1 Flat Clustering

Download the `clustering.ipynb` notebook and the provided datasets on Toledo. In this notebook, you will implement a distance-based clustering algorithm: k-means. You will run k-means on a small dataset, and compare the algorithm to a model-based algorithm: EM clustering. As a more fun exercise, you will also apply your k-means implementation on images of football players.

2 Hierarchical Clustering

2.1 Agglomerative Clustering

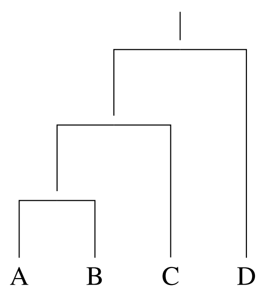
Apply agglomerative clustering with (1) single linkage and (2) complete linkage to a dataset with four points A, B, C, D with the distance matrix below. Also draw the corresponding dendrograms.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

- *Single linkage:* The distance between two clusters is the minimal distance for all pairs of points x_1, x_2 where x_1 is in cluster 1 and x_2 is in cluster 2. To cluster the four points, we start by creating one cluster for each point. That is, we have four clusters: {A}, {B}, {C}, {D}. We then merge pairs of clusters for which the single link distance is minimal, until we obtain one cluster that contains all four points:

  1. The closest pair of clusters is {A} and {B} with distance 1. Merge them into one cluster: {A, B}. The new clustering is {A, B}, {C}, {D}
  2. The closest pair is now {A, B} and {C} with distance 2 (the closest pair of points is B and C). The new clustering is {A, B, C}, {D}
  3. The closest pair is now {A, B, C} and {D} with distance 3. The final clustering is {A, B, C, D}

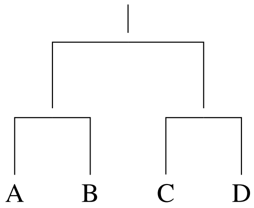
  The corresponding dendrogram is the following:



- *Complete linkage:* The distance between two clusters is the maximal distance for all pairs of points x_1, x_2. We again start from four clusters {A}, {B}, {C}, {D} and merge the clusters as follows:

  1. The closest pair of clusters is {A} and {B} with distance 1. Merge them into one cluster: {A, B}. The new clustering is {A, B}, {C}, {D}
  2. The closest pair is now {C} and {D} with distance 3. The new clustering is {A, B}, {C, D}
  3. The closest pair is now {A, B} and {C, D} with distance 6. The final clustering is {A, B, C, D}

The corresponding dendrogram is the following:



2.2 Integration with Distance-Based Clustering

(a) For the following two sets of points, write down the cluster feature (CF) that summarizes each set:

1. {(1,2), (2, 3), (3, 2), (2, 1)}: $\{4, (8, 8), (18, 18)\}$
2. {(2,4), (4, 3), (3, 4), (2, 2)}: $\{4, (11, 13), (33, 45)\}$

(b) Given the following cluster feature $\{5, (25, 30, 20), (147, 190, 96)\}$, do the following:

1. Compute the centroid of the cluster.

$$\left(\frac{25}{5}, \frac{30}{5}, \frac{20}{5}\right) = (5, 6, 4)$$

2. Compute the radius of the cluster.

$$\sqrt{\frac{147 + 190 + 96 - 2 \cdot (5, 6, 4) \cdot (25, 30, 20)}{5} + (5, 6, 4) \cdot (5, 6, 4)} = \sqrt{\frac{433 - 2 \cdot 385}{5} + 77} = 3.10$$

3. Give the Manhattan distance of the point $(3, 8, 2)$ to the centroid.

$$2 + 2 + 2 = 6$$

(c) Construct a CF tree using BIRCH for the following points: $\{(3, 4), (4, 5), (7, 4), (8, 4), (4, 7), (1, 1)\}$.
Use a radius threshold of 1.5 and branching factor of 2 for both leaf and non-leaf nodes.
The following shows the CF tree after adding each point:

- (3, 4)
  Create the top node of the tree with CF $\{1, (3, 4), (9, 16)\}$.

{1, (3, 4), (9,16)}

- (4, 5)
  Add the point to the CF of the top node. The new CF is $\{2, (7, 9), (25, 41)\}$ *with radius* 0.71*, which is smaller than the threshold* 1.5*, so there is no need to split the cluster.*

{2, (7, 9), (25,41)}

- (7, 4)
  Add the point to the CF of the top node. The new CF would be $\{3, (14, 13), (74, 57)\}$ *with radius* 1.76*, which exceeds the threshold. Hence, we do not add the point to this subcluster, but add it as new entry* $\{1, (7, 4), (49, 16)\}$ *to the root node.*

{2, (7, 9), (25,41)}	{1, (7, 4), (49,16)}

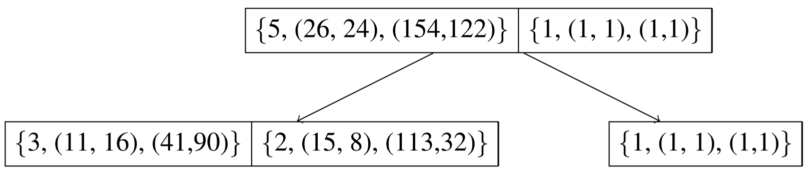
- (8, 4)
  To add the point to one of the clusters in the graph, we first need to find the closest cluster. The cluster with the smallest Manhattan distance is $\{1, (7, 4), (49, 16)\}$ *(the Manhattan distance between* $(8, 4)$ *and the centroid* $(7, 4)$ *is* 1*) so we add* $(8, 4)$ *to this cluster. The new CF then becomes* $\{2, (15, 8), (113, 32)\}$ *with radius* 0.5*, which is smaller than the threshold.*

{2, (7, 9), (25,41)}	{2, (15, 8), (113,32)}

- (4, 7)
  The closest cluster is $\{2, (7, 9), (25, 41)\}$*, so we add* $(4, 7)$ *to this cluster. The new CF becomes* $\{3, (11, 16), (41, 90)\}$ *with radius* 1.33*, which is smaller than the threshold.*

{3, (11, 16), (41,90)}	{2, (15, 8), (113,32)}

- (1, 1)
  The closest cluster is $\{3, (11, 16), (41, 90)\}$*. However, if we added the point to this cluster, the radius would become* 2.49*, which exceeds the threshold. Therefore, we should add this point as a new entry to the node. But the node already contains* 2 *entries, which is the maximum. Therefore, we need to split the root node. How exactly is the split made? We take the two CF fartherst apart, namely* $\{1, (1, 1), (1, 1)\}$ *and* $\{2, (15, 8), (113, 32)\}$*, as seeds. The third cluster,* $\{3, (11, 16), (41, 90)\}$*, is closer to the latter one, and they are therefore grouped together in the same node.*
  Note that when we split the root node, the height of the tree increases by 1*.*



The resulting clustering consists of three clusters: $\{(3, 4), (4, 5), (4, 7)\}$*,* $\{(7, 4), (8, 4)\}$ *and* $\{(1, 1)\}$*.*

3 General Questions

1. What does it mean that EM is a model-based clustering approach?
   That there are explicit models for each cluster and we can check the likelihood that an instance was generated by that model, i.e., that it belongs to the particular cluster.

2. What are the differences between agglomerative and divisive clustering? What are their relative strengths and weaknesses?
   Bottom-up splits (agglomerative) vs top-down merges (divisive). Finding the "best" split is more expensive than finding the "best" merge. Different choices for merge (single linkage, complete linkage, average linkage) lead to different clusters.