnie, implementacyjnie i obliczeniowo.
- Odporny na nadmierne dopasowanie.
- Niewymagający strojenia parametrów.
- Naruszenie założenia o niezależności nie wyklucza wartości predykcyjnej.
- Skuteczny jeśli:
  - trzeba uwzględnić nieznaczny wpływ znacznej liczby atrybutów (np. klasyfikacja tekstu),
  - liczba przykładów jest stosunkowo mała w porównaniu z liczbą atrybutów,
  - występują liczne brakujące wartości.