HW5 (Clustering)

Problem 1 (WEKA – Hierarchical and K-means)
You are given the dataset cancer.arff. It is a medical dataset containing 63 instances of persons having cancer. Attribute class registers the type of tumor (1=euroblastoma, 2=rhabdomyosarcoma, 3=non-Hodgkins lymphoma and 4=Ewing family of tumors). The interesting part of this dataset are the remaining 96 attributes that correspond to the 96 most important genes for diagnosing cancer.

Use WEKA to compare the quality of different clusterings when we ask for 4 clusters and we use classes to clusters evaluation. You will compare K-means (try 5 different random seeds) and Hierarchical (try all methods for measuring cluster distance, or linkType: SINGLE, COMPLETE, AVERAGE, etc).

For K-means, try more random seeds until you get the optimal clustering. Comment on the results: which algorithm variation is the best for the specific dataset?

Problem 2 (DBSCAN)
In the figure below, you are given 12 points with integer coordinates and names 1 through 12.

1. Apply DBSCAN by hand with eps=1.9 (use Euclidean distance) and minpts=4.
2. Apply DBSCAN using WEKA with the same parameter values. You must create a two column csv file with attribute names x and y having the coordinate values of the points you see in the figure. **IMPORTANT**: WEKA will normalize the values, thus, you should add points (0,0) and (10,10) in your dataset so that the normalized values equal to the 1/10 of the original ones. In other words, point (3,6) will become (0.3, 0.6). Obviously, eps should be set to 0.19. In addition, WEKA considers neighborhood of a point p all points q where distance(p,q) < eps and not distance(p,q) <= eps. In our case, this makes no difference..