**Written Assignment-4: Instance-Based Learning and Clustering**

**Question 1: Cross-Validation (20 points)**

Below is a dataset with one input *x* and one binary output *y*. We will use k-NN with Euclidean distance to predict *y* for a given *x*.



**(a)** **(5 pts)** What is the leave-one-out CV error of 1-NN on this dataset? Give your answer as the total number of misclassifications.

**Ans:**

6. For each x, if the nearest neighbor has a different label, X would be misclassified.

**(b)** **(5 pts)** What is the leave-one-out error of 3-NN on this dataset? Give your answer as the total number of misclassifications.

**Ans:** 5, For each x consider the majority vote of 3 nearest neighbors.

**(c)** **(5 pts)** What is the 5-fold CV error of 1-NN on this dataset? Give the error rate. The five folds should be created as follows. 1st fold: -0.1 and 0.7; 2nd fold: 1.0 and 1.6; 3rd fold: 2.0 and 2.5; 4th fold: 3.2 and 3.5; 5th fold: 4.1 and 4.9.

**(d)** **(5 pts)** What is the 2-fold CV error of 3-NN on this dataset? Give the error rate. The first fold should contain the leftmost five points, and the second fold should contain the remaining five points.

**Question 2: Hierarchical Clustering (10 points)**

Draw the clustering tree for the following sample of ten points in one dimension

Sample = (−2.2, −2.0, −0.3, 0.1, 0.2, 0.4, 1.6, 1.7, 1.9, 2.0).

Use single-link clustering (i.e. *d*(*Ci*, *Cj*) = *minx* ∈ *Ci*, *x*′∈ *Cj* ||*x* – *x*′||) . According to the clustering tree, what is the natural number of clusters for this dataset? Justify your answer.

**Question 3: More Hierarchical Clustering (10 points)**

Using agglomerative clustering, what is the maximum height of the tree required to cluster a set of *n* points? What is the minimum height? Give an example of a set of points that requires a tree of maximum height.

**Question 4: K-means Clustering (15 points)**

****

One disadvantage of K-means is that one has to specify the value of *k*. Consider the following strategy for picking *k* automatically: try all possible values of *k* ∈ {1, 2, . . . , *n*} and pick the *k* that minimizes the distortion function. Explain whether this is a good / bad idea.