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

2. Now run the same algorithm on the "square" data set from assignment 3. Hand in the same output as in 1c above.

   How are the results different from the Haberman survival data set? And is there overfitting?

3. Suppose embedding $f: R^d \to R^k$ satisfies the bounds of the JL-Lemma: For any two points $u, v \in S$ it is true that $||v-u||_2 \leq ||f(v)-f(u)||_2 \leq (1+\epsilon)||v-u||_2$.

   Does f preserve the area of triangles? That is, is it true that for every triple $u, v, w \in S$, and some constant c,
   area($< u, v, w >$) $\leq$ area($< f(u), f(v), f(w) >$) $\leq (1+c\epsilon)$area($< u, v, w >$)?

   If so, give a proof and derive a value for c. Otherwise give a counterexample.