

CSE 6363 - *Machine Learning*

Homework/Project 3- Spring 2019

Due Date: Apr. 30 2019, 11:59 pm

Hierarchical Clustering

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

$$D = \{ \begin{array}{l} ((170, 57, 32), \text{ W}), \\ ((192, 95, 28), \text{ M}), \\ ((150, 45, 30), \text{ W}), \\ ((170, 65, 29), \text{ M}), \\ ((175, 78, 35), \text{ M}), \\ ((185, 90, 32), \text{ M}), \\ ((170, 65, 28), \text{ W}), \\ ((155, 48, 31), \text{ W}), \\ ((160, 55, 30), \text{ W}), \\ ((182, 80, 30), \text{ M}), \\ ((175, 69, 28), \text{ W}), \\ ((180, 80, 27), \text{ M}), \\ ((160, 50, 31), \text{ W}), \\ ((175, 72, 30), \text{ M}), \end{array} \}$$

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

Self-Training

1. Consider the following linearly separable training data set:

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

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

- a) Implement a self-training system using a logistic regression classifier for this problem.
- b) Learn a classifier using the semi-supervised learning algorithm and compare it against a classifier learned only from the labeled data D_s using the following test set:

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