Drzewa decyzyjne w R

Michał Kaftanowicz

2017-11-07

Powtórzenie pojęć

Drzewo to spójny graf acykliczny

- Krawędzie tego grafu to gałęzie drzewa (branches)
- Wierzchołki tego grafu to węzły drzewa (nodes)
- Wierzchołki tego grafu z tylko jedną krawędzią to liście drzewa (leaves)

Drzewo decyzyjne (klasyfikacyjne)

- do węzła wchodzi jedna gałąź
- w węźle zachodzi test, od którego zależy dalsza ścieżka w drzewie
- liście reprezentują etykiety klasfikacji

Powtórzenie pojęć

Cechy drzew decyzyjnych

- proste w interpretacji
- algorytm zachłanny

Przykład drzewa decyzyjnego

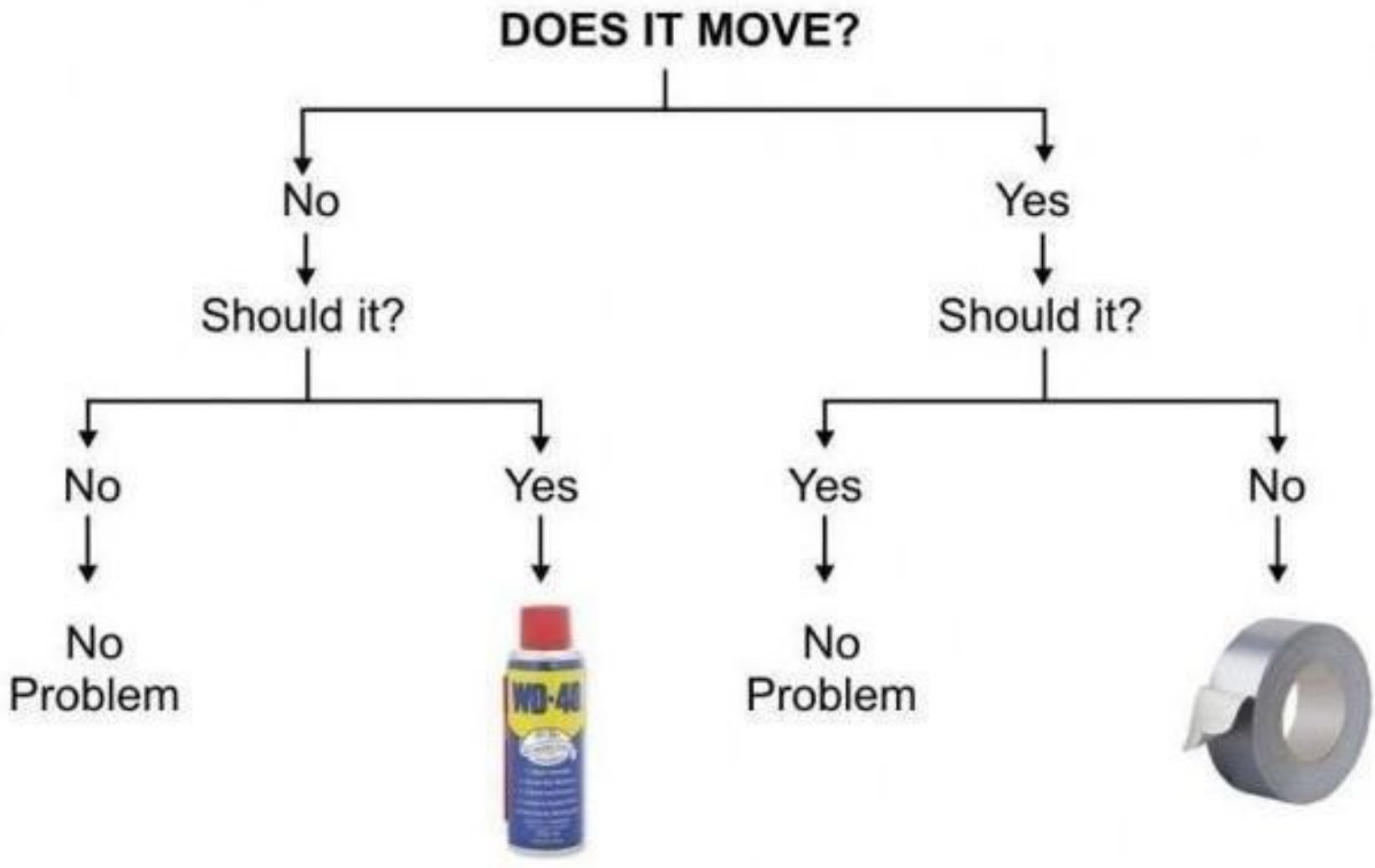

Figure 1: by Duncan Hull

Drzewa decyzyjne w R

Biblioteka RPART

```
library(rpart)
```

Recursive Partitioning and Regression Trees

Zbiór danych: pasażerowie Titanica

Kolumna	Definicja i wartości
survival	Przeżył(a) 0 = nie, 1 = tak
pclass	klasa biletu (1, 2, 3)
sex	płeć
age	wiek
sibsp	liczba rodzeństwa lub małżonków na pokładzie
parch	liczba rodziców lub dzieci na pokładzie
ticket	numer biletu
fare	opłata
cabin	numer kabinu
embarked	port, w którym pasażer(ka) wstąpił(a) na pokład: C = Cherbourg, Q = Queenstown, S = Southampton

Eksploracja danych

```
str(tdf)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthon
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..:
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4
```

Eksploracja danych

```
table(tdf$Survived)
```

```
##
##   0   1
## 549 342
```