CSCI4390-6390 Assign5

Assign5: Neural Nets (MLP)

Due Date: Nov 1, before midnight (11:59:59PM, Alofi Time; GMT-11)

You will use the Appliances energy prediction data set. You should ignore the first attribute, which is a date-time variable, and you should also remove the last attribute, which is a duplicate of the previous one. Use the first attribute (after removing the date-time variable), which denotes the **Appliances Energy Use**, as the response variable, with the remaining attributes as predictor variables. However, you have to discretize the response variable differently for the two sections as noted below, since CSCI4390 will implement binary classification, whereas CSCI6390 will implement multiclass classification via MLP.

CSCI4390: Binary Classification

You will implement binary classification via the MLP training Algorithm 25.1 (Chapter 25, page 655). This algorithm assumes sigmoid activation for hidden layer, and squared error for the loss. However you have to use ReLU activation for the hidden layer, and use binary cross entropy for the loss. Consequently, you have to modify lines 10 and 11 appropriately so that line 10 uses cross-entropy loss, and line 11 uses ReLU derivative. See section 25.1.1 for derivative of ReLU, and section 25.2.3 for the derivative of binary cross-entropy loss function.

Note that the **Appliances Energy Use** attribute takes values in the range $[10, 1080]$. However, for binary classification, we need only two values, so for the purpose of this assignment you should consider energy use less than or equal to 50 as the positive class (1), and energy use higher than 50 as negative class (0).

You should shuffle the data points before selecting 70% of the data training and 30% for testing, so that there is an equal mix of the classes in both.

CSCI6390: Multiclass Classification

You will implement multiclass classification via the deep MLP algorithm in Algorithm 25.2 (Chapter 25, page 668). You should use ReLU activation for all hidden layers, and softmax for the output layer. See section 25.1.1 for the derivative of the activation functions, and end of section 25.4.3 for derivative of the multiclass cross-entropy function.

Note that the **Appliances Energy Use** attribute takes values in the range $[10, 1080]$. However, for multiclass regression, we will convert these into four classes as follows: energy use less than or equal to 30 is class c_1, energy use greater than 30 but less than or equal to 50 is class c_2, energy use greater than 50 but less than or equal to 100 is class c_3, and finally energy use higher than 100 is class c_4. You need to do this conversion to create the categorical response variable, before you select the train (70%) and test (30%) subsets.

You should shuffle the data points before selecting 70% of the data training and 30% for testing, so that there is an equal mix of the classes in both.

What to submit

- Write a scripy named as **Assign5.py**, which will be run as

Assign5.py FILENAME ETA MAXITER HIDDENSIZE for CSCI4390

Assign5.py FILENAME ETA MAXITER HIDDENSIZE NUMHIDDEN for CSCI6390

FILENAME is the datafile name, ETA is the step size η, MAXITER is the number of epochs to train the model, HIDDENSIZE is the size of the hidden layer, and NUMHIDDEN is the number of hidden layers (this is for CSCI6390). CSCI6390 students may assume that all hidden layers use the same size.

You should implement the algorithms using NumPy; you cannot use any of the deep learning libraries (like keras, pytorch or tensorflow, and so on).

Your script should print out the weight matrices and biases for each layer, and also the final accuracy value on the test data, where you use 70% of the data for training and 30% for testing.

Note that accuracy is defined as the fraction of correct class label predictions. So for each test point, you should predict its class as the one that has the highest probability, and then you should count how many test points are correctly predicted. Divide that number by the test data size to get the accuracy. Report the results for the combination of ETA and HIDDENSIZE (and NUMHIDDEN for CSCI6390) that yields the best accuracy.

- Submit a PDF file named Assign5.pdf that should include your answers to each of the questions (just cut and paste the output from python). **Failure the submit the PDF will result in lost points.**
- Submit the scripts and pdf file via submitty

Policy on Academic Honesty

You are free to discuss how to tackle the assignment, but all coding must be your own. Please do not copy or modify code from anyone else, including code on the web. Any students caught violating the academic honesty principle will get an automatic F grade on the course and will be referred to the dean of students for disciplinary action.