Data Mining
Wykład 4

Indukcja drzew decyzyjnych -
Indeks Gini & Zysk informacyjny

Indeks Gini

- Popularnym kryterium podziału, stosowanym w wielu produktach komercyjnych, jest indeks Gini
- Algorytm SPRINT (IBM Inteligent Miner)
- Rozważmy przykładowy zbiór treningowy, w którym każdy rekord opisuje ocenę ryzyka, że osoba, która ubezpieczyła samochód, spowoduje wypadek. Ocena została dokonana przez firmę ubezpieczającą w oparciu o dotychczasową historię ubezpieczonego

Indeks Gini - Przykład

Zbiór treningowy

ID	Wiek	Typ_sam	Ryzyko
0	23	family	high
1	17	sport	high
2	43	sport	high
3	68	family	low
4	32	truck	low
5	20	family	high

Model może być wykorzystany do oceny ryzyka związanego z ubezpieczeniem nowego klienta

Klasyfikacja - Założenia

```
Partition(Data S) {
        if (all points in S are of the same class) then
                return;
        for each attribute A do
                evaluate splits on attribute A;

        Use best split found to partition S into S1 and S2

        Partition(S1);
        Partition(S2);
}

Initial call: Partition(Training Data)
```


Definicja 1: Wartość Indeksu Gini

$$gini(S) = 1 - \sum p_j^2$$

- gdzie:
  - **S** – zbiór przykładów należących do n klas
  - p_j – względna częstość występowania klasy j w S
- Przykładowo:

  dwie klasy, Pos i Neg, oraz zbiór przykładów S zawierający p elementów należących do klasy Pos i n elementów należących do klasy Neg

$$p_{pos} = p/(p+n) \qquad p_{neg} = n/(n+p)$$
$$gini(S) = 1 - p_{pos}^2 - p_{neg}^2$$

Definicja 2: Indeks podziału Gini

- Punkt podziału dzieli zbiór S na dwie partycje S1 i S2 – indeks podziału Gini jest zdefiniowany następująco:

$$gini_{SPLIT}(S) = (p_1 + n_1)/(p+n)*gini(S_1) + (p_2 + n_2)/(p+n)* gini(S_2)$$

gdzie p_1, n_1 (p_2, n_2) oznaczają, odpowiednio,
- p_1 - elementów w S_1 należących do klasy Pos,
- n_1 - liczba elementów w S_1 należących do klasy Neg,
- p_2 - elementów w S_2 nalężących do klasy Pos,
- n_2 - liczba elementów w S_2 należących do klasy Neg

Indeks Gini

- „Najlepszym" punktem podziału zbioru S jest punkt podziału, który charakteryzuje się **najmniejszą wartością indeksu podziału Gini $gini_{SPLIT}$**
- Dla każdego atrybutu, dla wszystkich możliwych punktów podziału, oblicz wartość indeksu podziału Gini – **wybierz punkt podziału o najmniejszej wartości $gini_{SPLIT}$**
- Wybrany punkt podziału **włącz do drzewa decyzyjnego**
- Punkt podziału dzieli zbiór S na **dwie partycje S1 i S2**.
- **Powtórz procedurę** obliczania indeksu podziału dla partycji S1 i S2 – znalezione punkty podziału włącz do drzewa decyzyjnego.
- **Powtarzaj procedurę** dla kolejnych partycji aż do osiągnięcia warunku stopu

Przykład (1)

ID	Wiek	Typ_sam	Ryzyko
0	23	family	high
1	17	sport	high
2	43	sport	high
3	68	family	low
4	32	truck	low
5	20	family	high

Przykład (2)

Lista wartości atrybutu Wiek

Wiek	ID	Ryzyko
23	0	high
17	1	high
43	2	high
68	3	low
32	4	low
20	5	high

Lista wartości atrybutu Typ_sam

Typ_sam	ID	Ryzyko
family	0	high
sport	1	high
sport	2	high
family	3	low
truck	4	low
family	5	high

Przykład (3)

- Możliwe punkty podziału dla atrybutu **Wiek**:
  Wiek ≤17, Wiek ≤ 20, Wiek ≤ 23,
  Wiek ≤ 32,Wiek ≤ 43, Wiek ≤ 68

klasa

Liczba krotek	High	Low
Wiek ≤ 17	1	0
Wiek > 17	3	2

gini(Wiek <= 17) = 1- (1^2+0^2) = 0
gini(Wiek > 17) = 1- $((3/5)^2+(2/5)^2)$ = 0,73
$gini_{SPLIT}$ = (1/6) * 0 + (5/6) * (0,73) = **0,6**

Przykład (4)

Liczba krotek	High	Low
Wiek <= 20	2	0
Wiek > 20	2	2

gini(Wiek <= 20) = 1- (1^2+0^2) = 0
gini(Wiek >20) = 1- $((1/2)^2+(1/2)^2)$ = 1/2
$gini_{SPLIT}$ = (2/6) * 0 + (4/6) * (1/8) = **1/3 = 0.33**

Liczba krotek	High	Low
Wiek <= 23	3	0
Wiek > 23	1	2

gini(Wiek <= 23) = 1- (12+02) = 0
gini(Wiek > 23) = 1- ((1/3)2+(2/3)2) = 4/9
$gini_{SPLIT}$ = (3/6) * 0 + (3/6) * (4/9) = **2/9 = 0.22**

Przykład (5)

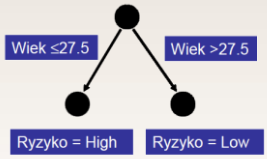
Liczba krotek	High	Low
Wiek <= 32	3	1
Wiek > 32	1	1

gini(Wiek ≤ 32) = 1- $((3/4)^2+(1/4)^2)$ = 3/8
gini(Wiek > 32) = 1- $((1/2)^2+(1/2)^2)$ = 1/2
$gini_{SPLIT}$ = (4/6) * (3/8) + (2/6) * (1/2) = **7/24 = 0.29**

Najmniejsza wartość indeksu podziału $gini_{SPLIT}$ posiada punkt podziału Wiek ≤ 23, stad, tworzymy wierzchołek drzewa decyzyjnego dla punktu podziału Wiek = (23+32) / 2 = **27.5**

Przykład (6)

- Drzewo decyzyjne, po pierwszym podziale zbioru treningowego, ma następująca postać:



- Zauważmy, że pierwsza partycja S1 jest partycją „czystą”, w tym sensie, że wszystkie rekordy należące do tej partycji należą do jednej klasy.

Przykład (7)

- Listę wartości atrybutów dzielimy w punkcie podziału:
  Listy wartości atrybutów dla Wiek >=27.5:

Wiek	ID	Ryzyko
17	1	high
20	5	high
23	0	high

Typ_sam	ID	Ryzyko
family	0	high
sport	1	high
family	5	high

- Listy wartości atrybutów dla Wiek > 27.5:

Wiek	ID	Ryzyko
32	4	low
43	2	high
68	3	low

Typ_sam	ID	Ryzyko
sport	2	high
family	3	low
truck	4	low

Przykład (8)

- **Ocena punktów podziału dla atrybutu kategorycznego**

Musimy dokonać oceny wszystkich punktów podziału atrybutu kategorycznego - 2^N kombinacji, gdzie N oznacza liczbę wartości atrybutu kategorycznego

Liczba krotek	high	low
Typ_sam={sport}	1	0
Typ_sam={family}	0	1
Typ_sam={truck}	0	1

$gini(Typ\_sam \in \{sport\}) = 1 - 1^2 - 0^2 = 0$
$gini(Typ\_sam \in \{family\}) = 1 - 0^2 - 1^2 = 0$
$gini(Typ\_sam \in \{truck\}) = 1 - 0^2 - 1^2 = 0$

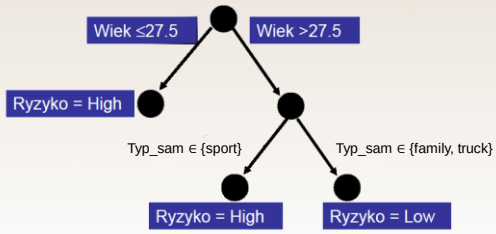
Przykład (9)

- gini(Typ_sam ∈ { sport, family }) = 1 - $(1/2)^2$ - $(1/2)^2$ = 1/2
- gini(Typ_sam ∈ { sport, truck }) = 1 - $(1/2)^2$ - $(1/2)^2$ = 1/2
- gini(Typ_sam ∈ { family, truck }) = 1 - 0^2 - 1^2 = 0

- $gini_{SPLIT}$(Typ_sam ∈ { sport }) = (1/3) * 0 + (2/3) * 0 = 0
- $gini_{SPLIT}$(Typ_sam ∈ { family }) = (1/3) * 0 + (2/3)*(1/2) = 1/3
- $gini_{SPLIT}$(Typ_sam ∈ { truck }) = (1/3) * 0 + (2/3)*(1/2) = 1/3
- $gini_{SPLIT}$(Typ_sam ∈ { sport, family}) = (2/3)*(1/2)+(1/3)*0= 1/3
- $gini_{SPLIT}$(Typ_sam ∈ { sport, truck}) = (2/3)*(1/2)+(1/3)*0= 1/3
- $gini_{SPLIT}$(Typ_sam ∈ { family, truck }) = (2/3)*0+(1/3)*0 = 0

Przykład (10)

- Drzewo decyzyjne po wprowadzeniu wierzchołka ma postać:



Zysk informacyjny

- Do wyboru atrybutu testowego w wierzchołku drzewa decyzyjnego wykorzystujemy miarę **zysku informacyjnego**
- Jako atrybut testowy (aktualny wierzchołek drzewa decyzyjnego) wybieramy **atrybut o największym zysku informacyjnym** (lub największej redukcji entropii)
- Atrybut testowy m**inimalizuje ilość informacji niezbędnej do klasyfikacji przy**kładów w partycjach uzyskanych w wyniku podziału

Zysk informacyjny – Oczekiwana ilość informacji (1)

- Niech S oznacza zbiór s przykładów. Załóżmy, że atrybut decyzyjny posiada m różnych wartości definiujących m klas, C_i (dla i=1, ..., m)
- Niech s_i oznacza liczbę przykładów zbioru S należących do klasy C_i
- Oczekiwana ilość informacji niezbędna do zaklasyfikowania danego przykładu:

$$I(s_1,\ldots,s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Zysk informacyjny – Oczekiwana ilość informacji (2)

- p_i oznacza prawdopodobieństwo, że dowolny przykład należy do klasy Ci (oszacowanie - s_i/s)
- Niech atrybut A posiada v różnych wartości:

  $\{a_1, a_2, \ldots, a_v\}$

  Atrybut A dzieli zbiór S na partycje $\{S_1, S_2, \ldots, S_v\}$, gdzie S_j zawiera przykłady ze zbioru S, których wartość atrybutu A wynosi a_j
- Wybierając atrybut A jako atrybut testowy tworzymy wierzchołek drzewa, którego łuki wychodzące posiadają etykiety $\{a_1, a_2, \ldots, a_v\}$ i łączą dany wierzchołek A z wierzchołkami zawierającymi partycje $\{S_1, S_2, \ldots, S_v\}$

Zysk informacyjny – Entropia

- Niech s_{ij} oznacza liczbę przykładów z klasy C_i w partycji S_j. Entropie podziału zbioru S na partycje, według atrybutu A definiujemy następująco:

$$E(A_1, A_2, \ldots A_v) = \sum_{j=1}^{v} \frac{(s_{1j} + s_{2j} + \ldots + s_{mj})}{s} I(s_{1j}, s_{2j}, \ldots, s_{mj})$$

Im mniejsza wartość entropii, tym większa „czystość" podziału zbioru S na partycje

Zysk informacyjny – waga j-tej partycji

- Współczynnik (s_{1j} + s_{2j} +...+s_{mj})/s stanowi wage j-tej partycji i zdefiniowany jest jako iloraz liczby przykładów w j-tej partycji (i.e. krotek posiadających wartość a_j atrybutu A) do całkowitej liczby przykładów w zbiorze S. Zauważmy, że dla danej partycji S_j,

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij})$$

- gdzie $p_{ij} = s_{ij}/|S_j|$ i określa prawdopodobieństwo, że przykład z S_j należy do klasy Ci

Zysk informacyjny – Gain(A)

- **Zysk informacyjny,** wynikający z podziału zbioru S na partycje według atrybutu A, definiujemy następująco:

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

- Gain(A) oznacza oczekiwana redukcje entropii (nieuporządkowania) spowodowana znajomością wartości atrybutu A

Przykład (1)

ID	wiek	dochód	student	status	kupi_komputer
1	<=30	wysoki	nie	kawaler	nie
2	<=30	wysoki	nie	żonaty	nie
3	31..40	wysoki	nie	kawaler	tak
4	>40	średni	nie	kawaler	tak
5	>40	niski	tak	kawaler	tak
6	>40	niski	tak	żonaty	nie
7	31..40	niski	tak	żonaty	tak
8	<=30	średni	nie	kawaler	nie
9	<=30	niski	tak	kawaler	tak
10	>40	średni	tak	kawaler	tak
11	<=30	średni	tak	żonaty	tak
12	31..40	średni	nie	żonaty	tak
13	31..40	wysoki	tak	kawaler	tak
14	>40	średni	nie	żonaty	nie

Przykład (2)

- Rozważmy przedstawiony zbiór treningowy opisujący klientów sklepu komputerowego
- Atrybut decyzyjny, **„kupi_komputer”**, posiada dwie wartości (tak, nie), stad, wyróżniamy dwie klasy (m=2)
  - C1 odpowiada wartości **tak** - s1 = 9
  - C2 odpowiada wartości **nie** - s2 = 5

$$I(s_1,s_2)=I(9,5)=-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}=0.94$$

Przykład (3)

- Następnie, obliczamy entropie każdego deskryptora Rozpocznijmy od atrybutu wiek:

dla wiek =‘<=30’
s_{11}=2 s_{21}=3 I(s_{11}, s_{21}) = 0.971

dla wiek =’31..40’
s_{12}=4 s_{22}=0 I(s_{12}, s_{22}) = 0

dla wiek =‘>40’
s_{13}=2 s_{23}=3 I(s_{13}, s_{23}) = 0.971

Przykład (4)

- Entropia atrybutu „wiek” wynosi:

E(„wiek”)=5/14 *I(s_{11}, s_{21}) +4/14* I(s_{12}, $s_2$2) + + 5/14* I(s_{13}, s_{23}) = 0.694

- Zysk informacyjny wynikający z podziału zbioru S według atrybutu wiek wynosi:

Gain(„wiek”) = I(s_1, s_2) – E(„wiek”) = 0.246

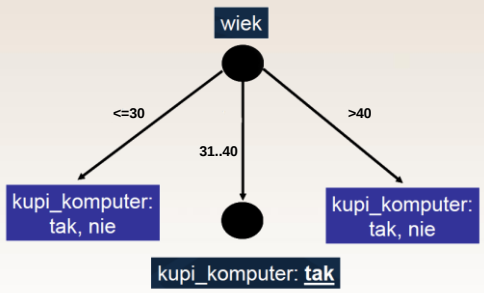
Przykład (5)

- Analogicznie obliczamy zysk informacyjny dla pozostałych atrybutów:

  Gain(*dochód*) = 0.029
  Gain(*student*) = 0.151
  Gain(*status*) = 0.048

- Ponieważ **„wiek" daje największy zysk informacyjny** spośród wszystkich deskryptorów, atrybut ten jest wybierany jako pierwszy atrybut testowy
- Tworzymy wierzchołek drzewa o etykiecie „wiek", oraz etykietowane łuki wychodzące, łączące wierzchołek „wiek" z wierzchołkami odpowiadającymi partycjom zbioru utworzonymi według atrybutu *„wiek"*

Przykład (6)



Przykład (7)