Techniki eksploracji danych

Krzysztof Gajowniczek

Rok akademicki: 2021/2022

Section 1

Wybór optymalnego podziału

Subsection 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)
}
```

Subsection 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 2

Przycinanie drzewa

- Nadmierne dopasowanie jest istotną praktyczną trudnością w przypadku modeli drzew decyzyjnych i wielu innych modeli predykcyjnych.
- Do nadmiernego dopasowania dochodzi, gdy algorytm uczący się nadal rozwija reguły decyzyjne, które zmniejszają błąd zbioru uczącego kosztem zwiększonego błędu na zbiorze testowym.
- Istnieje kilka podejść do uniknięcia nadmiernego dopasowania w drzewach decyzyjnych:

Pre-pruning

Przycinanie zstępujące (wczesne zatrzymanie), powoduje, że drzewo wcześniej przestaje rosnąć, zanim doskonale sklasyfikuje zbior treningowy.

Post-pruning

Przycinanie wstępujące, które pozwala drzewu idealnie sklasyfikować zbior treningowy, a następnie przyciąć drzewo.

Pre-pruning

- **Maksymalna głębokość**: Ten parametr służy do ustawiania maksymalnej głębokości drzewa. Głębokość to długość najdłuższej ścieżki od węzła głównego do węzła liścia. Ustawienie tego parametru spowoduje zatrzymanie wzrostu drzewa, gdy głębokość będzie równa danej wartości.
- **Minimalna liczba obserwacji w węźle**: Jest to minimalna liczba obserwacji, które muszą istnieć w węźle, aby nastąpił podział lub próba jego wykonania. Na przykład ustawiając minimalną liczbę obserwacji w podziale na 5, węzeł można dalej podzielić w celu uzyskania jednorodności, gdy liczba obserwacji w każdym podzielonym węźle jest większa niż 5.

- **Minimalna liczba obserwacji w liściu**: Jest to minimalna liczba obserwacji, które mogą znajdować się w liściu. Ustawiając minimalną liczbę obserwacji w lisicu na 5, powoduje się, że każdy liść powinien mieć co najmniej pięć obserwacji. Trzeba zadbać o to, aby model nie został nadmiernie dopasowany poprzez podanie tego parametru. Jest on częściej wykorzystywany niż parametr **minimalna liczba obserwacji w węźle**.
- **Złożoność drzewa**: Nie podejmuje się prób podziału, który nie zmniejsza ogólnego braku dopasowania o współczynnik zakładaną wartość (parametr cp w algorytmie `CART` w bibliotece `rpart`). Na przykład w przypadku problemu regresji oznacza to, że całkowity R^2 musi rosnąć o cp na każdym kroku. Główną rolą tego parametru jest oszczędność czasu obliczeniowego poprzez przycinanie podziałów, które oczywiście nie są opłacalne.

Post-pruning

- **Reduced Error Pruning** (REP): Ta metoda traktuje każdy z węzłów decyzyjnych w drzewie jako kandydatów do przycinania, polega na usunięciu poddrzewa zakorzenionego w tym węźle, czyniąc go węzłem liścia. Dostępne dane są podzielone na trzy części: zbior uczący, zbior walidacyjny użyty do przycinania drzewa, oraz zestaw obserwacji testowych służących do zapewnienia obiektywnej oceny dokładności w stosunku do przyszłych niewidocznych obserwacji (algorytm `C5.0` w bibliotece `C50`).

- **Pessimistic Error Pruning** (PEP): Opiera się na obserwacji, że błąd klasyfikacji uzyskany przez drzewo przy użyciu zbioru treningowego jest nadmiernie optymistyczny. Rozmiar drzewa jest zmniejszany, gdy poprawiona liczba błędnych klasyfikacji uzyskanych przez przycięte drzewo jest większa niż poprawiony błąd przed przycięciem powiększony o błąd standardowy.

Pessimistic Error

$$PE^{(Node)} = \frac{\overline{y}^{(Node)} + \frac{z^2}{2n^{(Node)}} + z\sqrt{\frac{\overline{y}^{(Node)}}{n^{(Node)}} - \frac{(y^{(Node)})^2}{n^{(Node)}} + \frac{z^2}{4(n^{(Node)})^2}}}{1 + \frac{z^2}{n^{(Node)}}}$$

gdzie z^2 jest kwantylem rozkładu normalnego dla przyjętego przedziału ufności.
Pod-drzewo jest przycięte gdy:

$$PE^{(Parent)} \leq \frac{n^{(Left)}}{n^{(Parent)}} * PE^{(Left)} + \frac{n^{(Right)}}{n^{(Parent)}} * PE^{(Right)}$$

- Minimum Error Pruning (MEP),
- Critical Value Pruning (CVP).

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.*