Course: *MATH 320- Machine Learning* – Professor: *Dr. Osei Tweneboah*

Homework #1

Instructions: While discussion with classmates is allowed and encouraged, please try to work on the homework independently and direct your questions to me.

Part A– Does not involve Python Programming

Instructions: Please show detailed work to receive full credit. Guesses receive no credits.

1. The distances between pairs of five items are given below:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0 \end{pmatrix} \end{array}$$

   Cluster the five items using each of the following procedures.

   (a) Single linkage hierarchical procedure.

   (b) Complete linkage hierarchical procedure.

   (c) Average linkage hierarchical procedure.

   Draw the dendrograms and compare the results in (a), (b), and (c).

Part B– Involves Python Programming

Unsupervised techniques are often used in the analysis of genomic data. This project will illustrate the Hierarchical and K-means clustering.

The file `NCI60_data.csv` is a 64 by 6830 matrix of the expression values, while the file `NCI60_labs.csv` is a vector listing the cancer types for the 64 cell lines.

1. Given the provided datasets (CSV files), load them in Python.
   (a) How many observations and features are in the dataset?
   (b) Are there any missing values?
2. Data preprocessing: Check to see if the data is standardized. If not, standardize the data matrix X so all variables are given a mean of zero and a standard deviation of one.
3. Add an index name as “cancer type" to nci_data and make it appear as “cancer type" when clustering is applied. (Cancer type is in file `NCI60_labs.csv`).
4. K-Means Clustering:
   (a) Perform K-means clustering and set `random_state=123` and `n_init=150`.
   (b) Create a cross table/contingency_table (use library `pd.crosstab`) to see how it is clustered according to the type of cancer.
   Hint: Use as column labels the `kmean.labels_`
5. Hierarchical Clustering:
   (a) Now apply SciPy’s Hierarchical Clustering to all features. Use ’Euclidean’ as the Distance method and comment on how clustering differs according to complete, single, and average linkage.
   (b) Which linkage produced better results?
   (c) Using the linkage that produced “better" clustering results, cut the dendrogram at the height that will yield a particular number of clusters. How many clusters are obtained?
   (d) Create a cross table/contingency_table (use library `pd.crosstab`) to see how it is clustered according to the type of cancer.
   Hint: Use the `fcluster` function from the scipy.cluster.hierarchy to first see what samples are assigned to the clusters. Use this as the column label for the cross table. In the `fcluster` function set `criterion='maxclust'`