Homework #4

Clustering

Due Date: November 19th, 2015, end of the day.

1. Consider the following set of data points in a 2-D data space: (6, 12), (19, 7), (15, 4), (11, 0), (18, 12), (9, 20), (19, 22), (18, 17), (5, 11), (4, 18), (7, 15), (21, 18), (1, 19), (1, 4), (0, 9), (5, 11). Answer the following questions in the context of these data points.
   1. Perform basic sequential clustering with these data points using the threshold “theta” value of 12 and the maximum number of clusters as four. Use the numbers from the first to the last. Show the data points on a 2-D plot with boundaries drawn around sets of points to show clusters.
   2. Do the same as in (a) but use the data points in the reverse order: from the last one to the first one.
   3. Use Rand index to find the difference between the two clusterings obtained in (a) and (b). Informally, identify the cluster in (a) that has been altered the most in clustering done in (b). Give informal explanation of why this cluster broke apart the most.
2. Use the same data points as in #1 above and perform the following:
   1. Perform hierarchical clustering of these points using single-link method and show the resulting dendrogram. What will be the three clusters identified in the data? Show them on a 2-D plot with boundaries drawn around clusters.
   2. Do the same as in (a) above but use complete-link method for clustering.
   3. Compute the sum of squared errors for the clusterings obtained in (a) and (b) above. Compare and comment on these two values. In each clustering identify the cluster that is contributing the most to the SSE value.
   4. Use the correlation analysis to determine which of these two clusterings (obtained in (a) and (b)) has higher correlation. Construct the proximity matrix for the points and then build a binary matrix marking 1’s wherever the point-pairs are in the same cluster. Find correlation between the two matrices. (Consult the slides used in class to see details). Comment on the correlation values obtained and your intuitive idea about the clusters in the dataset.
3. Consider the following set of 1-D data points: 1, 3, 5, 6, 8, 11, 12, 13, 14, 15, 16, 22, 28, 32, 33, 34, 35, 36, 37, 42, 58.
   1. Use DBSCAN algorithm to mark each point as Core, Border, or Noise. Use these markings to perform density based clustering of these points. On a number line show the clusters obtained. Use an epsilon value of 4 and Minpoints value of 3.
   2. Do the same as in (a) above but use Epsilon value of 6.
   3. Compare the two clusterings using Rand Index and show all your work.