Ariel University

Machine Learning

Homework 4

1. In the vector data-set, each vector is binary, and has eight dimensions. The final value of the vector is its label, which is 0 or 1. Implement decision trees on the binary vectors data set:

   Given a parameter k, the decision tree will have up to k levels, and so at most 2^k-1 nodes. At each node, one can either decide that all vectors reaching this node will be labelled 0 or 1 (and then the node is a leaf), or one can split the node into two nodes, based on one coordinate in the vector set. Implement two different splitting strategies:

   A. Brute-force. Construct all possible trees of k levels, and choose the one that has the smallest error on the vector set.

   B. Binary entropy. Begin with a single root node, and split this node into two leaves based on the best coordinate. The best split is the one that minimizes the sum of the binary entropies of the two created leaves. Then recursively split the leaves, until reaching k levels.

   In both cases, a node which has only vectors of one label need not be split.

   Run the algorithms in A and B on the vector data-set with k=3. Return the error returned by each one, and draw the tree achieving this error.

2. Suppose we are given access to a random number generator U, which generates independent random real variables distributed uniformly in the range [0,1). Show that this can be used to produce an embedding satisfying the properties of the JL-lemma, with high probability.

   Now show the same given a random number generator C, which generates independent random real variables according to the Chi-squared distribution with 1 degree of freedom.