EE 634/734 Intro to Neural Networks

Assignment #1

Due Date: Wednesday, September 13, 2023

In this project, we want to implement a neural network using the breast cancer dataset, published by the University of Wisconsin Hospitals. The goal is to build a binary classification model that can predict whether a tumor is malignant (1) or benign (0) based on various features.

The data consists of 699 rows and 11 columns, with each row representing a different patient. The fields in each row are as follows:

Column	Attribute	Range of Values
0	Patient ID number	7-digit number
1	Clump Thickness	1 - 10
2	Uniformity of Cell Size	1 - 10
3	Uniformity of Cell Shape	1 - 10
4	Marginal Adhesion	1 - 10
5	Single Epithelial Cell Size	1 - 10
6	Bare Nuclei	1 - 10
7	Bland Chromatin	1 - 10
8	Normal Nucleoli	1 - 10
9	Mitoses	1 - 10
10	Class	2 (benign) or 4 (malignant)

The following are the first 6 rows of the data file:
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2
1016277,6,8,8,1,3,4,3,7,1,2
1017023,4,1,1,3,2,1,3,1,1,2
1017122,8,10,10,8,7,10,9,7,1,4

Step 1: Load and Preprocess the Dataset:

- Input Data: Read the data from the file into a two-dimensional array.
- Preprocess the data: convert the “Class” column from 2 and 4 into 1 and 2.

Step 2: Shuffle and split the data set:

- Shuffle the rows of the array using the Numpy function: *np.random.shuffle(arr)*
- Split the data into training and testing subsets. For example, you can take the first 80% of the data rows as the training set and the remaining 20% as your test set.

Step 3: Train the model with the training data:

- Train the model using a one-hidden layer implementation of a neural network, with 8 neurons in the hidden layer.
- How many epochs are required to reduce the error to a minimum.
- Validate your trained model using the test data.
- What was the accuracy achieved on the test data by the trained model. Show the confusion matrix.

Step 4: Experimenting with different number of neurons in the hidden layer:

- What accuracy would be achieved on the test data if the model was trained with 5 neurons in the hidden layer.
- What accuracy would be achieved on the test data if the model was trained with 10 neurons in the hidden layer.

Step 5: Experimenting with different data split:

- Compare the accuracy achieved on the test data if the data was originally split into 60%-40% .
- Compare the accuracy achieved on the test data if the data was originally split into 90%-10% .

Step 6: Using a different cost function:

- In the given code, the algorithm uses cross-entropy to evaluate the loss. Does the accuracy change if we use the Root-Mean-Square-Error (RMSE) instead. Show by example.

Step 7: (Bonus question) Using different batch sizes for training the model:

- Modify the code so that the algorithm uses a batch size of 32 during training.
- Is the accuracy affected by using a batch size of 64.