Eksploracja danych

Piotr Lipiński

Lista zadań nr 1 – Minikurs Matlaba

Zadanie 0. *(rozgrzewkowe, 1 punkt, zadanie należy przesłać emailem)*
Utwórz skrypt definiujący poniższe zmienne:
a = [1, 2, 3, 4, …, 100] (wektor złożony z liczb całkowitych od 1 do 100)
b = [1, 3, 5, 7, …, 99] (wektor złożony z liczb całkowitych nieparzystych od 1 do 99)
c = [-1.00 * π, -0.99 * π, …, -0.01 * π, 0, 0.01 * π, …, 0.99 * π, 1.00 * π]
d = [-1.00 * π, -0.99 * π, …, -0.01 * π, 0.01 * π, …, 0.99 * π, 1.00 * π]
e = [e_1, e_2, …, e_{100}], gdzie e_i = sin(i), jeśli sin(i) > 0, lub e_i = 0 w przeciwnym przypadku
A = macierz rozmiaru 10 x 10 zawierająca liczby całkowite od 1 do 100: w pierwszym wierszu od lewej 1, 2, …, w drugim wierszu od lewej 11, 12, .., itd. (wskazówka: użyć polecenia `reshape`)
B = macierz trójdiagonalna rozmiaru 100 x 100 mająca na głównej przekątnej liczby całkowite od 1 do 100, a poniżej i powyżej głównej przekątnej liczby od 99 do 1
C = macierz trójkątna górna wypełniona jedynkami (łącznie z główną przekątną)
D = macierz rozmiaru 2 x 100, w której pierwszy wiersz zawiera elementy d_{1i} = 1 + 2 + … + i, a drugi wiersz zawiera elementy d_{2i} = i!
E = macierz rozmiaru 100 x 100 mająca 1 w pozycji (i, j), jeśli i dzieli j, lub 0 w przeciwnym przypadku.

Zadanie 1. *(1 punkt)*
a) Wygeneruj 10 000 liczb z rozkładu jednostajnego na odcinku [-1, 1]. Sporządź ich histogram z 100 przedziałami. Porównaj histogram z wykresem funkcji gęstości. Powtórz obliczenia dla 100 000 liczb.

b) Wygeneruj 10 000 liczb z rozkładu normalnego o średniej 5 i odchyleniu standardowym 3. Sporządź ich histogram z 100 przedziałami. Porównaj histogram z wykresem funkcji gęstości. Powtórz obliczenia dla 100 000 liczb.

c) Wygeneruj 10 000 punktów (x, y), których współrzędna x ma rozkład normalny N(2, 5), zaś współrzędna y ma rozkład normalny N(3, 1). Sporządź wykres tych punktów. Porównaj go z wykresem funkcji gęstości. Powtórz obliczenia dla 100 000 punktów.

d) Używając danych wygenerowanych w poprzednim punkcie oszacuj prawdopodobieństwo, że X < Y dla zmiennych losowych X z rozkładem normalnym N(2, 5) i Y z rozkładem normalnym N(3, 1). Uzyskaną wartość porównaj z dokładnym prawdopodobieństwem takiego zdarzenia obliczonym w oparciu o rachunek prawdopodobieństwa i statystykę.

Zadanie 2. *(1 punkt)*
Utwórz skrypt generujący losowe sudoku, tzn. macierz M rozmiaru 9x9 zawierającą liczby 1, 2, …, 9, taką, że w każdym wierszu każda z liczb występuje dokładnie jeden raz, w każdej kolumnie każda z liczb występuje dokładnie jeden raz oraz w każdej klatce 3x3, powstałej przez podział macierzy M dwoma liniami pionowymi i dwoma liniami poziomymi, każda z liczb występuje dokładnie jeden raz.

Zadanie 3. *(1 punkt)*
a) Niech **x**, **y**, **w** będą wektorami kolumnowymi ustalonej długości *d*. Policz:
- długość wektora **x**,
- średnią ważoną wektora **x** z wagami **w**,
- odległość euklidesową między wektorami **x** i **y**,

- iloczyn skalarny wektorów **x** i **y**.

Obliczenia przeprowadź dla losowo wygenerowanych wektorów **x**, **y**, **w** (dla d = 100).

b) Niech **X** będzie macierzą ustalonego rozmiaru *d* x *N* zawierającą *N* wektorów kolumnowych długości *d*. Niech **y** i **w** będą wektorami kolumnowymi długości *d*. Policz

- długości kolejnych wektorów z macierzy **X** (wyznacz wektor długości *N* zawierający te długości),
- średnią ważoną kolejnych wektorów z macierzy **X** z wagami **w** (wyznacz wektor długości *N* zawierający te średnie),
- odległości euklidesowe między kolejnymi wektorami z macierzy **X** i wektorem **y** (wyznacz wektor długości *N* zawierający te odległości),
- iloczyny skalarne kolejnych wektorów z macierzy **X** i wektora **y** (wyznacz wektor długości *N* zawierający te iloczyny).

Obliczenia przeprowadź dla losowo wygenerowanej macierzy **X** i losowo wygenerowanych wektorów **y** i **w** (dla d = 100 i N = 1000).

Zadanie 4. *(1 punkt)*
Napisz funkcję, która dla danych macierzy **X** i **Y** ustalonego rozmiaru *d* x *N* i *d* x *M* odpowiednio, zawierających *N* i *M* wektorów kolumnowych długości *d*, wyznacza macierz odległości euklidesowych między wektorami tych macierzy, tzn. macierz **D** rozmiaru *N* x *M*, gdzie **D**(*i*, *j*) to odległości między *i*-tym wektorem z macierzy **X** i *j*-tym wektorem z macierzy **Y**. Oblicz czas działania napisanej funkcji dla losowo wygenerowanych macierzy **X** i **Y** dla d = 100 i N = M = 1 000 oraz dla d = 100, N = 10 000 i M = 1 000.

Zadanie 5. *(1 punkt)*
Napisz funkcję, która dla danych macierzy **X** i **Y** ustalonego rozmiaru *d* x *N* i *d* x *M* odpowiednio, zawierających *N* i *M* wektorów kolumnowych długości *d*, wyznacza dla każdego wektora z macierzy **X** najbliższego sąsiada spośród wektorów z macierzy **Y**, tzn. zwraca wektor **h** rozmiaru 1 x *N*, gdzie **h**(*i*) to numer wektora z macierzy **Y** będącego najbliższym sąsiadem *i*-tego wektora z macierzy **X**. Napisz też wersję tej funkcji, która wyznacza *k* najbliższych sąsiadów, tzn. zwraca macierz **H** rozmiaru *k* x *N*, gdzie **H**(*i*, *j*) to numer wektora z macierzy **Y** będącego *i*-tym najbliższym sąsiadem *j*-tego wektora z macierzy **X** (czyli **h**(*j*) = **H**(1, *j*)).

Zadanie 6. *(bonusowe, 1 punkt)*
Zapoznaj się z paradoksem Monty'ego Halla (który był podstawą teleturniejów telewizyjnych "Let's make a deal", w polskiej wersji "Idź na całość"). Napisz skrypt symulujący taki teleturniej. Przeprowadź minimum 10 000 prób i oszacuj prawdopodobieństwo wygranej dla strategii pozostawania przy swoim wyborze oraz dla strategii zmiany wyboru.

UWAGA: Proszę nie korzystać z żadnych funkcji wbudowanych ani bibliotecznych liczących odległości, iloczyny skalarne, itp. Proszę sprawdzić działanie swoich funkcji na przykładowych danych (najlepiej dość dużych rozmiarów). Proszę spróbować ocenić efektywność swoich obliczeń.