Problem Set 6

MAS 622J/1.126J: Pattern Recognition and Analysis

Due Monday, 22 November 2010. Resubmission due, 24 November 2010

Note: All instructions to plot data or write a program should be carried out using MATLAB. In order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.

If you collaborated with other members of the class, please write their names at the end of the assignment. Moreover, you will need to write and sign the following statement: "In preparing my solutions, I did not look at any old homeworks, copy anybody's answers or let them copy mine."

Problem 1: Neural Nets [40 points]

In this problem, we are going to build a classifier to recognize handwritten digits. The datasets can be downloaded from the course webpage, they are from the UCI Machine Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

The 32 x 32 bitmaps of handwritten digits are preprocessed and are divided into non-overlapping blocks of 4 x 4 and the number of on-pixels are counted in each block. This generates an input matrix of 8 x 8 where each element is an integer in the range [0..16]. This reduces the dimensionality and gives invariance to small distortions.

The dataset was further processed, to rescale the input elements to a number between 0-1, and to transform the output values into 10 binary outputs. You get two arrays, data (input to neural network) and labels (output of neural network). The labels matrices contain the classification for the examples, as binary flags with zeros everywhere except for a 1 in the correct position (i.e. 0 0 0 0 0 0 1 0 0 0 is a 6 - positions correspond to: [0 1 2 3 4 5 6 7 8 9]).

For this problem you are free to write your own code or use any MATLAB toolboxes available for the purpose. You can use the default Neural Network toolbox available. Try help nnet, help nntool to help you get started. If you type demo on the MATLAB prompt, under the option Toolboxes, you can select Neural Networks to view some examples.

a. Train a two-layer neural network with sigmoidal hidden units (i.e. 1-hidden layer).

Train the network using the back-propagation algorithm with the provided training set. Test your network and report recognition results.

b. Experiment with different numbers of hidden units to optimize recognition accuracy.

c. Find a fixed number of hidden units, n that works well for the dataset and report the recognition results.

d. Comment on the effects of varying the number of hidden units on recognition accuracy. Suggestion: Plot the average percentage of recognition accuracy as a function of the number of neurons in the hidden layer to support your argument.

e. Train a 2-hidden layer model using n/2 hidden units in each layer, where n corresponds to the value you found in part (c). Test your network and report recognition results. Compare and comment on the results of the 1-hidden layer model (part c) and the 2-hidden layer model (part e)

Problem 2: Decision Trees [40 points]

We collected the following information from people who were accepted into the doctoral program at MedianLab and people who did not.

Hardworking	Cool	GPA	Accepted
N	N	4	N
Y	N	3.6	Y
N	Y	3.4	Y
Y	N	3	N
Y	Y	3.1	Y
N	Y	2.9	Y

a. Assuming binary splits (e.g. GPA $< t$ vs. GPA $\geq t$), what values of t do you need to consider to find the optimal split of the feature GPA?

b. Write a program for training an ID3 decision tree. Draw the decision tree and indicate the information gain of each split. (Consider binary splits as in the previous question.)

Problem 3: Feature Selection [20 points]

Given the following observations of the class label Y:

X_1	X_2	X_3	X_4	Y
1	0	0	1	1
1	1	0	1	1
0	0	1	0	1
0	0	0	1	0
0	0	0	0	0
1	1	1	0	0

a. What is the mutual information of each feature with the class label?

b. What features would you choose to reduce the dimensionality? Is mutual information the best criteria for feature selection?