Podstawy Przetwarzania Danych

Laboratoria 8: Usuwanie braków

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

Wersja 1.0
5 marca 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Usuwanie brakujących elementów

```
1 x<-na.omit(x)
```

- Funkcja na.omit usuwa puste dane ze zbioru danych.
- Jest to implementacja metody *List-Wise Deletion*
- Metoda powoduje usunięcie całego rekordu.

Wyliczanie statystyk dla brakujących elementów

```
1 m<-mean(x, na.rm=TRUE)
```

- Niektóre podstawowe funkcje pozwalają na wyliczanie statystyk ignorując braki.
- Dotyczy to na przykład funkcji sum i mean.
- Ustalenie wartości parametru na.rm na TRUE pozwala na implementację *Pair-Wise Deletion*.

Zadanie 1

- Dla danych z pliku wifi.csv wylicz kowariancję zmiennych wifi.1 i wifi.2.
- Wylicz kowariancję tylko dla pełnych danych w tych kolumnach.
- Następnie wylicz kowariancję obliczając średnią korzystając ze wszystkich dostępnych danych w kolumnie.

Drzewa CART

```
cart.tree <- rpart(y ~ ., data = data)
summary(cart.tree)
cart.tree.pred<-predict(cart.tree, data[,-1])
```

- Drzewa CART pozwalają na analizę danych zawierających braki.
- W podsumowaniu generowanym przez metodę summary uzyskujemy informacje o surogatach decyzji stosowanych w wypadku braku danych.
- Metoda predict pozwala na predykcję rekordów zawierających dane

Zadanie 2

- Dla danych z pliku wifi.csv zbuduj drzewo CART pozwalające na estymację zmiennej y na podstawie siły sygnałów Wi-Fi.
- Wylicz średni błąd regresji, używając danych uczących jako dane testowe.

Zadanie 3

- Dla danych z pliku wifi.csv zbuduj drzewo CART pozwalające na estymację zmiennej y na podstawie siły sygnałów Wi-Fi.
- Zastąp brakujące sygnały wartością średnią dla danego źródła.
- Wylicz średni błąd regresji, używając danych uczących jako dane testowe.

Algorytm MICE

```
require(mice)
miceAlg<-mice(data,m=5)
```

- Algorytm MICE jest implementowany przez metodę mice.
- Metoda pozwala przeprowadzić m krotną imputację

Parametry algorytmu MICE

```
1 miceAlg<-mice(wifi, m=5, method="cart", visitSequence="
      monotone")
```

- Metoda mice pozwala na parametryzację metody imputacji i kolejności przeglądania kolumn z danymi.
- Wartość method określa metodę imputacji pojedynczej stosowanej przez algorytm.
- Metoda może być określona osobno dla różnych typów danych.
- Wartość visitSequence pozwala określić kolejność przeglądania danych.
- Dostępne wartości:
  - roman, arabic - od lewej do prawej i odwrotnie
  - monotone, revmonotone - od największych braków i odwrotnie

Usuwanie braków

```
1 data.mice<-complete(miceAlg)
```

- Metoda mice tworzy obiekt opisujący proces imputacji.
- W celu uzupełnienia braków należy wywołać metodę complete.

Zadanie 4

- Dla danych z pliku wifi.csv zbuduj drzewo CART pozwalające na estymację zmiennej y na podstawie siły sygnałów Wi-Fi.
- Zastąp brakujące sygnały stosując metodę MICE z pięcioma iteracjami.
- Wylicz średni błąd regresji, używając danych uczących jako dane testowe.

Zadanie 5

- Dla danych z pliku wifi.csv zbuduj drzewo CART pozwalające na estymację zmiennej y na podstawie siły sygnałów Wi-Fi.
- Zastąp brakujące sygnały stosując metodę MICE z pięcioma iteracjami.
- Zastosuj inputację prostą metodą CART i sekwencję monotonną.
- Wylicz średni błąd regresji, używając danych uczących jako dane testowe.

Porównanie algorytmów regresji

- Algorytmy regresji porównywaliśmy stosując miarę średniego błędu.
- Jest to jednak miara jednostkowa, która może nie oddawać całkowicie charakteru uzyskanych wyników.
- Możemy wykreślić empiryczną funkcję dystrybucji (*Empirical distribution function*) dla każdego z porównywanych algorytmów, aby je porównać wizualnie.
- Pakiet ggplot2 oferuje metodę stat_ecdf pozwalającą na wykreślenie dystrybucji.
- Stosując rozgraniczenie poprzez parametry prezentacji, np. przez zmienną color możemy wykreślić dystrybucję dla kilku wyników na raz.

Zadanie 6

- Porównaj wyniki uzyskane w zadaniach 2-5.
- Wykreśl dystrybucję błędów metod.
- Korzystając z funkcji xlim i ylim ogranicz pole wykresu, aby podkreślić różnice między wynikami.