Quantized Machine Learning

Objective:

This lab aims to provide students with hands-on experience in applying quantization methods (Dynamic Quantization) to logistic regression in PyTorch.

Part 1: Setup and Data Preparation

1. **Environment Setup**:
   - Import necessary libraries such as sklearn, numpy and quantization tools from PyTorch.
2. **Data Loading**:
   - Use the torchvision datasets to load the MNIST dataset. Apply transformations to normalize the data.
   - https://scikit-learn.org/1.5/modules/generated/sklearn.datasets.load_digits.html
   - Make
     i. X = digits.data
     ii. y = digits.target
   - Split the dataset into training and test split.

Part 2: Model Building

1. **Use LogisticRegression to fit the model.**

Part 3: Report model accuracy, Model size, Inference time of Logistic regression model.

Part 4: Create a function name quantize_model, scale the weights of original model to 8-bit.

Part 5: Create another function to inference using the quantized model.

Keep scale_factor = 2 ** 7 (number of bits is 8)

Part 6: Report Quantized model accuracy, Quantized model size, Quantized inference time.

1. **Model Size Comparison**:
   - Compare and print the results.

References:
https://pytorch.org/blog/introduction-to-quantization-on-pytorch/

https://pytorch.org/tutorials/recipes/recipes/dynamic_quantization.html