Ariel University

Machine Learning

Homework 4

For this assignment, 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 Haberman's Survival data set.[1]
    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?

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

2. Now run the same algorithm on the "circle separator" data set from assignment 2. Hand in the same output as in 1c above.

    How are the results different from the Haberman data set, and why is this? And is there overfitting?

[1] https://archive.ics.uci.edu/dataset/43/haberman+s+survival