Ariel University

Machine Learning

Homework 4

1. Implement k-nearest neighbor on the rectangle data set.
    a. Sample half the points; these are the training points. The remaining points are the test set.
    b. For each of k=1,3,5,7,9 and $p=1,2,\infty$, 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 test set.
    c. Repeat steps (a) and (b) 100 times, and print the average empirical and true errors for each p and k.

Which parameters of p,k are the best? Do you see overfitting? Hand in code, printout, and answers to these two questions.

2. A table of frequencies of Hebrew letters, based on a large number of texts, is found here:

    https://www.sttmedia.com/characterfrequency-hebrew

    A. Use Huffman encoding to encode the Hebrew alphabet in binary. (For the purposes of the encoding, you can take כ,ך and צ,ץ etc. to be the same letter.) Show your work.
    B. Use this encoding to encode your first name in binary.
    C. Now take the frequencies of letters based on your name, and use them to compute Shannon's entropy bound for encoding your name in binary. How does this compare with #2?