CS7180: Algorithmic and Statistical Aspects of Deep Learning Fall 2020

Problem Set 1

Instructor: Hongyang R. Zhang *Due:* ***October 1, 2020, 11:59pm***

Policy: We encourage discussions and collaborations on the homework. You should write up the solution on your own, and remember to mention any fellow students you collaborated with. There are up to three late days for all the problem sets and project submissions. Use them wisely. After that, the grade depreciates by 20% for every extra day. Late submissions are allowed only under extreme situations. Please discuss with the instructor well in advance if you cannot meet the deadline. All homework submissions are subject to the Northeastern University Honor Code.

Submission: We will use Gradescope for the homework submissions. Login to gradescope through Canvas using your northeastern.edu account. We strongly recommend that you write up your solution in LaTeX, although you are allowed to send in scanned copies.

Length of submissions: Include as much of the calculations needed for us to understand the answer. After solving the problem, try to identify the main steps taken and critical points of proof and include them as a rule of thumb.

Problem 1

In this problem, we will set up a two-layer neural network and use the network to train a classifier for recognizing hand-written digits. We consider the MNIST dataset, which consists of 50,000 hand-written digits in the training dataset and 10,000 hand-written digits in the test dataset. Given the training dataset, we construct a two-layer neural network with rectified linear unit activations and use the cross-entropy loss (softmax plus negative log-likelihood) to evaluate the predicted result.

(a) **[10 points]** Let W_1 and W_2 denote the weight matrices of each layer, respectively. Given a sample with feature vector $x \in \mathbb{R}^{28\times 28}$ and label $y \in \{0, 1, 2, \ldots, 9\}$, what is the predicted output y^{pred} of the two-layer neural network? Write down the equation for y^{pred} as a function of W_1 and W_2, given input x.

(b) **[10 points]** Let $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ denote the training dataset, where $m = 50,000$, $x_i \in \mathbb{R}^{28\times 28}$, and $y_i \in \{0, 1, 2, \ldots, 9\}$ for every $i = 1, 2, \ldots, m$. What is the loss of the predicted label compared to the correct label? Write down the equation for the optimization objective over the training dataset as a function of W_1 and W_2 for the two-layer neural network.

(c) **[40 points]** Open the MNIST notebook that we handed out. Implement the above two-layer neural network using PyTorch. Follow the instructions in the notebook. Write down the discussions

mentioned in "Part 4: Ablation studies".

Problem 2

In this problem, we are going to do regression analysis using a two-layer neural network. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ denote the samples of the training dataset, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, for any $i = 1, 2, \ldots, m$.

(a) **[5 points]** Let W_1 and W_2 denote the weight matrices of the network, respectively. Suppose we use ReLU activation and quadratic loss function. Write down the equation for the training loss as a function of W_1 and W_2.

(b) **[15 points]** Describe the backpropagation algorithm. Write down the gradient formula of the training loss for W_1 and W_2. Show how to apply the backpropagation algorithm to compute the gradients.

(c) **[15 points]** Consider a teacher-student network setting. Our training dataset consists of data generated via a teacher network, and we use a student network to learn from the training dataset. Open the Synthetic notebook that we handed out. Implement the above teacher-student network setting. Follow the instructions in the notebook.

(d) **[5 points]** Vary the number of neurons in the hidden layer. Plot the test error as a function of the number of neurons.