CS540 Homework 3
Solution - Written part

Spring 2016

1 - Hierarchical Clustering (20 points)

a) (15 points)

(i) (5 points) Single linkage

Step	**Clusters**
0	{6} {8} {10} {20} {26} {30} {32} {33}
1	{6} {8} {10} {20} {26} {30} {32, 33}
2	{6, 8} {10} {20} {26} {30} {32, 33}
3	{6, 8, 10} {20} {26} {30} {32, 33}
4	{6, 8, 10} {20} {26} {30, 32, 33}
5	{6, 8, 10} {20} {26, 30, 32, 33}
6	{6, 8, 10} {20, 26, 30, 32, 33}
7	{6, 8, 10, 20, 26, 30, 32, 33}

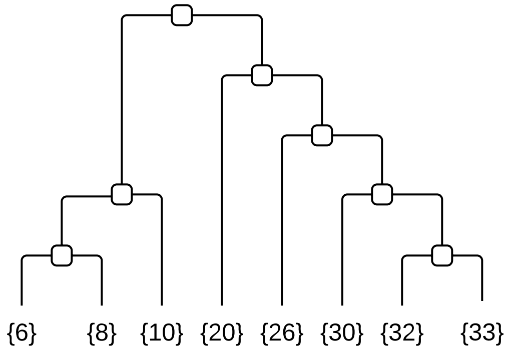
(ii) (5 points) Complete linkage

Step	**Clusters**
0	{6} {8} {10} {20} {26} {30} {32} {33}
1	{6} {8} {10} {20} {26} {30} {32, 33}
2	{6, 8} {10} {20} {26} {30} {32, 33}
3	{6, 8} {10} {20} {26} {30, 32, 33}
4	{6, 8, 10} {20} {26} {30, 32, 33}
5	{6, 8, 10} {20, 26} {30, 32, 33}
6	{6, 8, 10} {20, 26, 30, 32, 33}
7	{6, 8, 10, 20, 26, 30, 32, 33}

(iii) (5 points) Average linkage

Step	**Clusters**
0	{6} {8} {10} {20} {26} {30} {32} {33}
1	{6} {8} {10} {20} {26} {30} {32, 33}
2	{6, 8} {10} {20} {26} {30} {32, 33}
3	{6, 8} {10} {20} {26} {30, 32, 33}
4	{6, 8, 10} {20} {26} {30, 32, 33}
5	{6, 8, 10} {20} {26, 30, 32, 33}
6	{6, 8, 10} {20, 26, 30, 32, 33}
7	{6, 8, 10, 20, 26, 30, 32, 33}

b) (2 points) Binary tree for single linkage clustering looks as follows:



c) (3 points) One way to determine the best number of final clusters is to try several different values, e.g. $k = 1, 2, 5, 10, 20, 50, ...$, and for each value of k calculate the distortion of the final clustering. The lower the distortion, the better are instances grouped around the centroid that they are assigned to, the better is the final clustering.

However, we also have to have a mechanism to penalize large values of k. Otherwise we will end up with k being equal to the number of training instances where each instance is assigned to a different centroid. In that case the distortion will be equal to 0 but we cannot call such clustering "good".

2 - k Nearest Neighbor (10 points)

a) (6 points) Classification.

Three nearest neighbors for instance [100, 50, 7, ?] are

1. $[90, 52, 7, 10]$
2. $[130, 69, 9.5, 20]$
3. $[63, 51.5, 6.5, 10]$

Hence, the predicted label for this test instance will be $Age = 10$.

Three nearest neighbors for instance [120, 90, 9, ?] are

1. $[130, 69, 9.5, 20]$
2. $[145, 70, 11, 20]$
3. $[160, 69.5, 10, 20]$

Hence, the predicted label for this test instance will be $Age = 20$.

b) (4 points) Regression.

Three nearest neighbors for our test instances stay the same as in part a).

The label for instance [100, 50, 7, ?] is $Age = (10 + 20 + 10)/3 = 13.3$.

The label for instance [120, 90, 9, ?] is $Age = (20 + 20 + 20)/3 = 20.0$.