Techniki eksploracji danych

Krzysztof Gajowniczek

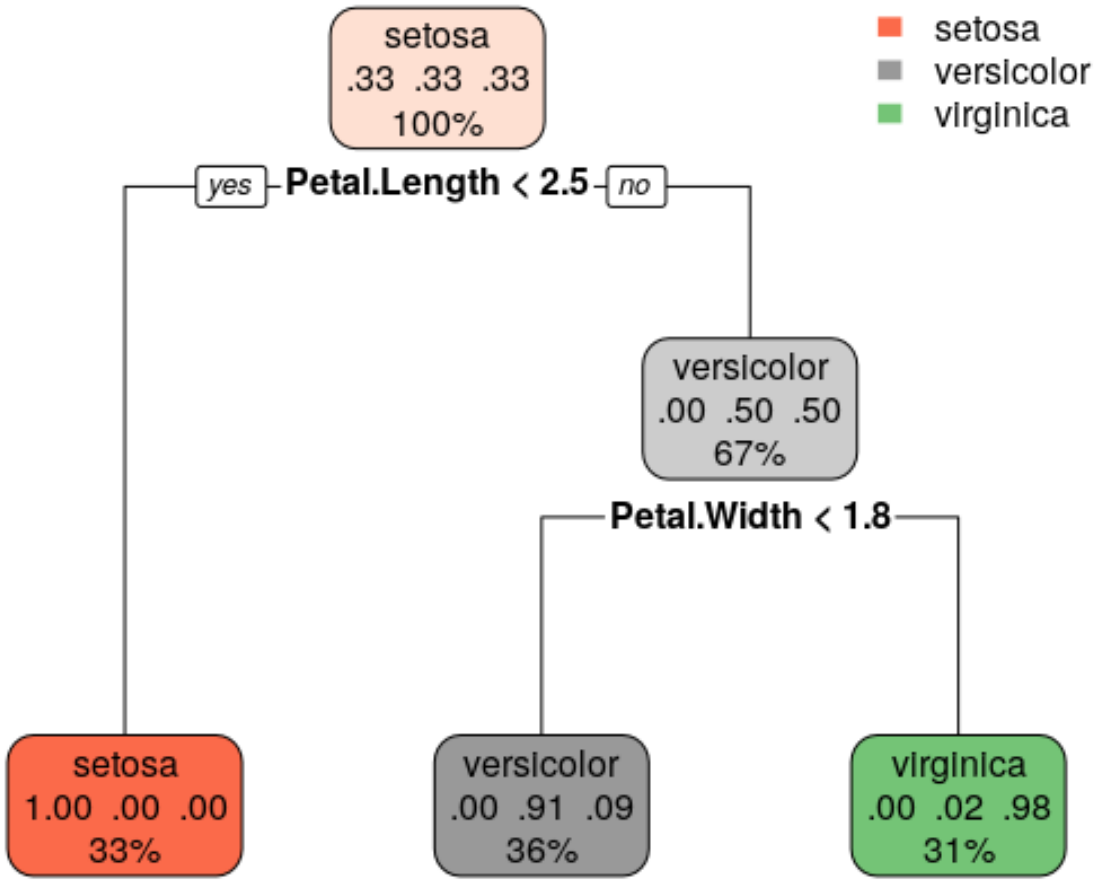
Rok akademicki: 2021/2022

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 dla danego atrybutu.
- Każda gałąź wychodząca z węzła odpowiada jednej z możliwych wartości atrybutu.
- Każdy węzeł na liściu przypisuje wartość przewidywaną.


setosa
versicolor
virginica
setosa
.33 .33 .33
100%
yes
Petal.Length < 2.5
no
versicolor
.00 .50 .50
67%
Petal.Width < 1.8
setosa
1.00 .00 .00
33%
versicolor
.00 .91 .09
36%
virginica
.00 .02 .98
31%

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 kategorycznych.
- **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.
- **Wykorzystywanie zmiennych ciągłych** - ponieważ drzewa decyzyjne mogą przyjmować stałe dane liczbowe, mogą nie być praktycznym sposobem wykorzystania/przewidywania takich wartości.

Section 2

Miary niejednorodności węzłów/liści

Subsection 1

Zagadnienie regresji

Suma kwadratów

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

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

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

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

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:

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

Indeks Giniego

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

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

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*