

Techniki eksploracji danych

Krzysztof Gajowniczek

Rok akademicki: 2020/2021

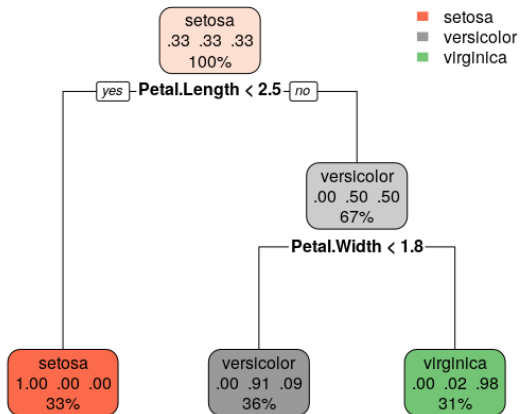
- 1 Drzewa klasyfikacyjne i regresyjne
- 2 Miary niejednorodności węzłów/liści
- 3 Rekurencyjny podział przestrzeni
- 4 Wybór optymalnego podziału
- 5 Literatura

Section 1

Drzewa klasyfikacyjne i regresyjne

- Drzewo decyzyjne jest jednym z szeroko stosowanych algorytmów w uczeniu maszynowym, zapewniającym solidną podstawę dla innych podejść.
- Podstawowym celem korzystania z drzewa decyzyjnego jest stworzenie modelu, który może przewidzieć docelową klasę lub wartość zmiennej poprzez naukę prostych reguł podejmowania decyzji wywnioskowanych z wcześniejszych danych (danych szkoleniowych).
- Używa wykresu przypominającego drzewo, aby pokazać prognozy wynikające z serii podziałów na podstawie cech.

- Jednym ze sposobów myślenia o drzewie decyzyjnym jest seria węzłów lub wykres kierunkowy, który zaczyna się od pojedynczego węzła u podstawy i rozciąga się na wiele węzłów liści reprezentujących kategorie, które drzewo może klasyfikować.
- Każdy węzeł w drzewie określa test atrybutu instancji.
- Każda gałąź wychodząca z węzła odpowiada jednej z możliwych wartości atrybutu.
- Każdy węzeł na liściu przypisuje wartość przewidywaną.



Zalety

- **Moc wyjaśniająca** - łatwe do wyjaśnienia i zinterpretowania, dane wyjściowe drzew decyzyjnych są łatwe do interpretacji. Może być zrozumiany przez każdego bez wiedzy analitycznej, matematycznej lub statystycznej.
- **Eksploracyjna analiza danych** - drzewa decyzyjne pozwalają analitykom szybko zidentyfikować istotne zmienne i istotne relacje między dwiema lub więcej zmiennymi, pomagając w ten sposób ujawnić sygnał, który zawiera wiele zmiennych wejściowych.
- **Minimalne czyszczenie danych** - ponieważ drzewa decyzyjne są odporne na wartości odstające i brakujące wartości, wymagają mniej czyszczenia danych niż inne algorytmy.

Zalety

- **Wszystkie typy danych** - drzewa decyzyjne mogą dokonywać klasyfikacji na podstawie zarówno zmiennych numerycznych, jak i kategoryalnych.
- **Nieparametryczne** - drzewo decyzyjne jest nieparametrycznym algorytmem, w przeciwieństwie np. do regresji logistycznej czy sieci neuronowych, które przetwarzają dane wejściowe przekształcone w tensor, używając dużej liczby współczynników, zwanych parametrami, poprzez mnożenie tensorów.

Wady

- **Przeuczenie** - częstym błędem w drzewach decyzyjnych jest nadmierne dopasowanie. Dwa sposoby regulowania drzewa decyzyjnego to ustawienie ograniczeń parametrów modelu i uproszczenie modelu poprzez przycinanie.
- **Przewidywanie zmiennych ciągłych** - ponieważ drzewa decyzyjne mogą przyjmować stałe dane liczbowe, mogą nie być praktycznym sposobem przewidywania takich wartości. Dlatego prognozy drzewa decyzyjnego należy podzielić na dyskretne kategorie, co prowadzi do utraty informacji przy stosowaniu modelu do wartości ciągłych.

Section 2

Miary niejednorodności węzłów/liści

Subsection 1

Zagadnienie regresji

Suma kwadratów

$$SS^{(Node)} = \sum_{i=1}^{n^{(Node)}} (y_i - \bar{y}^{(Node)})^2$$

gdzie $n^{(Node)}$ jest liczbą obserwacji w danym węźle oraz \bar{y} :

$$\bar{y}^{(Node)} = \frac{1}{n^{(Node)}} \sum_{i=1}^{n^{(Node)}} y_i$$

jest finalną wartością teoretyczną $\forall i \in Node, \hat{y}_i = \bar{y}^{(Node)}$.

Subsection 2

Zagadnienie klasyfikacji

Entropia Shannona

$$H_S^{(Node)} = - \sum_{l=1}^k \hat{y}^{(l)} \log_2 \hat{y}^{(l)}$$

gdzie $\hat{y}^{(l)}$ jest warunkowym prawdopodobieństwem przynależności obserwacji do danej klasy l , wyznaczanym jako udział obserwacji z danej klasy w danym węźle:

$$\forall l, \hat{y}^{(l)} = \frac{n^{(Node)}}{\sum_{i=1}^{n^{(Node)}} I(y_i = l)}$$

Indeks Giniego

$$Gini^{(Node)} = 1 - \sum_{l=1}^k (\hat{y}^{(l)})^2$$

Błąd klasyfikacji

$$Err^{(Node)} = 1 - \max_l \hat{y}^{(l)}$$

Subsection 3

Zysk informacyjny / spadek zróżnicowania

Zagadnienie regresji

$$SS^{(Parent)} - \left(\frac{n^{(Left)}}{n^{(Parent)}} * SS^{(Left)} + \frac{n^{(Right)}}{n^{(Parent)}} * SS^{(Right)} \right)$$

Zagadnienie klasyfikacji (przykład entropii)

$$H_S^{(Parent)} - \left(\frac{n^{(Left)}}{n^{(Parent)}} * H_S^{(Left)} + \frac{n^{(Right)}}{n^{(Parent)}} * H_S^{(Right)} \right)$$

Section 3

Rekurencyjny podział przestrzeni

Algorytm zewnętrzny

```
Tree( formula, data, param ){  
  
  StopIfNot( formula, data, param )  
  
  tree <- CreateTree()  
  
  AssignInitialMeasures( tree )  
  
  BuildTree( tree, formula, data, param )  
  
  return( tree )  
  
}
```

Algorytm wewnętrzny

```
BuildTree( node, formula, data, param ){  
  
  node <- AssignMeasures( formula, data, param )  
  bestsplit <- FindBestSplit( formula, data, param )  
  
  if( StopCond(bestsplit) == TRUE ){  
    return( node )  
  }else{  
    left <- CreateLeaf(node,formula,data,param,bestsplit)  
    BuildTree( left, formula, data, param )  
    right <- CreateLeaf(node,formula,data,param,bestsplit)  
    BuildTree( right, formula, data, param )  
  }  
}
```

Section 4

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 5

Literatura

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