DATA 220 Mathematical Methods for Data Analytics – Homework -3
Deadline – 11.59 PM – 11/06/2023
20 pts

Problem 1 (**1*3 = 3 pts): Determine whether the following transformation is linear or non-linear. Show all steps.**

a. $T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x-y \\ x+y \\ 4x \end{bmatrix}$

b. $T(x,y) = (3x+2y, 4x-y)$

c. $T(a,b) = (x^2, y^2)$

Problem 2 (**1+ 2+ 1+1+1 +1= 6 pts) (coding):** Use doc_similarity.txt data that includes 10 documents (a sentence or a paragraph) for this task. First preprocess the data using tokenization (you can use x.split() in python)and lower casing (can use x.lower()) the documents. Create a term-frequency document matrix and calculate document similarity matrix using 1. Euclidean Distance, 2. Manhattan Distance, and 3. Cosine Similarity. For this task, you must not use any direct package for creating term-frequency matrix like CountVectorizer. The output of this question should be three matrices. And mention which document is the most similar document to document 1 using the three different distance formula.

Problem 3 (**9 pts**) (Coding): Perform principal component analysis on the breast cancer dataset. You must **not** use any direct package **like PCA for this** analysis. You can use packages to calculate eigenvectors, eigenvalues, and covariance matrix. You can download the dataset from sklearn. For PCA (a-d) you should not use the target variable. However, you will need to use the target variable in the question e.

```
from sklearn.datasets import load_breast_cancer
cancer = load_breast_cancer()
```

a. Find the mean centered dataset. **1 pts**
b. Find eigenvalues and associated eigenvectors of the covariance matrix of the mean centered dataset. **2 pts**
c. Plot percent of variance explained by each of the principal components (Scree Plot) and explain. **2 pts**
d. What is the optimal number principal components you will retian for reduce the dimension of the original data? Give proper explanation **2 pts**
e. Create a scatter plot with each of the observations of the dataset projected onto the first two principal components and use a different color for each group and include a legend. **2 pts**

Problem 4 (1+.5+.5 =2 pts) (Coding): In problem 4, you have used covariance matrix to find the variability of the mean centered data. Now, apply correlation matrix (normalized version of the covariance matrix) instead of covariance matrix and plot the scree plot principal components. Finally, compare the two scree plots.

You are required to submit:

1. An MS/PDF/Scanned document:
    a. Include all the steps of your calculations.
    b. Attach screenshots of the code output.
2. Source code:
    a. Python (Jupyter Notebook)
    b. Ensure it is well-organized with comments and proper indentation.

- Failure to submit the source code will result in a deduction of 5 points.
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.