BIN2023R01 – INTRODUCTION TO DATAMINING & MACHINE LEARNING FOR BIOINFORMATICS

Lab Exercise 2- Similarity Measures and Principal Components Analysis

Aim: To perform Similarity measures and Principal Components Analysis for various datasets.

To do in the lab:

Write a python code:

1. using the packages numpy, pandas, sklearn (datasets, cosine_similarity, decomposition, PCA, load_iris, preprocessing, StandardScaler, OneHotEncoder), scipy (euclidean), matplotlib

2. Using the inbuilt iris dataset calculate dot product similaroty, cosine similarity, and euclidean similarity measures for any two variables.

3. For the given nutritional dataset, perform PCA. Plot the results as a biplot. Write your interpretations as to what PC's to be considered further.

Write the answers to the following questions in your notebook

Questions:

1. Compare and contrast the results obtained from dot product similarity, cosine similarity, and Euclidean distance in the provided code. Under what circumstances might each measure be more suitable?

2. How would you modify the code to compute similarity measures between multiple pairs of vectors, rather than just a single pair?

3. What is the primary goal of Principal Component Analysis (PCA)? In the context of PCA, what are principal components, and how are they related to the original features of the data?

4. Why is standardization of data recommended before applying PCA?

6. Explain the concept of explained variance in the context of PCA. How is it calculated, and what does it represent?

8. How does the one-hot encoder differ from the standard scaler, and what is its purpose?

9. What role does a biplot serve, and what do the length, angle, and direction of the arrows signify in the plot?

Soft copy deadline: 5th February 2024 11:59PM

Hard copy deadline: 6th February2024 3:15PM