

## **EXPERIMENT 8**

### **Aim :**

Develop a Machine Learning Method to Cluster Gene Expression Data.

### **Introduction :**

Almost every cell in the body of an organism has the same DNA. Genes are portions of this DNA that code for proteins or (less commonly) other large biomolecules. A gene is expressed through a two-step process in which the gene's DNA is first transcribed into RNA, which is then translated into the corresponding protein. A novel technology of gene-expression microarrays whose development started in the second half of the 1990s and is having a revolutionary impact on molecular biology allows one to monitor the DNA-to-RNA portion of this fundamental biological process.

### **Implementation :**

```
import scipy.cluster.hierarchy as hier
import pylab as pl

def fcluster( pts, ncluster, method = "average", criterion="maxclust" ):
    """ -> ( pts, Y pdist, Z linkage, T fcluster, clusterlists )
    ncluster = n1 + n2 + ... ( including n1 singletons )
    av cluster size = len( pts ) / ncluster
    """

    pts = np.asarray( pts )
    Y = scipy.spatial.distance.pdist( pts ) # ~ N^2 / 2
    Z = hier.linkage( Y, method ) # N-1
    T = hier.fcluster( Z, ncluster, criterion = criterion )
    # clusters = clusterlists( T )
    return ( pts, Y, Z, T )

hier.dendrogram( Z )
```