Techniki eksploracji danych

Krzysztof Gajowniczek

Rok akademicki: 2020/2021

Section 1

Zmienne ciągłe

- Aby obsłużyć ciągłe zmienne, algorytm tworzy próg s, a następnie dzieli wartości zmiennej na te, które są powyżej progu $x > s$ i te, które są od niego mniejsze lub równe $x \leq s$.
- Aby znaleźć najlepszy punkt podziału, algorytm musi wykonać $s - 1$ operacji, gdzie s jest liczbą unikalnych wartości danej zmiennej.

```
SplitNum <- function( Y, x, param ){
  spilts <- mozliwe_podzialy
  wyniki <- matrix(0, length(spilts), kolumny)
  for( i in 1:length(spilts) ){
    lewy <- x <= spilts[i]
    prawy <- x > spilts[i]
    wyniki[i,] <- FunkcjaWyniki(lewy,prawy)
  }
  return(wyniki)
}
```

Section 2

Zmienne nominalne

- Uporządkowane zmienne jakościowe (porządkowe) można traktować tak samo, jak zmienne ciągłe.
- Niestety, w przypadku nieuporządkowanych zmiennych jakościowych (nominalnych), standardowym podejściem jest rozważenie wszystkich 2 elementowych podziałów c kategorii/poziomów zmiennej.
- Każdy z tych $2^{c-1} - 1$ podziałów jest poddawany ocenie i wybierany jest najlepszy podział.

- Na całe szczęście dla problemu klasyfikacji binarnej oraz dla problemu regresyjnego istnieje heurystyka ograniczająca złożoność obliczeniową.
- Dla klasyfikacji binarnej, kategorie zmiennej nominalnej (c) sortuje się według proporcji przypadających klasie pozytywnej dla zmiennej celu.
- Dla regresji, kategorie zmiennej nominalnej (c) sortuje się według wartości średniej zmiennej celu.

```
c(kat1 = 0.5, kat2 = 0.1, kat3 = 0.4)
```

```
## kat1 kat2 kat3
##  0.5  0.1  0.4
```

```
sort(c(kat1 = 0.5, kat2 = 0.1, kat3 = 0.4))
```

```
## kat2 kat3 kat1
##  0.1  0.4  0.5
```

- Dla tak przygotowanych kategorii stosuje się algorytm dla zmiennych ciągłych.
- Można wykazać, że daje to optymalny podział dla entropii, indeksu Giniego i sumy kwadratów różnic.

Section 3

Literatura

- *Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.*