G53MLE Machine Learning, 2011/12 Session Coursework

25% of total course mark

Deadline 4:00 PM 4th May 2012

The coursework will use a subset of MNIST handwritten digit dataset that you have prepared in the lab sessions and you will perform the following tasks:

1. Design and implement a single layer perceptron to distinguish the digit 5 from the other 9 digits, i.e., when the input is 5, the output will be +1, and -1 otherwise. Do your experiment follow these guidelines:

    *   Split the data into a training set and a testing set.
    *   Plot at least 4 graphs (e.g., 2 different sets of initial weight, 2 different training rates) of error over time (%-incorrect classification on y axis, training epochs on x axis) for the training set.
    *   Plot the corresponding graphs for the testing set.
    *   Explain the stopping criteria you used to terminate training and discuss the problem of over fitting.

2. Repeat task 1 for ADLINE but do your experiment follow these guidelines:

    *   Split the data into a training set and a testing set.
    *   Plot at least 4 graphs (e.g., 2 different sets of initial weight, 2 different training rates) of mean squared error (MSE) over time (MSE on y axis, training epochs on x axis) for the training set.
    *   Explain the stopping criteria you used to terminate training and discuss the problem of over fitting.
    *   To use the trained ADLINE to distinguish 5 from the other 9 digits, you can set a threshold T such that if the ADLINE's output is greater than T the input is classified as 5 and others if the output is smaller than or equal to T. Implement a procedure to automatically determine the best value of T. Explain and justify your implementation.

3. Compare and contrast the classification results of task 1 and task 2.

4. Design and implement a K Nearest Neighbour (K-NN) classifier to classify the 10 digits.

    You should follow the following guideline

    *   Randomly divide the data into training set and testing set, say 70/30 split (you can try other combinations, e.g., 50/50. 40/60, 30/70 etc). Try out at least 5 different random splits and in each case

        *   Automatically pick the best K.
        *   Report the classification result.

Submission

What to submit (deadline 4:00 PM 4th May 2012)

o Submit an electronic copy (code and report) through the School's CW system and a hardcopy of report to the School office. Your report should be no more than 4 pages long (font size 11 or above including all figures and plots) which should include

    *   A brief explanation of your implementations
    *   Results and explanation of results
    *   Conclusions and reflections (or what you've learned about Machine Learning through doing this coursework, ½ page maximum)