

Metody przetwarzania i analizy danych w R

Łukasz Wawrowski

Contents

Wprowadzenie	5
1 Wprowadzenie	7
1.1 Narzędzie	7
1.2 Cele analiz	7
2 Testowanie hipotez	9
2.1 Hipoteza statystyczna	9
2.2 Poziom istotności i wartość p	9
2.3 Testy parametryczne i nieparametryczne	9
3 Regresja	11
3.1 Regresja prosta	11
3.2 Regresja wieloraka	11
4 Grupowanie	13
4.1 Metoda k-średnich	13
4.2 Metoda hierarchiczna	13
5 Klasyfikacja	15
5.1 Drzewa klasyfikacyjne	15
5.2 KNN	15

Wprowadzenie

Literatura podstawowa:

- Przemysław Biecek - *Przewodnik po pakiecie R*
- Marek Gągolewski - *Programowanie w języku R. Analiza danych, obliczenia, symulacje.*
- Garret Golemund, Hadley Wickham - *R for Data Science* (polska wersja)

Literatura dodatkowa:

- inne pozycje po polsku
- inne pozycje po angielsku

Internet:

- R-bloggers
- rweekly

Chapter 1

Wprowadzenie

1.1 Narzędzie

- darmowe
- wszechstronne
- wsparcie społeczności
- wersja desktopowa i serwerowa

czyli **R** - środowisko do obliczeń statystycznych i wizualizacji wyników

- strona projektu: r-project.org
- świetne IDE: RStudio
- wersja przeglądarkowa: rstudio.cloud

R + Python

1.2 Cele analiz

Podstawowe:

- wnioskowanie statystyczne - porównywanie grup
- regresja - poszukiwanie związków
- klasyfikacja - przyporządkowanie do grup
- grupowanie - poszukiwanie grup
- prognozowanie - patrzenie w przyszłość

Inne:

- analiza języka naturalnego
- rozpoznawanie obrazów
- analiza koszykowa
- ...

1.2.1 Eksporacja danych

Pakiet `tidyverse`

```
library(tidyverse)
```

- analiza częstości dla zmiennych jakościowych
- analiza struktury dla zmiennych ilościowych

Case study: Wybory 2018

Chapter 2

Testowanie hipotez

2.1 Hipoteza statystyczna

Przypuszczenie dotyczące własności analizowanej cechy, np. średnia w populacji jest równa 10, rozkład cechy jest normalny.

Formuluje się zawsze dwie hipotezy: hipotezę zerową (H_0) i hipotezę alternatywną (H_1). Hipoteza zerowa jest hipotezą mówiącą o równości:

$$H_0 : \bar{x} = 10$$

Z kolei hipoteza alternatywna zakłada coś przeciwnego:

$$H_1 : \bar{x} \neq 10$$

Zamiast znaku nierówności (\neq) może się także pojawić znak mniejszości ($<$) lub większości ($>$).

2.2 Poziom istotności i wartość p

Hipotezy statystyczne weryfikuje się przy określonym poziomie istotności α , który wskazuje maksymalny poziom akceptowalnego błędu (najczęściej $\alpha = 0,05$).

Większość programów statystycznych podaje w wynikach testu wartość p. Jest to prawdopodobieństwo uzyskania obserwowanych wyników przy założeniu prawdziwości hipotezy zerowej.

Generalnie jeśli $p < \alpha$ - odrzucamy hipotezę zerową.

Krytyka wartości p

2.3 Testy parametryczne i nieparametryczne

Testy statystyczne dzielą się na dwie grupy:

- parametryczne, które wymagają spełnienia założeń, ale są dokładniejsze,
- nieparametryczne, które nie wymagają tylu założeń, ale są mniej dokładne.

Chapter 3

Regresja

3.1 Regresja prosta

Na podstawie danych dotyczących informacji o doświadczeniu i wynagrodzeniu pracowników zbuduj model określający ‘widełki’ dla potencjalnych pracowników o doświadczeniu równym 8, 10 i 11 lat.

regresja_prosta.Rmd

cały projekt

3.1.1 Zadanie

Dla danych dotyczących sklepu nr 77 opracuj model zależności sprzedaży od liczby klientów. Ile wynosi teoretyczna sprzedaż w dniach, w których liczba klientów będzie wynosiła 560, 740, 811 oraz 999 osób?

3.2 Regresja wieloraka

Na podstawie danych dotyczących zatrudnienia opracuj model, w którym zmienną zależną jest bieżące wynagrodzenie. Jaka cecha ma największy wpływ na tę wartość?

Opis zbioru:

- id - kod pracownika
- plec - płeć pracownika (0 - mężczyzna, 1 - kobieta)
- data_urodz - data urodzenia
- edukacja - wykształcenie (w latach nauki)
- kat_pracownika - grupa pracownicza (1 - ochroniarz, 2 - urzędnik, 3 - menedżer)
- bwynagrodzenie - bieżące wynagrodzenie
- pwynagrodzenie - początkowe wynagrodzenie
- staz - staż pracy (w miesiącach)
- doswiadczenie - poprzednie zatrudnienie (w miesiącach)
- zwiazki - przynależność do związków zawodowych (0 - nie, 1 - tak)
- wiek - wiek (w latach)

regresja_wieloraka.Rmd

cały projekt

3.2.1 Zadanie

Na podstawie zbioru dotyczącego 50 startupów określ jakie czynniki w największym stopniu wpływają na przychód startupów.

Chapter 4

Grupowanie

Metody grupowania są wykorzystywane np. do segmentacji klientów, w przypadku, gdy nie jest znany końcowy podział.

4.1 Metoda k-średnich

Algorytm:

1. Wskaż liczbę grup k .
2. Wybierz dowolne k punktów jako centra grup.
3. Przypisz każdą z obserwacji do najbliższego centroidu.
4. Oblicz nowe centrum grupy.
5. Przypisz każdą z obserwacji do nowych centroidów. Jeśli któraś obserwacja zmieniła grupę - przejdź do kroku nr 4, a w przeciwnym przypadku zakończ algorytm.

Zalety:

- dobrze działa zarówno na małych, jak i dużych zbiorach
- efektywny

Wady:

- trzeba wskazać liczbę grup
- losowy wybór punktów początkowych

4.2 Metoda hierarchiczna

Algorytm:

1. Każda obserwacji stanowi jedną z N pojedynczych grup.
2. Na podstawie macierzy odległości połącz dwie najbliższe leżące obserwacje w jedną grupę ($N - 1$ grup).
3. Połącz dwa najbliższe sobie leżące grupy w jedną ($N - 2$ grup).
4. Powtórz krok nr 3, aż do uzyskania jednej grupy.

Zalety:

- prosty sposób ustalenia liczby grup
- praktyczny sposób wizualizacji

Wady:

- nieodpowiedni dla dużych zbiorów

4.2.1 Zadanie

Na podstawie zbioru zawierającego informacje o klientach sklepu dokonaj grupowania klientów.

Opis zbioru:

- klientID - identyfikator klienta
- plec - płeć
- wiek - wiek
- roczny_dochod - roczny dochód wyrażony w tys. dolarów
- wskaznik_wydatkow - klasyfikacja sklepu od 1 do 100

grupowanie.Rmd

cały projekt

4.2.2 Zadanie 2

Dokonaj grupowania danych dotyczących 32 samochodów według następujących zmiennych: pojemność, przebieg, lata oraz cena.

4.2.3 Zadanie 3

Rozpoznawanie czynności na podstawie danych z przyspieszeniomierza w telefonie: User Identification From Walking Activity Data Set

Chapter 5

Klasyfikacja

A visual introduction to machine learning - niestety powstała tylko jedna część.

5.1 Drzewa klasyfikacyjne

Zalety:

- łatwa interpretacja
- nie trzeba normalizować cech
- rozwiązuje problemy liniowe i nieliniowe

Wady:

- mała efektywność przy małych zbiorach danych
- łatwo można przeuczyć

5.2 KNN

Algorytm:

1. Określ liczbę sąsiadów - K
2. Wyznacz K sąsiadów dla nowego punktu na podstawie wybranej odległości
3. Oblicz liczbę sąsiadów, w każdej z grup
4. Przypisz nową obserwację do grupy, w której ma więcej najbliższych sąsiadów

Zalety:

- łatwa interpretacja
- szybki i efektywny

Wady:

- trzeba określić liczbę sąsiadów

5.2.1 Zadanie

Zbuduj model klasyfikacyjny dla zbioru danych dotyczących cech internautów oraz informacji czy zamówili reklamowany produkt czy nie.

Przeprowadź imputację braków danych dla zbioru pracowników.