Podstawy Przetwarzania Danych

Laboratoria 10: Próbkowanie danych

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.

- Wyliczenie wielkości próby dla danych nienormatywnych można przeprowadzić według wzoru

$$n = Z^2 * s^2 / E^2$$

- s - odchylenie standardowe
- Z - poziom ufności
- E - margines błędu

Zadanie 1

- Dla danych z pliku FEB15.csv wylicz wielkość próbki potrzebnej do szacowania średniej prędkości AverageSpeed.
- Sprawdź wyniki dla $Z = 0.05$ i $E = 0.05$ oraz $E = 0.01$

Zadanie 2

- Dla danych z pliku FEB15.csv załóż, że poprawne pomiary to te dla których DataQuality wynosi 1.
- Na tej podstawie wylicz współczynnik odpowiedzi.
- Wylicz rozmiar próbki którą należy zebrać dla parametrów $Z = 0.05$ i $E = 0.01$ uwzględniając współczynnik odpowiedzi.

Próbkowanie

```
1 sample(x, size, replace = FALSE, prob = NULL)
```

- Metoda sample tworzy próbkę ze zbioru x
- Zmienna x może być liczbą, wtedy próbkowany jest zbiór 1:x
- Wybieranych jest size elementów z próbki

Zadanie 3

- Przygotuj zbiór danych według następujących instrukcji.
- Dane z pliku FEB15.csv podziel na zbiór uczący i testowy względem mediany daty Date.
- Ze zbioru uczącego wylosuj próbkę danych o rozmiarze wyliczonym w poprzednim ćwiczeniu, uwzględniając współczynnik odpowiedzi.
- Ze zbioru testowego i z utworzonej próbki usuń elementy dla których DataQuality jest różne od 1.
- Przed losowaniem próbki wywołaj set.seed(100)

Zadanie 4

- Dla przygotowanego zbioru danych zbuduj drzewo decyzyjne estymujące średnią prędkość AverageSpeed na podstawie zmiennych Date,TimePeriod i LinkLength
- Drzewo wytrenuj na stworzonej próbce danych.
- Oblicz błąd średni drzewa uzyskany na danych testowych.

Tworzenie sekwencji

```
1 seq(from = 1, to = 1, by = ((to - from)/(length.out -
      1)))
```

- Polecenie seq tworzy sekwencję wartości od parametru from do parametru to z krokiem by.

Zadanie 5

- Powtórz przygotowanie zbioru danych, ale dobieraj próbkę ze zbioru uczącego sekwencyjnie, w równych odstępach
- Zbuduj drzewo decyzyjne estymujące średnią prędkość AverageSpeed na podstawie zmiennych Date,TimePeriod i LinkLength
- Drzewo wytrenuj na stworzonej próbce danych.
- Oblicz błąd średni drzewa uzyskany na danych testowych.

Parametry próbkowania

```
1 sample(x, size, replace = FALSE, prob = NULL)
```

- Parametr replace pozwala przeprowadzić próbkowanie z powtórzeniami.
- Parametr prob jest wektorem, który pozwala nadać elementom zróżnicowane prawdopodobieństwo wyboru.

Zadanie 6

- Powtórz przygotowanie zbioru danych, ale dobieraj próbkę stratyfikacyjnie, dobierając taką samą liczbę odczytów dla każdego odcinka autostrady LinkRef.
- Jeżeli dla któregoś odcinka brakuje danych to zastosuj próbkowanie z powtórzeniami.
- Zbuduj drzewo decyzyjne estymujące średnią prędkość AverageSpeed na podstawie zmiennych Date,TimePeriod i LinkLength
- Drzewo wytrenuj na stworzonej próbce danych.
- Oblicz błąd średni drzewa uzyskany na danych testowych.

Zadanie 7

- Porównaj wyniki dla wszystkich metod próbkowania.
- Wykreśl skumulowany rozkład błędów.
- Powtórz każde próbkowanie 10 razy i wykreśl wykres pudełkowy porównujący średnie wyniki uzyskane przez metody
- Która metoda daje najlepsze wyniki?
- Jak wygląda odchylenie standardowe tych metod?