Nile University
Communication and Information Technology Program
Spring 2017

CIT651 –Introduction to Machine Learning and Statistical Data Analysis

Machine Learning - Assignment #2

(Due on: May 13, 2017 at mid-night)

a) Implement a *K*-Nearest Neighbor (KNN) classifier that uses the 80%-20% cross validation approach for determining the best *K* value. Apply the classifier to the training data of the 26 lower-case characters provided in the file "Problem 2 Dataset.zip". The zip file contains two folders: "Noise Train" and "Noise Test". The "Noise Train" folder contains 7 images for each lower-case character while the "Noise Test" folder contains 2 images for each lower-case character. The images in the "Noise Train" folder should be used in the cross validation. In your analysis, you should examine 10 different 80%-20% datasets that are randomly determined. Use maximum K of 100.

Deliverables:

- Your code.
- A plot of the classification error obtained for the training data during the validation process versus the choice of *K*. Name your file "KNN.jpg".

b) Use the test data to test your classifier. Apply your KNN classifier with the best value of *K* as obtained from part (a).

Deliverables:

- Your code.
- A plot of the number of images classified correctly for each character. The x-axis should show the character while the y-axis should show the count. Name the plot "Accuracy.jpg".

Important Notes:

- Do not use R functions for KNN classifier. You have to implement your own version of all needed functions. However, you are allowed to use the function that computes the norm of a vector or its equivalent.
- This is an individual assignment. It is not a team assignment.
- **To speed up the process of your function, in part (a), you should first compute the distance between each image in the 20% with all other images in the 80% and store such values in some data structure. You can then start changing K and get the nearest neighbors of each image from the values you stored instead of re-computing the distances with every change of K.**