Politechnika Wrocławska

Statystyka

Zastosowanie programu MATLAB w zagadnieniach inżynierskich

Wojciech Myszka

Katedra Mechaniki, Inżynierii Materiałowej i Biometrycznej

listopad 2024

HR EXCELLENCE IN RESEARCH

1. Statystyka

2. Demonstracja

3. Lepszy wielowymiarowych przykład danych kategorycznych

Statystyka

Politechnika Wrocławska

Dane

- Można czytać z „zewnątrz” — formatów danych jest nieskończoność
- Można korzystać z danych przykładowych dostępnych wraz z MATLABem

Politechnika Wrocławska

Typy danych

- numeryczne (była już o tym mowa)
- tekstowe (będzie szybkie przypomnienie)
- „kategoryczne” w odróżnieniu od danych ilościowych definiują jakieś cechy obiektu (płeć, kolor oczu,...)
- cell array

Dane tekstowe (przypomnienie) I

Przypominam o drobnej różnicy między cudzysłowem pojedynczym a podwójnym

```
>> a="ala ma kota"
a =
    "ala ma kota"
>> size(a)
ans =
     1     1
>> class(a)
ans =
    'string'
>> b='ala ma kota'
b =
```

Dane tekstowe (przypomnienie) II

```
    'ala ma kota'
>> size(b)
ans =
     1    11
>> class(b)
ans =
    'char'
```

- **podwójny** to `string`
- **pojedynczy** to `char`

Dane tekstowe (przypomnienie) III

Więcej na ten temat w dokumentacji.

String arrays

- `string`
- `strings` — utworzenie pustej tablicy
- `join` — łączenie elementów tablicy typu `string`
- `plus` — konkatenacja napisów (można też użyć znaku +, na przykład tak: `"ala" + " " + "ma" +" " + "kota"`)

Dane tekstowe (przypomnienie) IV

Character arrays

- `char` — tablica znaków ale też konwersja (`s = char("ala");`)
- `cellstr` — konwersja do typu tablica komórek
- `blanks` — tablica złożona z pustych znaków o zadanych rozmiarach
- `newline` — nowa linia
- konkatenacja: `'|'  blanks(5) '|'` da w efekcie |␣␣␣␣␣|

Politechnika Wrocławska

Konwersje

Convert Input Arguments

`convertCharsToStrings`	Konwertuj tablice znakowe na tablice łańcuchowe, pozostawiając inne tablice bez zmian
`convertStringsToChars`	Konwertuj tablice napisów na tablice znakowe, pozostawiając inne tablice bez zmian
`convertContainedStringsToChars`	Konwertuj łańcuchy znaków na dowolnym poziomie tablicy lub struktury komórek

Politechnika Wrocławska

Cell array I

Cell array to *tablica komórek* czyli typ danych zawierający

- indeksowane kontenery danych zwane komórkami,
- przy czym każda komórka może zawierać dane dowolnego typu.
- Tablice komórek zwykle zawierają
  - listy tekstu,
  - kombinacje tekstu i liczb lub
  - tablice numeryczne o różnych rozmiarach.

Odwołania do zestawu komórek: indeks w nawiasach okrągłych ()

Dostęp do zawartości komórek — nawiasy klamrowe {}

Politechnika Wrocławska

Cell array II

```
C = {1,2,3;
     'text',rand(5,10,2),{11; 22; 33}}
C =
  2×3 cell array
    {[   1]}    {[          2]}    {[     3]}
    {'text'}    {5×10×2 double}    {3×1 cell}
>> C{2,2}(1,1,1)
ans =
    0.6837
```

DEMONSTRACJA

Demonstracja

Politechnika Wrocławska

Dane kowidowe

1. Temat już raczej zszedł z agendy, ale...
2. Ciągle dostępny jest spory zestaw (w miarę) wiarygodnych danych dotyczących epidemii
3. Serwis ourworldindata.org zawiera bardzo dużo interesujących zestawów danych; w tym dane dotyczące epidemii SARS-CoV-2
4. Z COVIDem jest ten kłopot, że przestano w niego wierzyć i trudno teraz o obiektywne dane.
5. Plik z danymi, którym posługiwałem się na tych zajęciach przestał być uaktualniany i od około trzech miesięcy nic się tam nie dzieje.
6. Serwis Our World in Data ciągle dostarcza jakiś zestaw danych

Politechnika Wrocławska

„Stary” zbiór danych... I

1. Zaczniemy więc od „starego zbioru danych” znajdującego się pod adresem covid.ourworldindata.org/data/owid-covid-data.csv a stworzonego na podstawie danych w serwisie GitHub: github.com/owid/covid-19-data/tree/master/public/data
2. Dane dostępne są (były) w jednym z trzech formatów:
   2.1 CSV
   2.2 XSLX (Excel)
   2.3 JSON
3. Pierwszy i drugi nie wymagają komentarza

Politechnika Wrocławska

„Stary” zbiór danych... II

JSON...

- ...to skrót pochodzący od **JavaScript Object Notation** formatu danych opracowanego specjalnie na potrzeby języka JavaScrip używanego powszechnie przez przeglądarki internetowe.
- plik tekstowy kodowany jako UTF-8 o strukturze zależnej od danych i ustalonych regułach tworzenia tej struktury

Politechnika Wrocławska

„Stary” zbiór danych... III

przykład

```
{
  "menu": {
    "id": "file",
    "value": "File",
    "popup": {
      "menuitem": [
        {"value": "New", "onclick": "CreateNewDoc()"},
        {"value": "Open", "onclick": "OpenDoc()"},
        {"value": "Close", "onclick": "CloseDoc()"}
      ]
    }
  }
}
```

Import danych CSV

1. Polecam aplikację Import Tool

Import danych Excel I

readtable()

```
start = tic;
A = readtable('owid-covid-data.xlsx');
toc(start)
Elapsed time is 101.016576 seconds.
size(A)
ans = 1x2
      429435          67
```

Import danych Excel II

readmatrix()

```
start = tic;
B = readmatrix('owid-covid-data.xlsx');
toc(start)
Elapsed time is 66.848213 seconds.
size(B)
ans = 1x2
      429435          67
```

Import danych Excel III

readcell()

```
start = tic;
C = readcell("owid-covid-data.xlsx");
toc(start)
size(C)
```

1. Te polecenia zrestartowały mi komputer. Najprawdopodobniej zabrakło pamięci
2. Kiedyś, gdy zbiór nie był tak wielki trwało długo, ale importowało...

Politechnika Wrocławska

Import pliku JSON I

1. Nie ma funkcji pozwalającej bezpośrednio czytać pliki typu `.json`.
2. Trzeba przeczytać **cały** plik w trybie tekstowym

   a następnie
3. Jest funkcja `jsondecode()`, która rozkodowuje dane
4. Wszystko działa, ale danym nadawana jest postać drzewiasta i konwertowane są one na złożoną strukturę
5. Jest też funkcja `jsonencode()` pozwalająca zapisać złożoną strukturę MATLAB w formacie json.

Czytanie plików tekstowych w MATLABie przypomina analogiczną operację w językach C++/C

Import pliku JSON II

```
start = tic;
fname = 'owid-covid-data.json';
fid = fopen(fname);
raw = fread(fid,inf);
str = char(raw');
fclose(fid);
val = jsondecode(str);
toc(start)
```

```
Elapsed time is 12.988704 seconds.
```

Import Tool I

MATLAB dokonuje analizy danych i podejmuje sensowne deyzje. Do importu używa funkcji readtable() ale deklaruje niektóre z kolumn jako kategorie:

- kod kraju (iso_code)
- kontynent (continent)
- location

Kolumnę z datami poprawnie rozpoznaje jako obiekt typu datetime.

Wiele danych poprawnie traktuje jako double, ale **ma problem z kolumnami, które na początku pliku są puste** (bo nie analizuje całego pliku, tylko kilka pierwszych wierszy. Gdy wygenerujemy skrypt — możemy to poprawić, albo odpowiednich konwersji dokonywać później.

Odpowiadają za to linie

Politechnika Wrocławska

Import Tool II

```
% Specify column names and types
opts.VariableNames = ["iso_code", "continent", "location", "date"
...
opts.VariableTypes = ["categorical", "categorical", "categorical"
...
```

Import jest bardzo szybki.

Politechnika Wrocławska

Analiza danych

Jeżeli pomiary mają przypisane kategorie wybór danych jest bardzo prosty.

Kod ISO Polski to `POL` i wybranie z całej tabeli danych dotyczących Polski jest bardzo łatwe:

- najpierw wskazujemy wiersze, które dotyczą Polski (nie muszą stanowić one obszaru spóknego)

```
rows = owidcoviddata.iso_code == 'POL';
```

- następnie z tablicy wyciągamy tylko dane dotyczące <polski:

```
Polska = owidcoviddata(rows,:);
```

I wszystko

Politechnika Wrocławska

„Nowy zbiór” danych

- Serwis Our World in Data w dalszym ciągu udostępnia dane COVIDoww w nieco innym formacie.
- Nie jest to jedyny dostępny tam zestaw danych. Jest ich wiele.
- Cechą charakterystyczną zestawów danych jest wyposażenie ich w metadane (dane o danych)
- Metadane są w formacie JSON
- Dane zawierają również informacje współczesne (ale już raczej uzupełniane rzadziej)

Lepszy wielowymiarowych przykład danych kategorycznych

Politechnika Wrocławska

samochody I

1. Wraz z MATLABem dostajemy przykładowe zestawy danych.
2. Należą do nich `carsmall` i `carbig`
3. pierwszy zawiera informacje o stu konkretnych modelach samochodów, drugi o czterystu
4. Zapisane tam parametry to:
   - przyśpieszenie
   - liczba cylindrów
   - pojemność skokowa
   - moc silnika
   - wytwórca
   - model samochodu
   - rok

samochody II

- zużycie paliwa (w milach z galona, bo to USA)
- kraj pochodzenia
- waga

Dane te (w pewnym sensie) opisują obiekt, którym jest *jakiś samochód*

Dane są wielowymiarowe i nie da się łątwo pokazać zależności parametrów samochodu od innych jego cech.

gplotmatrix() I

1. Funkcja `gplotmatix()` pozwala pokazać zależności różnych numerycznych parametrów
2. Jej wywołanie nie jest bardzo proste
3. ale efekty są interesujące
4. Wybieramy zestaw danych `carsmall`

```
load carsmall
X = [Acceleration Displacement Horsepower MPG Weight];
xnames = {'Acceleration','Displacement','Horsepower','MPG','Weigh
[h,ax] = gplotmatrix(X,[],Cylinders,[],[],[],[],'variable',xnames
title('Car Data')
```

- `x` to numeryczne parametry, które będą obrazowane na tym wykresie
- `xnames` ich ludzkie nazwy

Politechnika Wrocławska

gplotmatrix() II

- trzeci parametr funkcji `gplotmatric` oznacza dodatkowe grupowanie kolorami w zależności od liczby cylindrów
- na jednym wykresie dostajemy zależności między sześcioma zmiennymi numerycznymi

gplotmatrix() III