Programming Methodology(UoG-PM) 2020-21

Assessed Exercise: Neural Network Machine Learning in C programming

Introduction

The goal of this AE exercise is to familiarize yourselves with the design, implementation and performance testing of C programming with the given medical dataset in the Internet of Things (IoT) edge computing machine learning landscape. Given the problem statement with the dataset, you will be required to go through the whole cycle of problem definition, problem analysis, algorithm and pseudocode design. You will implement with testing and demonstrate your classification/regression algorithm in C programming.

Dataset and Software

You will be working on a dataset **Group1_4.txt** for **Group1** to **Group 4 , Group5_8.txt** for **Group5** to **Group 8** and **Group 9_13.txt** for **Group 9** to **Group 13** that comprises of

- 100 rows with 10 columns of information[1] . Based on WHO 2010 criteria, they represent 100 volunteers' semen analysis with 9 attributes to determine the fertility of the volunteers as sperm concentration are related to the environmental factors, socio-demographic data, health status and life habits. These 9 attributes and semen fertility outcome are as follows:

Attributes	Data Format	Data Representation
Season of analysis	Real number	Winter : -1 Spring : -0.33 Summer : 0.33 Fall : 1
Age of Analysis	Real number	Age 18 to 36 : 0 to 1
Childish Disease (Chicken pox, measles, mumps, polio	Integer	Yes : 0 No : 1
Accident or serious trauma	Integer	Yes : 0 No : 1
Surgical Intervention	Integer	Yes : 0 No : 1
High fevers in last year	Integer	Less than three months ago : -1 More than three months ago : 0 No : 1
Frequency of alcohol consumption	Real number	Several times a day : 0.2 Every day : 0.4

[1] 2013 Data from Lucentia Research Group and Department of Biotechnology, University of Alicante, Spain

		Several times a week :0.6 Once a week :0.8 Hardly ever or never : 1
Smoking Habit	Integer	Never : -1 Occasional : 0 Daily : 1
Number of hours spent sitting per day	Real number	Ene-16 (0 range to 1)
Semen Diagnosis	Integer	Normal : 0 Altered : 1

- The dataset will be uploaded and be available in the SIT-Xsite Dropbox Folder under the course assessment folder and you should work on the dataset that belongs to your grouping according to the grouping list in the **Assessed Exercise Grouping List.pdf.** The programming language is to be used is the same standard C programming language where Microsoft Visual Studio Code (VSC) is the baseline editor and GNU Compiler Collection (GCC) is the standard compiler. Your C Program is free to call/activate any plotting tool such as GNU plot at http://www.gnuplot.info Pls refer to the website on how to install the gnuplot for Windows and Mac (For Mac, it is better to use homebrew for ease of installation).

Assessed task

Assuming you are the data scientist and you are tasked by a particular local hospital doctor to implement a classification /regression algorithm for the above dataset to help the doctor to improve the fertility diagnostic accuracy for her patient. At this point of time, there is only 100 data records as mentioned in previous section, but her database will grow as her patient record on fertility cases increases. As such, the accuracy of her diagnostic must get better as data record increases. You are tasked to implement the diagnostic algorithm on her laptop (edge computing) so that she can eventually arrive immediately at her patient fertility diagnostic conclusion whenever she keys in on the spot the answer for the 9 attributes with the patient and match with the patient's semen analysis outcome.

Your main task is multiple fold

1) Implement the edge computing machine learning using neuron network (NN) with **one perceptron** in C programming to train and test a perceptron with the given dataset, learning rate η, sigmoid activation function $\sigma(x)$ and targeted Mean Absolute Error (MAE). For a start, you can set $\eta = 0.05$ and targeted $MAE = 0.25$
2) Calculate the MAE on training data set for every iteration and plot the MAE graphically till it reach MAE of 0.25. Hence, determine the iteration number to reach such accuracy. Fine tune the η and targeted MAE to assess how low your MAE can go.

3) Determine the Minimum Mean Square Error (MMSE) and obtain the confusion matrix based for both training and testing data set.
4) Calculate the time required for your entire whole program from extracting the given dataset, perform perceptron training till plotting the error estimation rate.

Looking at the dataset and problem statement above, you should first perform problem definition, problem analysis followed by the algorithm/pseudocode before proceed to the C programming coding. Your C program should be of originality with modularity and functionality with the NN concept described pictorially in next section. Below serves as a guideline but not limited to.

1) Input
   a) Write a function to open the dataset file (Hint: File pointer) and extract the respective attribute data of each volunteer data $(x_{1,1}, x_{1,2}, x_{1,3},, x_{1,4}, x_{1,5}, x_{1,6}, x_{1,7}, x_{1,8}, x_{1,9}, x_{1,10}) \cdots$ $(x_{100,1}, x_{100,2}, x_{100,3},, x_{100,4}, x_{100,5}, x_{100,6}, x_{100,7}, x_{100,8}, x_{100,9}, x_{100,10})$ into multidimensional data array set.
   b) Split into training set and testing set such as first 90 volunteer data is training set while the last 10 volunteer data as testing set
   c) For each set, further split into feature and output set for NN training and testing
2) Write various modular functions to perform perceptron training such as
   a) Calculate linear regression $\boxed{z_i(t) = \sum_{j=1}^{n} w_j^t\, x_{ij} + b_i^t = \vec{\mathbf{w}}^t . \vec{\boldsymbol{x}}_i + b_i^t}$ based on the input $\vec{\boldsymbol{x}}_i$, new weight $\vec{\mathbf{w}}^t$ and new bias b^i for each iteration t. $\vec{\mathbf{w}}^t = [w_1^t, w_2^t \cdots w_n^t], \vec{\boldsymbol{x}}_i = [x_{i1}, x_{i2} \cdots x_{in}]^T$ while n is the number of input features and i is the i^{th} data samples where $i = 1 \cdots I$. For example, in training set, $I = 90$ based on 90:10 data split.
   b) Calculate the sigmoid activation function output, $\boxed{\sigma(z_i(t)) = \frac{1}{1+e^{-z_i(t)}}}$ for each iteration t.
   c) Calculate the MAE for each iteration where $\boxed{MAE^t = \frac{\sum_{i=1}^{I} |\hat{y}_i{}^t - d_i|}{I}}$ where $\hat{y}_i{}^t = \sigma(z_i(t))$ and d_i is the true label or output of the i^{th} data sample.
   d) Backward propagation weight and bias update for each next iteration $t + 1$ if MAE does not meet expectation tolerance of 0.25 with $\boxed{\vec{\mathbf{w}}^{t+1} = \vec{\mathbf{w}}^t \text{ - } \eta \nabla_w E^t}$, $\boxed{b_i^{t+1} = b_i^t - \eta \nabla_b E^t}$ where

$$\nabla_w E^t = \left[\frac{\partial E^t}{\partial w_1^t}, \frac{\partial E^t}{\partial w_2^t}, \cdots \frac{\partial E^t}{\partial w_n^t}\right], \nabla_b E^t = \frac{\partial E^t}{\partial b_i^t} \quad , E^t = \frac{1}{2I} \sum_{i=1}^{I} \left(\hat{y}_i{}^t - d_i\right)^2$$

   and

$$\frac{\partial E^t}{\partial w_j^t} = \frac{\partial E^t}{\partial \hat{y}^t} . \frac{\partial \hat{y}^t}{\partial z_i(t)} . \frac{\partial z_i(t)}{\partial w_j^t} = \boxed{\frac{1}{I} \sum_{i=1}^{I} \left(\hat{y}_i{}^t - d_i\right) . \frac{e^{z_i(t)}}{\left(1+e^{z_i(t)}\right)^2} x_{ij}} \text{ where } j = 1 \cdots n$$

$$\frac{\partial E^t}{\partial b_i^t} = \frac{\partial E^t}{\partial \hat{y}^t} . \frac{\partial \hat{y}^t}{\partial z_i(t)} . \frac{\partial z_i(t)}{\partial b_i^t} = \boxed{\frac{1}{I} \sum_{i=1}^{I} \left(\hat{y}_i{}^t - d_i\right) . \frac{e^{z_i(t)}}{(1 + e^{z_i(t)})^2}}$$

   e) Calculate MMSE for both training & testing dataset and confusion matrix parameters where

$$\boxed{MMSE = \frac{1}{I}\sum_{i=1}^{I}\left(\hat{y}_i^{\,t} - d_i\right)^2}$$

For example, $I = 90$ and $I = 10$ for training and testing set respectively based on 90:10 data split.

For Confusion matrix, determine the four classes namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)

True Positive : No of times predict correctly that patient is not normal

True Negative: No of times predict correctly that patient is normal

False Positive: No of times predict wrongly that patient is not normal but patient is actually normal

False Negative: No of times predict wrongly that patient is normal but patient is actually not normal

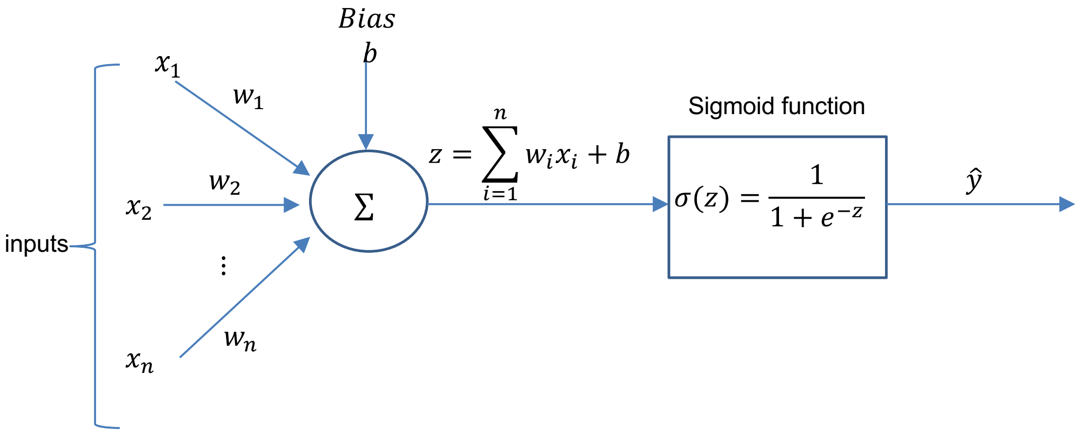
https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

3) The output program shall also print the following result
   a) Plot the MAE^t for every iteration t till the error accuracy of 0.25 is reached
   b) MMSE for training set and testing set before trained and after trained.
   c) Confusion matrix values
   d) Time taken to complete entire program

As such as possible, make full use of pointer, dynamic memory allocation, array concept in your program and functions if possible.

Basic Neural network Principle

Artificial Intelligence (AI) such as Deep learning (DL) tries to mimic the neurons in human brain to form large number of interconnected neurons that learn, train and produce a decision in many applications such as image recognition or robotics. In this assignment, one neuron is named as perceptron is used to predict the fertility output of the patient based on past data samples. A perceptron can be mathematically modelled as follows:



$Bias$
b
x_1
w_1
x_2
w_2
inputs
⋮
x_n
w_n
Σ
$z = \sum_{i=1}^{n} w_i x_i + b$
Sigmoid function
$\sigma(z) = \frac{1}{1+e^{-z}}$
$\hat{y}$