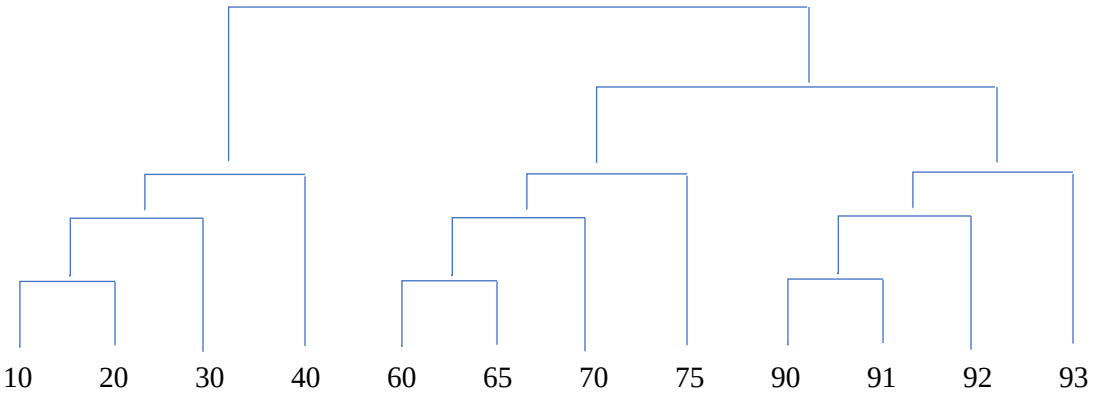
Roy Chan
chanroyc@sfu.ca
301202770

CMPT 459 Fall 2017
Data Mining
Martin Ester
TA: Zhilin Zhang

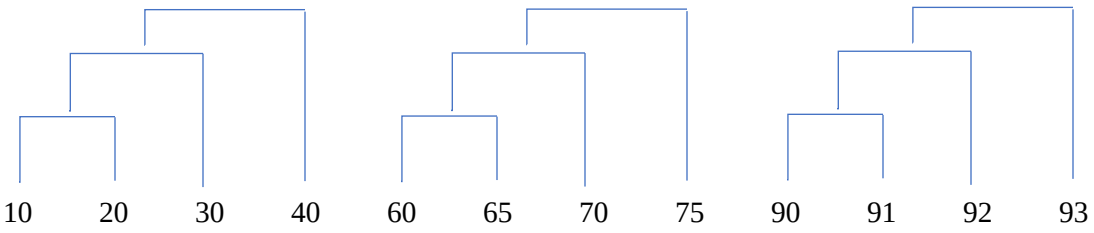
Assignment 2

Assignment 2.1

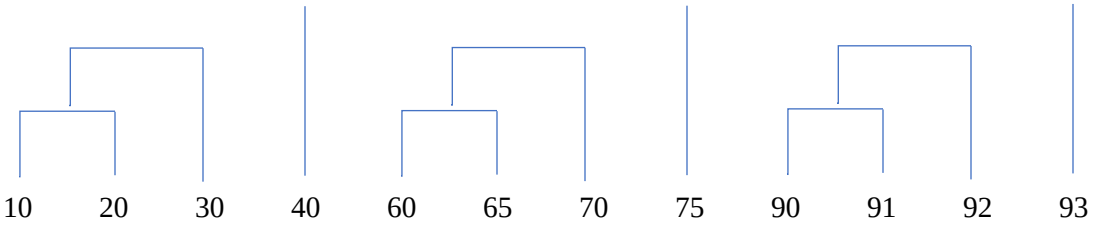
(a)



(b)



(c)



Roy Chan
chanroyc@sfu.ca
301202770

(d)

The purity of a class C is computed as the percentage of elements in C belonging to the dominant cluster, where the dominant cluster is the one among the resulting clusters that has the maximum overlap with C, where C1 = {10, 20, 30, 40}, C2 = {60, 65, 70, 75}, C3 = {90, 91, 92, 93}. In (c), we have C1 = {10, 20, 30}, C2 = {40}, C3 = {60, 65, 70}, C4 = {75}, C5 = {90, 91, 92} and C6 = {93}. The purity of the classes C1, C2 and C3, with regards to C1, C2, C3, C4, C5 and C6 is 75%.

Assignment 2.2

Let o be a core object in cluster C, and point P density reachable from point o with regard to Eps, Minpts. If point P is in neighbourhood $N_{EPS}(o)$, then $N_{EPS}(o)$ must be larger than MinPts by definition, since MinPts is the least amount of points within the neighbourhood. The statement must be true initially in order for the neighbourhood to be a valid neighbourhood. P is density reachable from point o if there is a chain of points P1, … , Pn, P1 = o, Pn = P such that Pi+1 is directly density reachable from P1.

Assume that there is a core object in o not in cluster C, the set of objects density-reachable from o is equal to C. Then by definition, if the set of objects where o is a core object is density reachable, then the distance will be the sum of distance between o, P1, … , Pn > C, which refutes the initial assumption that the set of objects density-reachable from is equal to c. Then o must be a core object in cluster C where the set of objects are density-reachable from o is equal to C.

Assignment 2.3

The original agglomerative Hierarchical clustering is as follows. Agglomerative Hierarchical clustering forms initial clusters consisting of a singleton object, and computes the distance between each pair of clusters. Merge clusters which have the minimum distance, then calculate the distance between the new cluster and all other clusters. If there is only one cluster containing all objects, then the agglomerative hierarchical clustering is complete.

In semi-supervised clustering, user can provide domain knowledge in the form of must-link and cannot-link constraints on pairs of objects where must-links, as the name suggests, are pairs of objects that must be linked together. As opposed to cannot-links, as the name suggests, are pairs of objects that cannot be linked together.

The variant agglomerative hierarchical clustering should first form initial clusters consisting by applying the must-link pairs of objects first. Then the algorithm can form initial clusters consisting of a singleton object, and compute the distance between each pair of clusters. Once a pair of clusters is chosen with the minimum distance, the algorithm can check a list of cannot-links. If the pair of objects cannot be linked, then continue searching for initial clusters consisting of singleton objects, and compute the distance between each pair of clusters. Repeat until there is only one cluster containing all objects, where pairs of objects from must-links will be neighbours, whereas pairs of objects from cannot-links will not be neighbours.

All we need to do in the variant is to force the must-links to link together first, proceed with the agglomerative hierarchical clustering, with the addition of checks for cannot-links of pairs of objects.