Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

ЗВІТ

про виконання лабораторної роботи №1
з дисципліни «Інтелектуальний аналіз даних»

Виконала:

Студентка III курсу
Групи КА-76
Хиленко В.В.

Перевірила:
Недашківська Н.І.

Київ – 2020

Варіант №17

Завдання:

Розглянути критерій якості кластеризації - ентропію розбиття:

$$PE = -\frac{\sum_{j=0}^{N}\sum_{k=1}^{g} u_{kj}\ln u_{kj}}{N},$$

де N - задана кількість об'єктів, які кластеризуються, $1 \le g \le N$ - задана кількість кластерів, $U = \{(u_{kj})|k = 1,\dots,g, j = 1,\dots N\}$ - матриця розбиття, $u_{kj} \in (0, 1]$, причому $u_{kj} = 1$ означає приналежність j-го об'єкту k-му кластеру. $\sum_{k=1}^{g} u_{kj} = 1, \sum_{j=1}^{N} u_{kj} < N.$

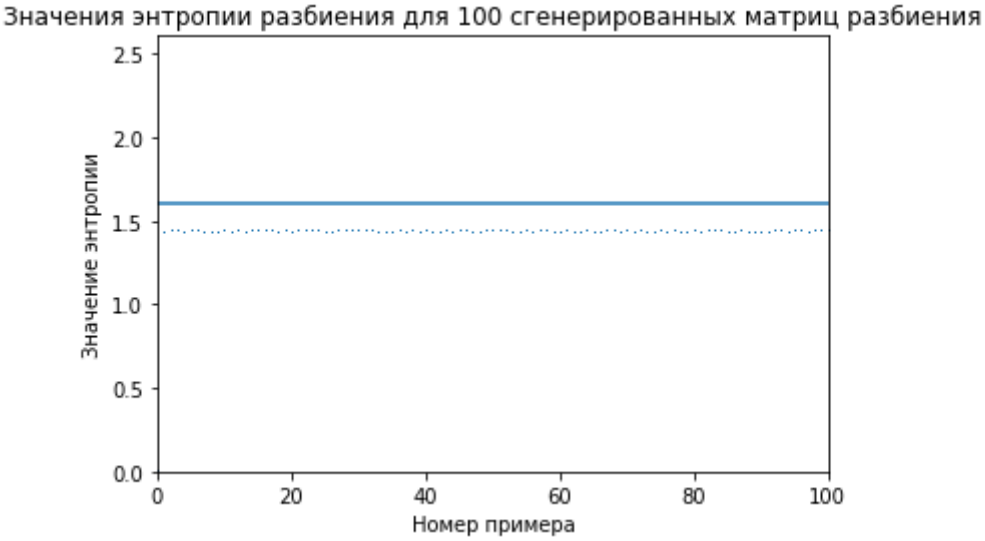
Використовуючи результати моделювання великої кількості матриць розбиття, показати, що

$$PE \in [0, \ln g]$$

Текст програми:

```
import numpy as np
import matplotlib.pyplot as plt
import math
import random

N = 1000 #кол-во кластеризируемых обьектов
g = 5 #кол-во кластеров

def ugen():
    u_matrix = np.random.rand(g,N)
    u_matrix = u_matrix / u_matrix.sum(axis = 0,keepdims = 1)
#keepdims returned array

    ln_u = np.log(u_matrix)  # matrix with ln(uij)
    pe = - np.sum(u_matrix* ln_u) / N
    return pe

i = 0
resdata = np.array([0])
while i < 100:
    resdata = np.append(resdata, ugen())
    i = i + 1

#print(resdata)

plt.plot(np.arange(101), resdata, ',')
plt.axhline(y=math.log(g))
plt.title(u'Значения энтропии разбиения для 100
сгенерированных матриц разбиения')

plt.ylabel(u'Значение энтропии')
plt.xlabel(u'Номер примера')
plt.axis([0, 100, 0, math.log(g) + 1])
plt.show()
```

Результати:



Графік демонструє те, що значення ентропії для 100 сгенерованих мариць розбиття лежить в межах від 0 до ln *g.*