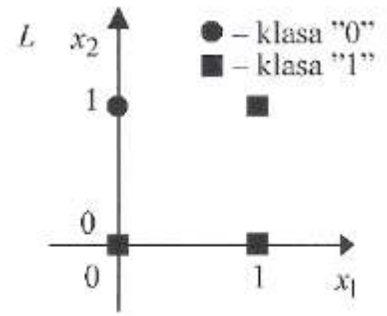


Systemy Inteligentne 2

Drzewa decyzyjne

Polecenie 5:

- podać analityczną postać zbioru L przedstawionego graficznie obok,
- podać zależności określające $E_t(P)$ oraz $E_{t=r}(P)$,
- przeprowadzić procedurę zstępującego konstruowania drzewa decyzyjnego dla zbioru L z wykorzystaniem testu tożsamościowego,
- podać ilustrację graficzną uzyskanego drzewa decyzyjnego,
- podać zbiorczą regułę określającą mechanizm wykrywania d . *



Ad a).

$$L = \{([x_1, x_2], d)^{(k)}\}_{k=1}^{K=4} = \{$$

$$([0,0], 1)^{(1)},$$

$$([1,0], 1)^{(2)},$$

$$([1,1], 1)^{(3)},$$

$$([0,1], 0)^{(4)}\}$$

Ad b).

- Entropia zbioru przykładów P ze względu na test „ t ”:

$$E_t(P) = \sum_{r \in R_t} \frac{|P_{t=r}|}{|P|} E_{t=r}(P)$$

- Entropia zbioru przykładów P ze względu na wynik „ r ” testu „ t ”:

$$E_{t=r}(P) = \sum_{d \in C} \left(-\frac{|P_{t=r}^d|}{|P_{t=r}|} \log_2 \frac{|P_{t=r}^d|}{|P_{t=r}|} \right)$$

Ad c).

KROK 1(1): Kryterium stopu:

- $P = L$
- $P \neq \emptyset$ i P nie zawiera rekordów należących wyłącznie do tej samej klasy
- Kryterium stopu niespełnione GOTO 1(2).

KROK 1(2): Utworzenie nowego węzła, wybór testu oraz realizacja rozgałęzień.

- $P: P = L, |P| = 4$
- $Z = \{tx_1, tx_2\}$
- $E_t(P)$ dla $t \in Z$

Test $tx_1, R_{tx1} = \{0,1\}$:

$$|P_{tx1=0}| = 2, |P_{tx1=0}^0| = 1, |P_{tx1=0}^1| = 1$$

$$|P_{tx1=1}| = 2, |P_{tx1=1}^0| = 0, |P_{tx1=1}^1| = 2$$

$$E_{tx1=0}(P) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$E_{tx1=1}(P) = -\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$E_{tx1}(P) = \frac{2}{4} * 1 + \frac{2}{4} * 0 = \frac{2}{4} = \frac{1}{2}$$

Test $tx_2, R_{tx_2} = \{0,1\}$:

$$|P_{tx_2=0}| = 2, |P_{tx_2=0}^0| = 0, |P_{tx_2=0}^1| = 2$$

$$|P_{tx_2=1}| = 2, |P_{tx_2=1}^0| = 1, |P_{tx_2=1}^1| = 1$$

$$E_{tx_2=0}(P) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$$E_{tx_2=1}(P) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$E_{tx_2}(P) = \frac{2}{4} * 0 + \frac{2}{4} * 1 = \frac{2}{4} = \frac{1}{2}$$

Wybór testu minimalizującego $E_t(P)$:

$$\hat{t} = \arg \min_{t \in Z} E_t(P) = \arg \min_{t \in Z} \{0.5, 0.5\} = tx_1$$

Realizacja rozgałęzień:

- dla $tx_1 = 0$ GOTO KROK 2(1)
- dla $tx_1 = 1$ GOTO KROK 3(1)

KROK 2(1): Kryterium stopu:

- $P: P = \{1,4\}$
- $P \neq \emptyset$ i P nie zawiera rekordów wyłącznie jednej klasy
- Kryterium stopu niespełnione GOTO KROK 2(2).

KROK 2(2): Utworzenie nowego węzła, wybór testu oraz realizacja rozgałęzień.

- $P: P = \{1,4\}, |P| = 2$
- $Z = \{tx_2\}$

Adnotacja: Wybór testu minimalizującego nie będzie nam już potrzebny, ponieważ jest tylko jeden test tx_2 .

Realizacja rozgałęzień:

- dla $tx_2 = 0$ GOTO KROK 4(1)
- dla $tx_2 = 1$ GOTO KROK 5(1)

KROK 3(1): Kryterium stopu:

- $P: P = \{2,3\}$
- $P \neq \emptyset$, ale P zawiera rekordy wyłącznie jednej klasy
- Kryterium stopu spełnione:
 - utwórz liść z etykietą klasy $y=1$
 - STOP

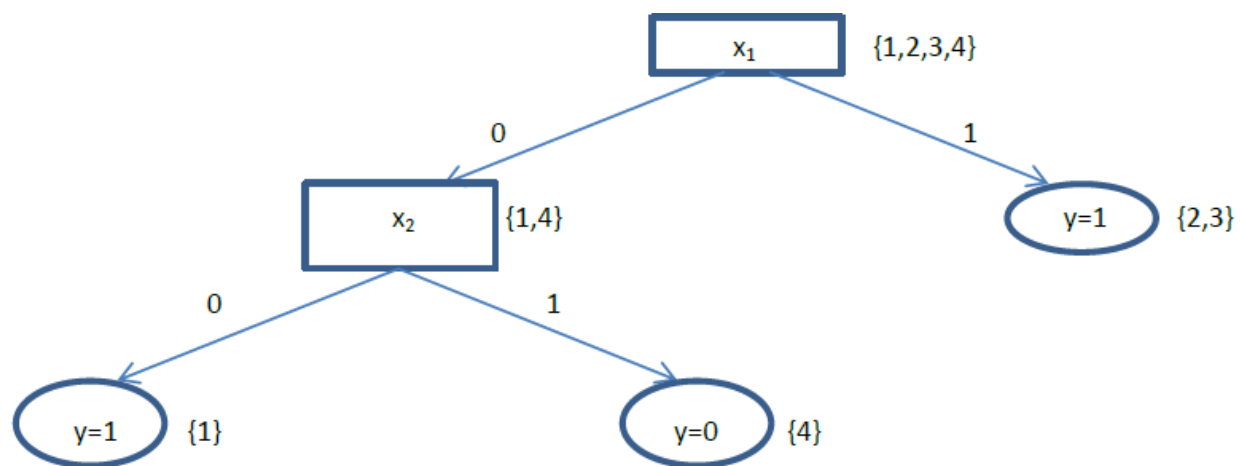
KROK 4(1): Kryterium stopu:

- $P: P = \{1\}$
- $P \neq \emptyset$, ale P zawiera rekordy wyłącznie jednej klasy
- Kryterium stopu spełnione:
 - utwórz liść z etykietą klasy $y=1$
 - STOP

KROK 5(1): Kryterium stopu:

- $P: P = \{4\}$
- $P \neq \emptyset$, ale P zawiera rekordy wyłącznie jednej klasy
- Kryterium stopu spełnione:
 - utwórz liść z etykietą klasy $y=0$
 - STOP

Ad d).

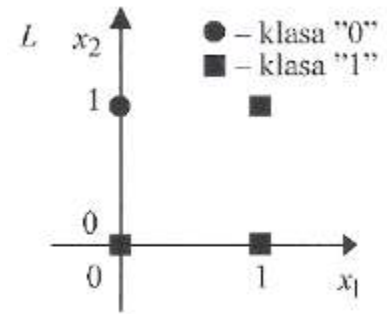


Ad e).

$JEŚLI \{[x_1 = 0] \text{ I } [x_2 = 1]\} \text{ TO } y = 0$
 $JEŚLI \{[x_1 = 1] \text{ LUB } \{[x_1 = 0] \text{ I } [x_2 = 0]\}\} \text{ TO } y = 1$

Polecenie 6:

- podać analityczną postać zbioru L przedstawionego graficznie obok,
- podać zależności określające $E_t(P)$ oraz $E_{t=r}(P)$,
- przeprowadzić procedurę zstępującego konstruowania drzewa decyzyjnego dla zbioru L z wykorzystaniem testu nierównościowy,
- podać ilustrację graficzną uzyskanego drzewa decyzyjnego.
- podać zbiorczą regułę określającą mechanizm wykrywania d . *



Ad a).

$$L = \{([x_1, x_2], d)^{(k)}\}_{k=1}^{K=4} = \{$$

$$([0,0], 1)^{(1)},$$

$$([1,0], 1)^{(2)},$$

$$([1,1], 1)^{(3)},$$

$$([0,1], 0)^{(4)}\}$$

Ad b).

- Entropia zbioru przykładów P ze względu na test „t”:

$$E_t(P) = \sum_{r \in R_t} \frac{|P_{t=r}|}{|P|} E_{t=r}(P)$$

- Entropia zbioru przykładów P ze względu na wynik „r” testu „t”:

$$E_{t=r}(P) = \sum_{d \in C} \left(-\frac{|P_{t=r}^d|}{|P_{t=r}|} \log_2 \frac{|P_{t=r}^d|}{|P_{t=r}|} \right)$$

Ad c).

KROK 1(1): Kryterium stopu:

- $P = L$
- $P \neq \emptyset$ i P nie zawiera rekordów należących wyłącznie do tej samej klasy
- Kryterium stopu niespełnione GOTO 1(2).

KROK 1(2): Utworzenie nowego węzła, wybór testu oraz realizacja rozgałęzień.

- $P: P = L \mid |P| = 4$
- $Z = \{nx_1, nx_2\}$
- $E_t(P)$ dla $t \in Z$

Test nx_1 :

$$nx_1(\theta) = \begin{cases} 0, & \text{gdy } x_1 \leq \theta \\ 1, & \text{gdy } x_1 > \theta \end{cases}$$

k:	4	1	2	3
x_1 :	0	0	1	1
d:	0	1	1	1

↑

$$\theta = \{0\}$$

$$|P_{nx_1=0}| = 2 \quad |P_{nx_1=0}^{(0)}| = 1 \quad |P_{nx_1=0}^{(1)}| = 1$$

$$|P_{nx_1=1}| = 2 \quad |P_{nx_1=1}^{(0)}| = 0 \quad |P_{nx_1=1}^{(1)}| = 2$$

$$E_{nx_1=0}(P) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$E_{nx_1=1}(P) = -\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$E_{nx_1}(P) = \frac{2}{4} * 1 + \frac{2}{4} * 0 = \frac{2}{4} = \frac{1}{2}$$

Test nx_2 :

$$nx_2(\theta) = \begin{cases} 0, & \text{gdy } x_2 \leq \theta \\ 1, & \text{gdy } x_2 > \theta \end{cases}$$

k:	1	2	3	4
x_2 :	0	0	1	1
d:	1	1	1	0

↑

$$\theta = \{1\}$$

$$|P_{nx_2=0}| = 4 \quad |P_{nx_2=0}^{(0)}| = 1 \quad |P_{nx_2=0}^{(1)}| = 3$$

$$|P_{nx_2=1}| = 0 \quad |P_{nx_2=1}^{(0)}| = 0 \quad |P_{nx_2=1}^{(1)}| = 0$$

$$E_{nx_2=0}(P) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.5 + 0.32 = 0.82$$

$$E_{nx_2=1}(P) = 0$$

$$E_{nx_2}(P) = \frac{4}{4} * 0.82 = 1 * 0.82 = 0.82$$

Wybór testu minimalizującego $E_{nx_1(\theta)}(P)$:

$$\hat{\theta} = \arg \min_{\theta} E_{nx_1(\theta)}(P) = \arg \min_{\theta} \{0.5, 0.82\} = 0$$

Wybór testu minimalizującego $E_t(P)$:

$$\hat{t} = \arg \min_{t \in Z} E_t(P) = \arg \min_{t \in Z} \{0.5\} = nx_1$$

Realizacja rozgałęzień:

- dla $nx_1(\hat{\theta}) = 0$, czyli $x_1 \leq 0$ GOTO KROK 2(1)
- dla $nx_1(\hat{\theta}) = 1$, czyli $x_1 > 0$ GOTO KROK 3(1)

KROK 2(1): Kryterium stopu:

- $P: P = \{1, 4\}$
- $P \neq \emptyset$ i P nie zawiera rekordów należących wyłącznie do tej samej klasy
- Kryterium stopu niespełnione GOTO 2(2).

KROK 2(2): Utworzenie nowego węzła, wybór testu oraz realizacja rozgałęzień.

- $P: P = \{1, 4\}, |P| = 2$
- $Z = \{nx_2\}$
- $E_t(P)$ dla $t \in Z$

Test nx_2 :

$$nx_2(\theta) = \begin{cases} 0, & \text{gdy } x_2 \leq \theta \\ 1, & \text{gdy } x_2 > \theta \end{cases}$$

k:	1	4
x_2 :	0	1
d:	1	0

↑

$$\theta = \{0.5\}$$

$$|P_{nx_2=0}| = 1 \quad |P_{nx_2=0}^{(0)}| = 0 \quad |P_{nx_2=0}^{(1)}| = 1$$

$$|P_{nx_2=1}| = 1 \quad |P_{nx_2=1}^{(0)}| = 1 \quad |P_{nx_2=1}^{(1)}| = 0$$

$$E_{nx_2=0}(P) = 0$$

$$E_{nx_2=1}(P) = 0$$

$$E_{nx_2}(P) = 0$$

Adnotacja: Wybór testu minimalizującego nie będzie nam już potrzebny, ponieważ jest tylko jeden test nx_2 .

Realizacja rozgałęzień:

- dla $nx_2(\hat{\theta}) = 0$, czyli $x_2 \leq 0$ GOTO KROK 4(1)
- dla $nx_2(\hat{\theta}) = 1$, czyli $x_2 > 0$ GOTO KROK 5(1)

KROK 3(1): Kryterium stopu:

- $P: P = \{2,3\}$
- $P \neq \emptyset$ ale P zawiera rekordy wyłącznie jednej klasy.
- Kryterium stopu spełnione:
 - utwórz liść z etykietą klasy $y=1$
 - STOP

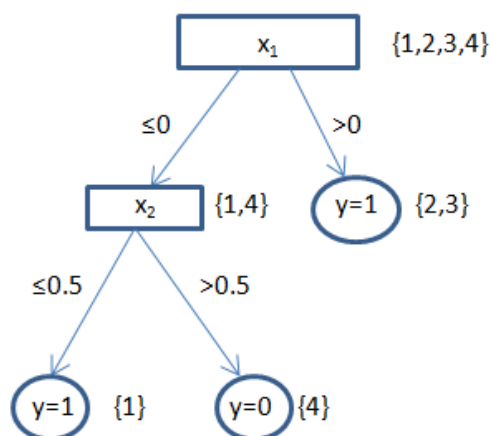
KROK 4(1): Kryterium stopu:

- $P: P = \{1\}$
- $P \neq \emptyset$ ale P zawiera rekordy wyłącznie jednej klasy.
- Kryterium stopu spełnione:
 - utwórz liść z etykietą klasy $y=1$
 - STOP

KROK 5(1): Kryterium stopu:

- $P: P = \{4\}$
- $P \neq \emptyset$ ale P zawiera rekordy wyłącznie jednej klasy.
- Kryterium stopu spełnione:
 - utwórz liść z etykietą klasy $y=0$
 - STOP

Ad d).



Ad e).

$$\begin{aligned}
 & \text{JEŻELI } \{[x_1 \leq 0] \text{ I } [x_2 > 0.5]\} \text{ TO } y = 0 \\
 & \text{JEŻELI } \{[x_1 > 0] \text{ LUB } \{[x_1 \leq 0] \text{ I } [x_2 \leq 0.5]\}\} \text{ TO } y = 1
 \end{aligned}$$

Opracowano dn. 8.12.2017
Bartłomiej Osak, Tomasz Odzimek