Homework 5 STOR 665

Due April 10, 2022

Submission Instruction

- Upload a **zip** folder that contains **the tex, the pdf, the py, the readme** file to Sakai.
- Use a new tex file to generate the pdf report and limit the number of pages of the pdf report to **2**.
- Also, upload your code and specify how to run your code in a README file (e.g., md).
- You are allowed to work with other students but homework should be in your own words. Identical solutions will receive a **0** in grade and will be investigated.

Goal

In this homework you will practice putting together a simple image classification pipeline to classify **handwritten digit 0 and 1**, based on the logistic regression or the SVM classifier. Please keep the batch size untouched. The goals of this homework are as follows:

(a) understand the basic Image classification pipeline and the data-driven approach (train/predict stages)

(b) understand how to use pytorch to build binary classifiers.

(c) understand how to exploit pytorch's autograd mechanics to do optimization.

(d) implement and apply a logistic regression classifier.

(e) implement and apply a binary Support Vector Machine (SVM) classifier.

Data

We use MNIST digit classification dataset. Pytorch/torchvision has provide a useful dataloader to automatically download and load the data into batches. In this homework, we need two class, digit 0 and digit 1, for binary classification. We have written the data loader for you as follow. You can find it in the attached python fille.

Problem Description

In this homework, you are asked to implement and solve linear Logistic Regression model and Linear SVM model (without regularization term) on MNIST dataset. In this task, you only need to perform binary classification on digit 0 and 1. Details of these models could be found in lecture 15 and lecture 17 slides. We provide a skeleton code for data loading and iterations of training data. You are asked to implement the rest of training in Pytorch code. You are required to optimize the model by using SGD and Momentum methods. Detailed requirements:

(a) For each of the model, report:

$$\frac{1}{B}\sum_{b=1}^{B}\sum_{d=1}^{D_b}\frac{loss(y_{bd}, f(\boldsymbol{x}_{bd})}{D_b}$$

for each training epoch.

- B: the total number of batches
- D_b: the number of observations in b-th batch
- f: the model (Logistic regression or Linear SVM)
- $(\boldsymbol{x}_{bd}, y_{bd})$: the d-th pair of input data and label in b-th batch
- An epoch is defined as one iteration of all observations in training dataset

In summary, you record the average training loss of each batch after you update the model, and report the average of batch losses after each epoch. You could plot the results as a figure or simply list down. Please at least report **10** epochs.

(b) Report the final testing accuracy of trained models (Logistic regression or Linear SVM).

(c) Please compare results for 2 optimizer (SGD and SGD-Momentum)).

(d) Try different step sizes and discuss your findings.

Resources

- You can follow the setup instructions at here.
- A useful tutorial on learning pytorch by examples at here.
- More illustrations of different optimizers could be found here.