UNIVERSITY OF CALIFORNIA, DAVIS
DEPARTMENT OF COMPUTER SCIENCE

ECS 171: Homework Set 2

Instructor: Ilias Tagkopoulos
TAs: Ameen Eetemadi and Minseung Kim
{eetemadi, msgkim}@ucdavis.edu

Homework is due on October 29, 2015

General Instructions: The homework should be submitted electronically through Smartsite. Each submission should be a zip file that includes the following: (a) a report in pdf format ("report_HW2.pdf") that includes your answers to all questions, plots, figures and any instructions to run your code, (b) the matlab/octave code files. Please note: (a) do not include any other files, for instance files that we have provided such as datasets, (b) each function should be written in a separate file, with the appropriate remarks in the code so it is generally understandable (what it does, how it does it), (c) do not use any toolbox unless is it explicitly allowed in the homework description. Shared/copied code from any source is not allowed, as it is considered plagiarism. There is a 20% penalty per day for late homeworks.

1 WHERE DID THE BAKER GO? [100PT]

In this exercise, you will build a classifier that can find the localization site of a protein in yeast, based on 8 attributes (features). You will use the "Yeast" dataset ("yeast.data" file; 1484 proteins, 8 features, 10 different classes; no missing data) that is available in the UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/machine-learning-databases/yeast/

Perform and report (code and results) the following:

1. Construct a 3-layer artificial neural network (ANN) and specifically a feed-forward multi-layer perceptron to perform multi-class classification. The hidden layer should have 3 nodes. Split your data into a random set of 65% of the samples as the training set and the rest 35% as the testing set. Please note that you will never train with the testing set; the ANN will only take into account the training set for updating the weights. Plot how the weights, error and output changes at each iteration for both the training set and the testing set. Use stochastic gradient descent with back-propagation. You can use any published code that is freely available online (GNU license or similar), with no restrictions on functionality or language used.[30pt]

2. Now re-train the ANN with all your data (all 1484 samples). What is your training error? Provide the final activation function equations (i.e. a_i) after training. [10pt]

3. For the ANN that you have built (3 layers, 1 hidden layer, 3 hidden nodes) calculate the first round of weight update with back-propagation with paper and pencil for all weights but for only the first sample. Confirm that the numbers that you calculated are the same with those produced by the code and provide both your calculations and the code output. Provide both calculations made by hand (scanned image is fine) and corresponding output from the program that shows that both are in agreement. [30pt]

4. Increase the number of hidden layers from 1 to 2 and then to 3. Then increase the number of hidden nodes per layer from 3 to 6, then to 9 and finally to 12. Create a 3x4 matrix with the number of hidden layers as rows and the number of hidden nodes per layer as columns, with each element (cell) of the matrix representing the testing set error for that specific combination of layers/nodes. What is the optimal configuration? What you find the relationship between these attributes (number of layers, number of nodes) and the generalization error (i.e. error in testing data) to be? [25pt]

5. Which class does the following sample belong to?[5pt]

   Unknown Sample 0.50 0.49 0.52 0.20 0.55 0.03 0.50 0.39

6. Can you come up with a quantitative measure of uncertainty for each classification? What is the uncertainty for the unknown sample of the previous question? Justify your assumptions and method [5pt bonus]

GOOD LUCK!