CMSC-478 Machine Learning - Spring 2018
Homework Assignment 1
Due at the start of class on February 20^{th}

(1) Decision Trees (20 points): Consider the following dataset with four binary attributes and one binary class label.

x_1	x_2	x_3	x_4	class
0	1	1	0	+
1	0	1	0	+
0	0	1	1	+
1	1	1	1	+
1	0	1	1	-
0	0	0	0	-
0	1	0	0	-
1	1	1	0	-

Use equation 3.4 from the Mitchell chapter on decision trees to compute information gain for each attribute to choose the root split for the tree. Give the computed information gain for each attribute and indicate which attribute should be used at the root of the tree.

Draw the full, unpruned tree that would be learned from this dataset. There is no need to do the full information gain computation for the splits below the root. Just "eyeball" the data and the correct splits should be obvious.

(2) Nearest Neighbors (20 points): Consider a dataset where for each instance $x = (x_1, x_2)$ and $y \in \{+, -\}$. In a 2D plot, let x_1 be the horizontal axis, and x_2 the vertical axis. The instances are as follows:

- + instances: (2, 2), (3, 4), (4, 6)
- - instances: (3, 0), (4, 3), (5, 5)

Plot these points and draw the decision boundary for the 1NN classifier. Then plot the points after multiplying x_1 in each instance by 5 and show the corresponding 1NN decision boundary. Explain qualitatively what the boundary approaches as the multiplier on x_1 gets larger and larger.

(3) k-means (20 points): Suppose you have a dataset in which the instances are 1-dimensional. The instances are $\{1, 2, 4, 5, 10, 11, 12, 25\}$. Run k-means clustering on this dataset for $k = 3$ with initial centroids on the first 3 instances in the dataset (i.e., 1, 2, 4). Draw the points on a number

line and show which points belong to which clusters for each interation of k-means. Run the algorithm until two consecutive iterations yield the same assignment of points to clusters.