Data Science Tools and Software
Model Answer
Assiment #2
Dr. Mohamed Abdelhafeez

Question 1: Data Preprocessing

a) Given the following dataset

i. Compute the Euclidian distance d_e (x_1, x_3) and d_e (x_2, x_4)

de(x1,x3) = sqr_root((85-80)^2+(0.7-0.2)^2) =5.024

de(x2,x4)= sqr_root((65-75)^2+(0.8-0.9)^2) = 10

ii. Comment on the computed distances above

iii. Normalize the given dataset using min-max

	math	physics
x1	85	0.7
x2	65	0.8
x3	80	0.2
x4	75	0.9

Answer

$x' = (x - \min(x))/(\max(x) - \min(x))$

	math	physics
x1	1	0.71
x2	0	0.86
x3	0.75	0
x4	0.5	1

b) Given the following dataset X with missing values denoted a and b

x_1=[a? 60] x_2=[11 75] x_3=[5 75] x_4=[5 80] x_5=[7 b?]

Show how to replace the missing data denoted a and b with proper values using each of the following methods:

i. The mean value

a = (11+5+5+7)/4 = 7

b = (60+75+75+80)/4 = 72.5

ii. The most probable

a =5

b=75

iii. kNN regression with k=2.

Distance(x1,x2)= |60-75|= 15

Distance(x1,x3)=15

Distance(x1,x4)=20

Nearest neighbors for x1 are x2 and x3

a=(11+5)/2=8

distance(x5,x1)=∣7−8∣=1

distance(x5,x2)=∣7−11∣=4

distance(x5,x3)=∣7−5∣=2

distance(x5,x4)=∣7−5∣=2

Nearest neighbors for x5 are x3 and x4.

b=(75+80)/2=77.5

c) Calculate a normalized dissimilarity (distance) between the following two symbolic objects *x* and *y* having 4 attributes where the first attribute is a string of 5 characters, the second is an interval, the third is a set and the fourth is a binary number of 5 bits as follows:

x = [“abcdg” 10:15 {a,b,c} 11100] and y = [“abcef” 10:30 {d,c,e} 01001]

1- Dissimilarity for the String Attribute

x="abcdg"

y="abcef"

Hamming distance = 2 mismatches.

Normalized dissimilarity = 2/5=0.4.

2- Dissimilarity for the Interval Attribute

|Midpoint(x)-Midpoint(y)| / range of combined intervals

|12.5-20|/20=0.375

3- Dissimilarity for the Set Attribute

Jaccard Dissimilarity= 1- (∣x∩y∣ / ∣x∪y∣)

Jaccard dissimilarity=1−1/5=1−0.2=0.8

4- Dissimilarity for the Binary Attribute
Hamming distance = 3 mismatches
Normalized dissimilarity = 3/5 =0.6

Total Dissimilarity=40.4+0.375+0.8+0.6=42.175=0.54375

Question 2) Feature Extraction

Given the following term frequencies in a corpus D that contains 3 documents D1..D3

Document 1 (D1)	
Term	Term Count
Caw	2
Sudan	1
Camel	1

Document 2 (D2)	
Term	Term Count
Sudan	3
Caw	2
Nile	1

Document 3 (D3)	
Term	Term Count
Egypt	2
Nile	2
Caw	1

a) Build a dataset matrix of size 3 objects (documents) by 5 attributes (terms) using binary term frequency.

	Caw	Sudan	Camel	Nile	Egypt
D1	1	1	1	0	0
D2	1	1	0	1	0
D3	1	0	0	1	1

b) Create a distance matrix using squared Euclidian distance.

	D1	D2	D3
D1	0	2	4
D2	2	0	2
D3	4	2	0

c) Identify the first nearest neighbour of the document D3 using hamming distance

D2

Question 3 Mongo DB

1. What is MongoDB?
    - A. Relational database
    - B. Document-oriented database
    - C. NoSQL database
    - [✓ D. Both B and C]
2. In MongoDB, what is a document equivalent to in a SQL database?
    - Table
    - [✓ Record]
    - Field
    - Column
3. Which method is used to insert a single document into a MongoDB collection using PyMongo?
    - add_one()
    - insert_single()
    - [✓ insert_one()]
    - add_document()
4. What is the purpose of the PyMongo package in Python with respect to MongoDB?

    □ A. Web development
    □ B. Data visualization
    [✓ □ C. MongoDB driver for Python]
    □ D. Machine learning
5. In MongoDB, what does RUD stand for?

    [✓ □ A. Create, Retrieve, Update, Delete]
    □ B. Connect, Read, Update, Delete
    □ C. Collect, Retrieve, Use, Delete
    □ D. Create, Read, Upload, Delete
6. How do you update a document in MongoDB using PyMongo?

    □ A. update_single()
    □ B. modify_one()
    [✓ □ C. update_one()]
    □ D. change_document()
7. In PyMongo, what does the $set operator do in the context of updating a document?

    □ A. Sets the document to null
    □ B. Adds a new field to the document
    [✓ □ C. Updates a specific field in the document]
    □ D. Sorts the document in ascending order
8. Which method is used to delete a single document from a MongoDB collection in PyMongo?

    [✓ □ A. delete_one()]
    □ B. remove_single()

- □ C. erase_one()
- □ D. discard_one()

9. What is the purpose of the sort() method in MongoDB when using PyMongo?
   - □ A. Group documents in a collection
   - □ B. Filter documents based on a condition
   - □ C. Order the result in ascending or descending order
   - □ D. Limit the number of documents returned

Question 4 Text Analysis

Given the following term frequencies in a corpus D that contains 3 documents D1..D3, answer the following questions 1 to 6 :-

Document 1 (D1)	
Term	Term
Caw	2
Sudan	1
Camel	1

Document 2 (D2)	
Term	Term
Sudan	3
Caw	2
Nile	1

Document 3 (D3)	
Term	Term Count
Egypt	2
Nile	2
Caw	1

1. The resulting data matrix will be of size
   a) 3×5 b) 4 × 4 c) 5×5 d) 5×4
2. The normalized term frequency of tf (“camel”,D1) is
   a) 0.20 b) 3 c) 4 d) 0.25
3. The inverse document frequency idf(“Camel”,D)
   a) 3 b) 1 c) 1/3 d) 0
4. what is the tflogidf(“caw”,D)
   a) 0 b) 1 c) 3 d) 5
5. The resulting distance matrix will be of size
   a) 3×5 b) 4 × 4 c) 5×5 d) 3×3
6. The corresponding feature vector of document D1 using binary term frequency is
   a) [1 1 1 0 0] b) [1 0 0 0 1] c) [1 0 1 1] d) [2 1 1]