BT3041 – Analysis and Interpretation of Biological Data

Assignment 2

Download Microarray gene expression data set from Khan et al., 2001 from this website (http://bioinf.ucd.ie/people/aedin/R/full_datasets/khan_train.csv). This dataset contains 2308 rows x 64 column. Khan et al., 2001 used cDNA microarrays to study the expression of genes in of four types of small round blue cell tumours of childhood (SRBCT). These were neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma, a subset of non-Hodgkin lymphoma (BL), and the Ewing family of tumours (EWS). Gene expression profiles from both tumour biopsy and cell line samples were obtained and are contained in this dataset. Perform the below mentioned agglomerative clustering of NB, RMS, BL and EWS:

1. Using Euclidean distance and complete linkage
2. Using Euclidean distance and average linkage
3. Using Euclidean distance and simple linkage
4. Using Euclidean distance and centroid linkage
5. Using Manhattan distance and complete linkage
6. Using Manhattan distance and average linkage
7. Using Manhattan distance and simple linkage
8. Using Manhattan distance and centroid linkage

Compare and contrast how the clustering changes based on distance metric and linkage metric and comment your interpretation.

Please provide your analyses results in a report form, specifically answering each of the above 8 questions with relevant dendrograms, etc. Also, do state any assumptions made clearly in the report. Attach the Google drive link to your software codes (MATLAB/Python) used for performing calculations with the report.

Submission due: 27/04/2023 11:59:59 PM.