Eksploracja danych

Piotr Lipiński

Lista zadań nr 1 – Minikurs Pythona z NumPy

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)*
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),

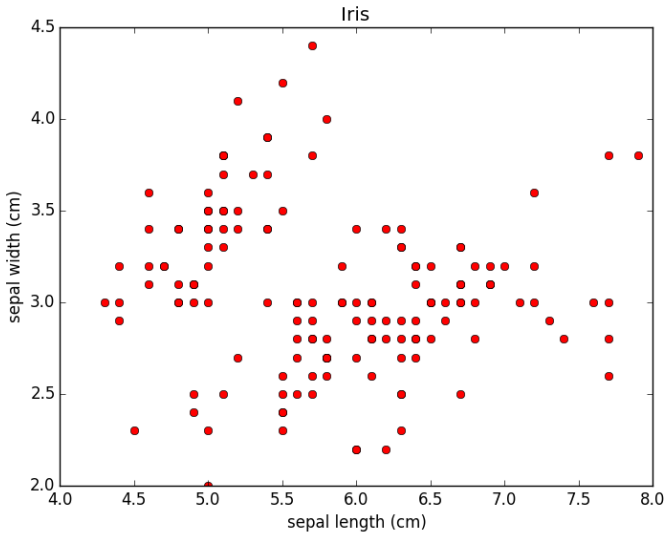
- średnią ważoną kolejnych wektorów z macierzy **X** z wagami **w** (wyznacz wektor długości *N* zawierający te średnie),
- odległości euklidesowe między kolejnymi wektorami z macierzy **X** i wektorem **y** (wyznacz wektor długości *N* zawierający te odległości),
- iloczyny skalarne kolejnych wektorów 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 3. *(1 punkt)*
a) Wczytaj dane IRIS. Można to szybko zrobić korzystając z biblioteki SciKit za pomocą polecenia:

```
from sklearn import datasets
iris = datasets.load_iris()
```

Zobacz co zawiera `iris.data`, `iris.target`, `iris.feature_names`, `iris.target_names`.
b) Przedstaw wczytane dane na wykresie w poniższy sposób (zwróć uwagę na kolory i typ znaczników, opisy osi i tytuł wykresu):



c) Zmień zakres osi: oś X powinna pokazywać wartości od 3 do 9, a oś Y od 1 do 5.
d) Zmień podziałki na osiach, tak aby zaznaczone były tylko liczby całkowite.
e) Każdy gatunek irysa zaznacz innym kolorem.
f) Zapisz rysunek do pliku zadanie1.png.

Zadanie 4. *(1 punkt)*
a) Zrób rysunek podobny do tego z poprzedniego zadania, ale umieść na nim tylko irysy gatunku *setosa* i *versicolor* (nie rysuj irysów gatunku versicolor).
b) Dodaj do rysunku prostą o równaniu y = 2 x – 8.
c) Irysy gatunku setosa znajdujące się pod narysowaną linią zaznacz na czerwono, pozostałe na zielono. Irysy gatunku virginica znajdujące się nad narysowaną linią zaznacz na czerwono, pozostałe na zielono.
d) Zapisz rysunek do pliku zadanie2a.png.
e) Spróbuj zmienić równanie prostej z punktu b) tak, aby zmniejszyć liczbę czerwonych punktów.
f) Zapisz rysunek do pliku zadanie2b.png.

Zadanie 5. *(1 punkt)*
a) Zrób rysunek przedstawiający 10 punktów o następujących współrzędnych (1, 10), (2, 10), (3, 11), (4, 12), (5, 18), (6, 18), (7, 19), (8, 26), (9, 19), (10, 26).
b) Dodaj do rysunku prostą o równaniu y = 2 x + 5.
c) Zapisz rysunek do pliku zadanie3a.png.
d) Dla każdego punktu danych policz jego odległość od wyznaczonej prostej.

e) Spróbuj zmienić równanie prostej z punktu b) tak, aby zmniejszyć sumę odległości punktów danych od prostej.
f) Zapisz rysunek do pliku zadanie3b.png.

Zadanie 6. *(1 punkt)*
Napisz program generujący zestaw dwuwymiarowych danych losowych złożony z K chmur punktów, taki że:
a) każda chmura punktów składała się z 1000 punktów o współrzędnych (x, y), gdzie x pochodzi z rozkładu normalnego $N(a_i, 1)$, y pochodzi z rozkładu normalnego $N(b_i, 1)$, zaś (a_i, b_i) to centrum i-tej chmury punktów,
b) centra chmur punktów (a_i, b_i) tworzą wielokąt foremny o boku o zadanej długości d.
Uruchom program dla K = 7, K = 11 i K = 23 oraz d = 5, d = 10 i d = 15. Zrób rysunki przedstawiające wyniki.

Zadanie 7. *(1 punkt)*
Dla danych z poprzedniego zadania policz odległość każdego punktu danych od każdego centrum chmury i na sporządzonych wykresach zaznacz kolorem czerwonym te punkty danych, które znajdują się bliżej centrum innej chmury niż chmury, z której pochodzą, a kolorem zielonym pozostałe punkty danych. Jak zależy frakcja punktów czerwonych od długości d? Jaka powinna być wartość d, żeby punkty czerwone stanowiły około 10% wszystkich punktów danych?

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.