# CSE 6363 - *Machine Learning*

Homework/Project 3- Spring 2021

Due Date: May. 03 2021, 3:00 pm

## Hierarchical Clustering

1. Consider an unlabeled version of our height/weight/age data set used in the early assignments (and shown below).

$$
\begin{aligned}
D = \{ \quad & (170, 57, 32), \\
& (190, 95, 28), \\
& (150, 45, 35), \\
& (168, 65, 29), \\
& (175, 78, 26), \\
& (185, 90, 32), \\
& (171, 65, 28), \\
& (155, 48, 31), \\
& (165, 60, 27), \\
& (182, 80, 30), \\
& (175, 69, 28), \\
& (178, 80, 27), \\
& (160, 50, 31), \\
& (170, 72, 30) \quad \}
\end{aligned}
$$

   a) Apply hierarchical clustering with single (minimum) linkage to this data and show the resulting cluster hierarchy. Indicate the order of the merge operations and the distance (linkage) value between the merged sets at each merge. You can do this on the cluster hierarchy tree if you want. You can use an existing implementation for hierarchical clustering as well as for the display of the cluster hierarchy tree.

   b) Repeat the clustering using complete (maximum) linkage. Again, make sure you indicate the linkage value for every cluster merge.

   c) For both cluster results, evaluate how well the clusters formed using both linkage criteria fit the original class labels (indicated in the corresponding dataset below). For this, indicate for the cases of 2, 3, and 4 clusters how many errors would occur if in each of the clusters the most frequent class label would be used to predict the class for all elements. Which of the two linkage criteria better reflects the original classes ?

$$
\begin{aligned}
D = \{ \quad & ((170, 57, 32), \quad W), \\
& ((190, 95, 28), \quad M), \\
& ((150, 45, 35), \quad W), \\
& ((168, 65, 29), \quad M), \\
& ((175, 78, 26), \quad M), \\
& ((185, 90, 32), \quad M), \\
& ((171, 65, 28), \quad W), \\
& ((155, 48, 31), \quad W), \\
& ((165, 60, 27), \quad W), \\
& ((182, 80, 30), \quad M), \\
& ((175, 69, 28), \quad W), \\
& ((178, 80, 27), \quad M), \\
& ((160, 50, 31), \quad W), \\
& ((170, 72, 30), \quad M), \quad \}
\end{aligned}
$$

## Self-Training

1. Consider the following linearly separable training data set:

$$
\begin{aligned}
D_s = \{ \quad & ((170, 57, 32), \quad W), \\
& ((190, 95, 28), \quad M), \\
& ((150, 45, 35), \quad W), \\
& ((168, 65, 29), \quad M), \\
& ((175, 78, 26), \quad M), \\
& ((185, 90, 32), \quad M), \\
& ((171, 65, 28), \quad W), \\
& ((155, 48, 31), \quad W), \\
& ((165, 60, 27), \quad W) \quad \}
\end{aligned}
$$

$$
\begin{aligned}
D_u = \{ \quad & (182, 80, 30), & (175, 69, 28), & (178, 80, 27), \\
& (160, 50, 31), & (170, 72, 30), & (152, 45, 29), \\
& (177, 79, 28), & (171, 62, 27), & (185, 90, 30), \\
& (181, 83, 28), & (168, 59, 24), & (158, 45, 28), \\
& (178, 82, 28), & (165, 55, 30), & (162, 58, 28), \\
& (180, 80, 29), & (173, 75, 28), & (172, 65, 27), \\
& (160, 51, 29), & (178, 77, 28), & (182, 84, 27), \\
& (175, 67, 28), & (163, 50, 27), & (177, 80, 30), \\
& (170, 65, 28) \ \}
\end{aligned}
$$

a) Implement a self-training system using a logistic regression classifier for this problem. You need to implement the self-training approach but can use either your previous or an existing library implementation for the Logistic Regression algorithm.

b) Learn a classifier using the semi-supervised learning algorithm and compare it agains a classifier learned only from the labeled data $D_s$ using the following test set:

$$
\begin{aligned}
D_t = \{ \quad & ((169, 58, 30), \quad W), \\
& ((185, 90, 29), \quad M), \\
& ((148, 40, 31), \quad W), \\
& ((177, 80, 29), \quad M), \\
& ((170, 62, 27), \quad W), \\
& ((172, 72, 30), \quad M), \\
& ((175, 68, 27), \quad W), \\
& ((178, 80, 29), \quad M) \quad \ \}
\end{aligned}
$$