

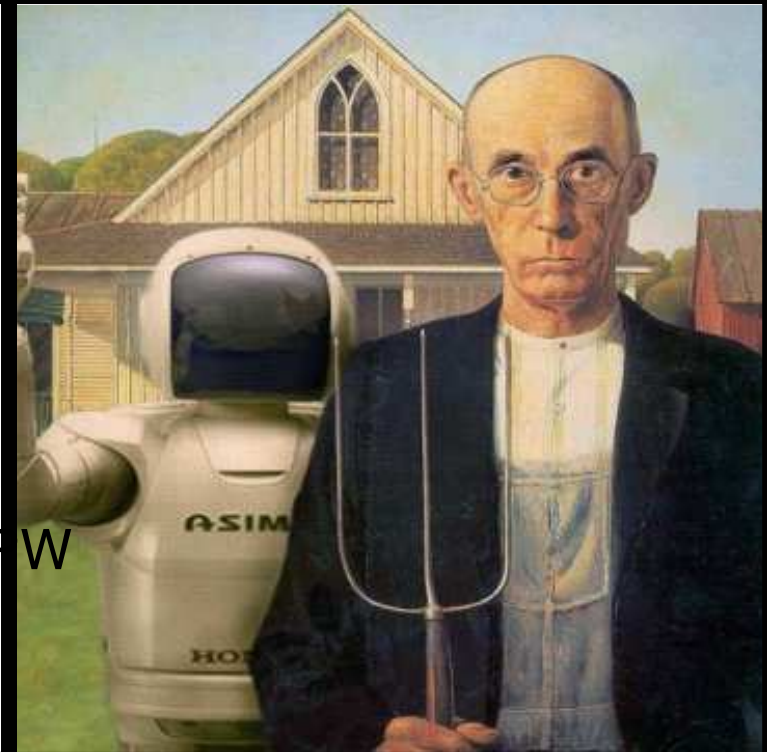
Sztuczna Inteligencja

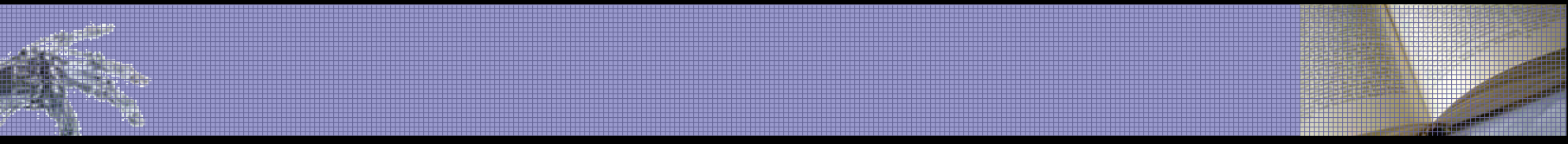
Wykład 5

Piotr Wąsiewicz

Zakład Sztucznej Inteligencji - ISE PW

pwasiewi@elka.pw.edu.pl





Konstrukcja drzew etykietujących



Zstępujące konstruowanie drzewa

funkcja *buduj-drzewo*(P, d, S)

argumenty wejściowe:

- P - zbiór przykładów etykietowanych pojęcia c ,
- d - domyślna etykieta kategorii,
- S - zbiór możliwych testów;

zwraca: drzewo decyzyjne jako hipotezę przybliżającą c na zbiorze P ;

jeśli *kryterium-stopu* (P, S) to

utwórz liść l ;

$d_l := \text{kategoria}(P, d)$;

zwróć l ;

koniec jeśli

utwórz węzeł n ;

$t_n := \text{wybierz-test}(P, S)$;

$d := \text{kategoria}(P, d)$;

dla wszystkich $r \in R_{t_n}$ wykonaj

$n[r] := \text{buduj-drzewo}(P_{t_n, r}, d, S - \{t_n\})$;

koniec dla

zwróć n

Kryterium stopu i wyboru kategorii

Kryterium stopu przyjmuje następującą postać:

$$P = \phi \vee S = \phi \vee |\{d' \in C | (\exists x \in P) \quad c(x) = d'\}| = 1$$

Operacja *wyboru kategorii* liścia natomiast taką:

$$\text{kategoria } (P, d) == \begin{cases} d & \text{jeśli } P = \phi, \\ \operatorname{argmax}_{d'} |P^{d'}| & \text{w przeciwnym przypadku} \end{cases}$$



Wybór testu dla największego przyrostu informacji

Wybór testu tworzącego węzeł lub liść zależy od przyrostu informacji $v_t(P)$ dla danego zbioru P i atrybutu t .

Informację zawartą w zbiorze etykietowanych przykładów P można wyrazić następująco:

$$I(P) = \sum_{d \in C} -\frac{|P^d|}{|P|} \log \frac{|P^d|}{|P|}$$

Z kolei *entropię* zbioru przykładów P ze względu na wynik r testu t określa się jako:

$$E_{tr}(P) = \sum_{d \in C} -\frac{|P_{tr}^d|}{|P_{tr}|} \log \frac{|P_{tr}^d|}{|P_{tr}|}$$
$$E_t(P) = \sum_{r \in R_t} \frac{|P_{tr}|}{|P|} E_{tr}(P)$$

Przyrost informacji wynikający z zastosowania testu t do zbioru przykładów etykietowanych P jest określony jako różnica:

$$g_t(P) = I(P) - E_t(P)$$

Jeśli *przyrost informacji* podzielimy przez *wartość informacyjną* $IV_t(P)$ testu t dla zbioru przykładów P , to otrzymamy *współczynnik przyrostu informacji* zdefiniowany jako:

$$v_t(P) = \frac{g_t(P)}{IV_t(P)}, \text{ gdzie } IV_t(P) = \sum_{r \in R_t} -\frac{|P_{tr}|}{|P|} \log \frac{|P_{tr}|}{|P|}$$



Zbiór trenujący T

x	aura	temperatura	wilgotność	wiatr	$c(x)$
1	słoneczna	ciepła	duża	słaby	0
2	słoneczna	ciepła	duża	silny	0
3	pochmurna	ciepła	duża	słaby	1
4	deszczowa	umiarkowana	duża	słaby	1
5	deszczowa	zimna	normalna	słaby	1
6	deszczowa	zimna	normalna	silny	0
7	pochmurna	zimna	normalna	silny	1
8	słoneczna	umiarkowana	duża	słaby	0
9	słoneczna	zimna	normalna	słaby	1
10	deszczowa	umiarkowana	normalna	słaby	1
11	słoneczna	umiarkowana	normalna	silny	1
12	pochmurna	umiarkowana	duża	silny	1
13	pochmurna	ciepła	normalna	słaby	1
14	deszczowa	umiarkowana	duża	silny	0

Współczynnik informacji

Obliczenia współczynnika przyrostu informacji dla testu tożsamościowego na wartościach atrybutu wilgotność.

$$\begin{aligned}|T^1| &= |\{3, 4, 5, 7, 9, 10, 11, 12, 13\}| = 9 \\|T^0| &= |\{1, 2, 6, 8, 14\}| = 5 \\|T_{wilgotno, normalna}| &= |\{5, 6, 7, 9, 10, 11, 13\}| = 7 \\|T_{wilgotno, normalna}^1| &= |\{5, 7, 9, 10, 11, 13\}| = 6\end{aligned}\tag{1}$$

$$\begin{aligned}|T_{wilgotno, normalna}^0| &= |\{6\}| = 1 \\|T_{wilgotno, duza}| &= |\{1, 2, 3, 4, 8, 12, 14\}| = 7 \\|T_{wilgotno, duza}^1| &= |\{3, 4, 12\}| = 3 \\|T_{wilgotno, duza}^0| &= |\{1, 2, 8, 14\}| = 4\end{aligned}\tag{2}$$

$$E_{wilgotno, normalna}(P) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0,592$$

$$E_{wilgotno, duza}(P) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0,985$$

$$I(T) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0,940$$

$$E_{wilgotno}(T) = \frac{7}{14} * 0,592 + \frac{7}{14} * 0,982 = 0,788$$

$$g_{wilgotno}(T) = I(T) - E_{wilgotno}(T) = 0,152$$

$$IV_{wilgotno}(T) = -\frac{7}{14} \log_2 \frac{7}{14} - \frac{7}{14} \log_2 \frac{7}{14} = 1$$

$$v_{wilgotno}(T) = \frac{g_{wilgotno}(T)}{IV_{wilgotno}(T)} = \frac{0,152}{1} = 0,152$$

Współczynnik przyrostu informacji

Obliczenia współczynnika przyrostu informacji dla testu tożsamościowego na wartościach atrybutu aura.

$$|T^1| = |\{3, 4, 5, 7, 9, 10, 11, 12, 13\}| = 9$$

$$|T^0| = |\{1, 2, 6, 8, 14\}| = 5$$

$$|T_{aura,soneczna}| = |\{1, 2, 8, 9, 11\}| = 5 \quad (3)$$

$$|T_{aura,soneczna}^1| = |\{9, 11\}| = 2$$

$$|T_{aura,soneczna}^0| = |\{1, 2, 8\}| = 3$$

$$|T_{aura,pochmurna}| = |\{3, 7, 12, 13\}| = 4 \quad (4)$$

$$|T_{aura,pochmurna}^1| = |\{3, 7, 12, 13\}| = 4$$

$$|T_{aura,pochmurna}^0| = |\{\phi\}| = 0$$

$$|T_{aura,deszczowa}| = |\{4, 5, 6, 10, 14\}| = 5 \quad (5)$$

$$|T_{aura,deszczowa}^1| = |\{4, 5, 10\}| = 3$$

$$|T_{aura,deszczowa}^0| = |\{6, 14\}| = 2$$

$$E_{aura,soneczna}(P) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0,971$$

$$E_{aura,pochmurna}(P) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$E_{aura,deszczowa}(P) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0,971$$

$$E_{aura}(T) = \frac{5}{14} * 0,971 + \frac{4}{14} * 0 + \frac{5}{14} * 0,971 = 0,694$$

$$g_{aura}(T) = I(T) - E_{aura}(T) = 0,940 - 0,694 = 0,246$$

Współczynnik przyrostu informacji

Obliczenia współczynnika przyrostu informacji dla testu tożsamościowego na wartościach atrybutu temperatura.

$$|T^1| = |\{3, 4, 5, 7, 9, 10, 11, 12, 13\}| = 9$$

$$|T^0| = |\{1, 2, 6, 8, 14\}| = 5$$

$$|T_{temp, ciepła}| = |\{1, 2, 3, 13\}| = 4 \quad (6)$$

$$|T_{temp, ciepła}^1| = |\{3, 13\}| = 2$$

$$|T_{temp, ciepła}^0| = |\{1, 2\}| = 2$$

$$|T_{temp, umiarkowana}| = |\{4, 8, 10, 11, 12, 14\}| = 6 \quad (7)$$

$$|T_{temp, umiarkowana}^1| = |\{4, 10, 11, 12\}| = 4$$

$$|T_{temp, umiarkowana}^0| = |\{8, 14\}| = 2$$

$$|T_{temp, zimna}| = |\{5, 6, 7, 9\}| = 4 \quad (8)$$

$$|T_{temp, zimna}^1| = |\{5, 7, 9\}| = 3$$

$$|T_{temp, zimna}^0| = |\{6\}| = 1$$

$$E_{temp, ciepła}(P) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

$$E_{temp, umiarkowana}(P) = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0,918$$

$$E_{temp, zimna}(P) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0,811$$

$$E_{temp}(T) = \frac{4}{14} * 1 + \frac{6}{14} * 0,918 + \frac{4}{14} * 0,811 = 0,911$$

$$g_{temp}(T) = I(T) - E_{temp}(T) = 0,940 - 0,911 = 0,029$$

Współczynnik przyrostu informacji

Obliczenia współczynnika przyrostu informacji dla testu tożsamościowego na wartościach atrybutu wiatr.

$$\begin{aligned} |T^1| &= |\{3, 4, 5, 7, 9, 10, 11, 12, 13\}| = 9 \\ |T^0| &= |\{1, 2, 6, 8, 14\}| = 5 \\ |T_{wiatr,saby}| &= |\{1, 3, 4, 5, 8, 9, 10, 13\}| = 8 \end{aligned} \quad (9)$$

$$\begin{aligned} |T_{wiatr,saby}^1| &= |\{3, 4, 5, 9, 10, 13\}| = 6 \\ |T_{wiatr,saby}^0| &= |\{1, 8\}| = 2 \\ |T_{wiatr,silny}| &= |\{2, 6, 7, 11, 12, 14\}| = 6 \\ |T_{wiatr,silny}^1| &= |\{7, 11, 12\}| = 3 \\ |T_{wiatr,silny}^0| &= |\{2, 6, 14\}| = 3 \end{aligned} \quad (10)$$

$$E_{wiatr,saby}(P) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0,811$$

$$E_{wiatr,silny}(P) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1$$

$$E_{wiatr}(T) = \frac{8}{14} * 0,811 + \frac{6}{14} * 1 = 0,892$$

$$g_{wiatr}(T) = I(T) - E_{wiatr}(T) = 0,940 - 0,892 = 0,048$$

$$IV_{wiatr}(T) = -\frac{8}{14} \log_2 \frac{8}{14} - \frac{6}{14} \log_2 \frac{6}{14} = 0,985$$

$$v_{wiatr}(T) = \frac{g_{wiatr}(T)}{IV_{wiatr}(T)} = \frac{0,048}{0,985} = 0,049$$

Kolejne kroki konstrukcji drzewa

1. Pierwsze wywołanie funkcji:
buduj-drzewo ($T, 1, \{aura, temperatura, wilgotność, wiatr\}$).
2. Kryterium stopu dla zbioru $P = T$ nie jest spełnione.
3. Tworzony jest nowy węzeł, dla którego na podstawie obliczonych wcześniej współczynników przyrostu informacji wybierany jest test tożsamościowy atrybutu *aura* o największym współczynniku.
4. Większościową etykietą w zbiorze P jest 1 i dalej jest przekazywana jako etykieta.
5. Następuje wywołanie rekurencyjne dla wyniku *słoneczna* testu *aura* :
 - *buduj-drzewo* ($P, 1, \{temperatura, wilgotność, wiatr\}$), gdzie $P = \{1, 2, 8, 9, 11\}$ i nie jest spełnione kryterium stopu.
 - Tworzony jest nowy węzeł dla którego wybierany jest test o najmniejszej entropii (w przypadku wątpliwości o największym współczynniku przyrostu informacji) tzn.: atrybut *wilgotność* :

$$E_{temp, zimna}(P) = -\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0$$

$$E_{temp, ciepła}(P) = -\frac{0}{2} \log_2 \frac{0}{2} - \frac{0}{2} \log_2 \frac{0}{2} = 0$$

$$E_{wilg, duża}(P) = -\frac{0}{3} \log_2 \frac{0}{3} - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

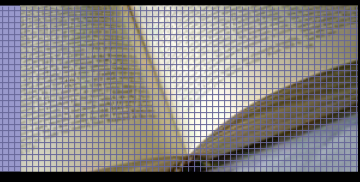
$$E_{wiatr, silny}(P) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$E_{temp, umiarkowana}(P) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$E_{wilg, normalna}(P) = -\frac{2}{2} \log_2 \frac{2}{2} - \frac{0}{2} \log_2 \frac{0}{2} = 0$$

$$E_{wiatr, słaby}(P) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0,918$$

$$E_{temp}(T) = 0,4; E_{wilgotno}(T) = 0; E_{wiatr}(T) = 0,951$$



3. Ciąg dalszy rekurencyjnego wykonania głównej funkcji dla wyniku *słoneczna* testu *aura* czyli punktu 5:
- Większościową etykietą kategorii w zbiorze P jest 0 i będzie ona przekazana dalej.
 - Dla wyniku *normalna* testu *wilgotność* następuje wykonanie rekurencyjne: *buduj-drzewo* ($P, 0, \{temperatura, wiatr\}$), gdzie $P = \{9, 11\}$ i jest spełnione kryterium stopu, gdyż zbiór P ma jedną etykietę 1. Jest tworzony liść z etykietą 1 i zwracany jako wynik funkcji.
 - Dla wyniku *duża* testu *wilgotność* następuje wykonanie rekurencyjne: *buduj-drzewo* ($P, 0, \{temperatura, wiatr\}$), gdzie $P = \{1, 2, 8\}$ i jest spełnione kryterium stopu, gdyż zbiór P ma jedną etykietę 0. Jest tworzony liść z etykietą 0 i zwracany jako wynik funkcji.
 - Zwracany jest jako wynik węzeł z testem *wilgotność*.
 - Następuje wywołanie rekurencyjne dla wyniku *pochmurna* testu *aura* dla $P = \{3, 7, 12, 13\}$ w wyniku czego powstaje liść z etykietą 1.
 - Następuje wywołanie rekurencyjne dla wyniku *deszczowa* testu *aura* dla $P = \{4, 5, 6, 10, 14\}$ w wyniku czego powstaje węzeł w testem *wiatr*, a następnie po dwóch rekurencyjnych wywołaniach powstają liście z etykietą 1 dla wyniku *słaby* przy czym $P = \{4, 5, 10\}$ oraz z etykietą 0 dla wyniku *silny* przy czym $P = \{6, 14\}$.



Skonstruowane drzewo decyzyjne

aura=słoneczna:

$$P = \{1, 2, 8, 9, 11\}$$

wilgotność=normalna:= 1

$$\text{dla } P = \{9, 11\}$$

wilgotność=duża:= 0

$$\text{dla } P = \{1, 2, 8\}$$

aura=pochmurna: 1

$$\text{dla } P = \{3, 7, 12, 13\}$$

aura=deszczowa:

$$P = \{4, 5, 6, 10, 14\}$$

wiatr=słaby:= 1

$$\text{dla } P = \{4, 5, 10\}$$

wiatr=silny:= 0

$$\text{dla } P = \{6, 14\}$$

