## CMPE 452 Assignment 2: Implement a Backpropagation Network

## Andrew Litt - 10150478, 14asl

Design Choice Explanations

**Section 1: Design Parameters, 2 Marks**

*Initial Weights:*

For this system, the weights were initialized to random values between -0.2 to 0.2. Given a feed of 11 input features to the hidden layer then 7 hidden activations to each output, having many large values being fed to the activation function could cause the sigmoid activation to appear almost binary. Smaller initial weight vectors can help improve the range of values the activation function will output, and avoid ‘saturation’, so to speak.

*Node Output Function:*

Each node in the output and hidden layers uses a sigmoid activation function. This choice was used as sigmoid functions are continuously differentiable – a requirement for the use of gradient descent – and is the most commonly used activation function for backpropagation networks.

*Learning Rate:*

For this application, the learning rate was varied within the common range of 0.1 to 0.9. After continuous testing of the system, the learning rate was most effective around 0.9.

*Terminating Criteria:*

The terminating criteria for training was set to be when the system reached a minimum mean-squared error (MSE) or hit a maximum number of iterations through the training patterns. These thresholds were set to be a minimum verification set MSE of 0.03, or a maximum 2000 iterations.

**Section 2: Data Preprocessing, 2 Marks**

Before being entered into the backpropagation algorithm, each data category was normalized between 0 and 1 in order for each category to start with the same level of impact to the hidden layer activation function. In addition, an equal amount of each class was fed into a sub-dataset to partition the classes into statistically equal sets. Overall, the preprocessing went as follows:

1. Import data
2. Normalize each pattern between 0 and 1
3. Sort into the 3 separate classifications
4. Randomly shuffle the order of the separate classes
5. Create a dataset using equal parts of each class
6. Randomly shuffle concatenated dataset
7. Map outputs 5,7,8 to [1,0,0],[0,1,0],[0,0,1], respectively
8. Split into training, validation, and testing classes

While this approach may reduce the total amount of data points being used in each class, it should make each class and training set statistically indistinguishable from one another.

**Section 3: Data Split, 2 Marks**

With partitioning the datasets to have an equal amount of each class, this had a great impact on the amount of data points available. To have equal amounts of each class, this limited the dataset to be 525 points to split amongst. As given by the Baum and Haussler condition, a desired target patterns for a 11-7-3 network would be 525 for 80% accuracy. Unfortunately, even with a 80/10/10 split for training, verification and testing respectively, all that can be given is 420 training points before significantly shrinking the size of the verification sets.

In order to balance each class without leaving too few points for training and verification, the split was left as 80/10/10.

While a broadening of the data used to not include each class proportionally could have been explored, it is unknown which balance of extra data would aid with training without disproportionally training with one class over another.

**Section 4: Number of Layers & Nodes, ½ Mark**

For this dataset, a BP network with 1 hidden layer was used. Given that many of the features relate to measured acidity or sulfurs, I concluded that there would only need to be one hidden layer necessary to group features together. While there are many ‘rules of thumb’ for choice of hidden nodes, a common one is to choose a value roughly halfway between the number of features and output nodes. With this, my choice was 6 nodes in the hidden layer.

**Section 5: Momentum Value, ½ Mark**

The choice of momentum value should help fulfill its purpose of avoiding local minimums while avoiding constant overshoot, similar to the learning rate. As seen in lecture 9 slide 38, a momentum value of was chosen for Example 3.71. This was an adequate value, also chosen for this application.

**Section 6: Precision & Recall, Confusion Matrix, Final Weight Vectors, Results – 2 Marks Total**

See results.xlsx, with the results spread across various tabs.