Ariel University

Machine Learning

Homework 3

For this assignment, we will use the Iris data set in the last homework. As in the last homework, we will use only Versicolor and Virginica, and only the second and third coordinates.

hand in your python code, and also hand in the output of the algorithm and the answers to the questions below in a separate file.

1. Implement k-nearest neighbor on the data set.

   A. Sample a training set with half the points. The remaining points are the test set.
   B. For each of k=1,3,5,7,9 and p=1,2,∞, evaluate the k-NN classifier on the test set, under the l_p distance. (The base set of the classifier is the training set.) Compute the classifier error on the training and test sets.
   C. Repeat steps (a) and (b) 100 times, and output the average empirical and true errors for each k and p. Also output the difference between them.

Which parameters of k,p are the best? Why is this?

How do you interpret the results? And is there overfitting?

2. Write decision tree algorithms for this 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** with k=3. (This tree has a root, its children, and leaf grandchildren.) Return the error returned by **A** and **B**, and draw the trees achieving this error. Analyze the results.