

CSCI/ARTI 8950 Machine Learning

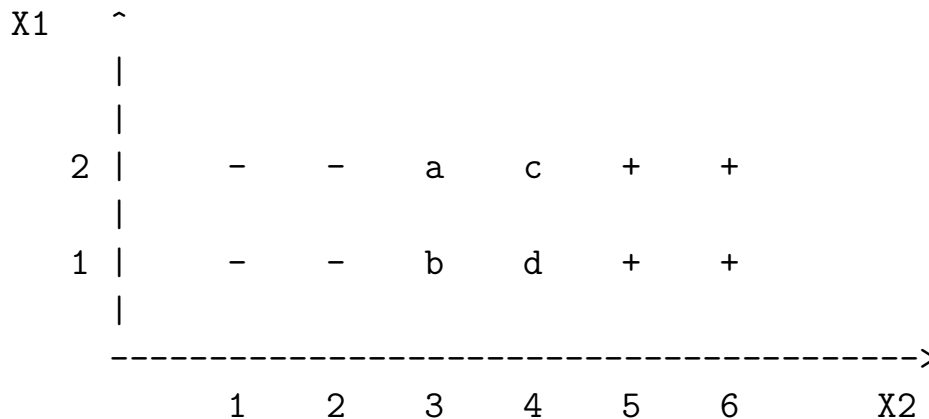
Assignment Number 2: Due 2/18/2022 (by eLC)

1. [20 points][MID] Consider the following training set of samples for machine learning:

Example	A1	A2	A3	A4	label
1	1	2	2	2	-
2	1	1	1	1	a
3	2	3	2	1	b
4	1	3	3	3	c
5	3	1	2	1	d
6	1	1	1	2	-

The attributes **A1** through **A4** are integers with values in the range [1,2,3] each.

- For the label assignment (a=-,b=+,c=-,d=+) give a minimal size (measured by the total number of nodes) decision tree that can correctly classify all the training examples.
 - How would the tree given in Part (a) above classify the following examples: (1,2,2,3) and (3,2,1,1)?
 - Propose a label assignment for a, b, c, and d that will make attribute **A4** better than attribute **A3** according to the ID3 information gain measure.
2. [20 points][MID] Consider the following diagram of a set of samples for machine learning:



- For the label assignment (a=-,b=+,c=+,d=-) can all the given samples (including a, b, c, and d) be correctly classified by the 3-nearest neighbor rule? This rule labels any sample with the majority label of its 3 nearest neighbors in Euclidean distance from the given samples (the sample itself is NOT considered one of the neighbors). If your answer is NO, briefly explain why.

- (b) For the label assignment ($a=+$, $b=-$, $c=+$, $d=-$) can all the given samples (including a, b, c, and d) be correctly classified by a properly trained (or computed in any possible way) binary decision tree with at most 2 levels (2 decisions along each path)? The decision at each level will be of the form $X_i \leq V$ where i is 1 or 2 and V is a variable threshold. If your answer is YES, sketch one such decision tree. If your answer is NO, briefly explain why.

3. [20 points][MID] Short answers please

- (a) Give one advantage for using 3-nearest neighbor over 1-nearest neighbor for classification.
- (b) Give one advantage for using 1-nearest neighbor over 3-nearest neighbor for classification.
- (c) Give one advantage to using the information GainRatio measure over the information Gain measure for constructing decision trees.
- (d) Give one advantage for using two-fold cross-validation over ten-fold cross-validation.
- (e) Give one advantage for using ten-fold cross-validation over two-fold cross-validation.