Deep Learning - Final Exam (Take-Home)

Due 11:59 May 7, 2020

Problem 1. (8 points) Sketch a typical learning curve for the training and validation sets, for a setting where overfitting occurs at some point. Assume that the training set and the validation set are of the same size. Label all the axes, and label the curves that you sketch. State how to apply early stopping in the context of using a validation set?

Problem 2. (12 points) Consider solving a binary classification problem using a fully connected network $y = f(x, \theta)$ with two hidden layers of 100 hidden units each and with the ReLU activation. Write down the function $y = f(x, \theta)$ as a composition of functions defining hidden variables and clearly identify the trainable variables involved. For a training dataset $\{(x_i, y_i)\}_{i=1}^N$, write down a suitable loss function $\mathcal{L}(\theta)$. Find $\frac{\partial \mathcal{L}}{\partial b_1}$, where b_1 is the bias associated with the first hidden layer. You may express the derivative as a product and identify each term in the product.

Problem 3. (14 points) For $f(\mathbf{x}) = x_1^2 + x_2^2 - \cos(x_1 + x_2)$ where $\mathbf{x} = [x_1, x_2]$,

1. show that $[0, 0]$ is a local minimum of $f(\mathbf{x})$;
2. starting with $[1, 2]$, carry out two iterations of the gradient descent algorithm with a learning rate of 0.01;
3. what is approximately the rate of convergence of the gradient descent?

Problem 4. (8 points) When training using mini-batch gradient descent algorithm, it sometimes happens that the training loss goes up after performing an update iteration. What are two possible reasons that can cause this to happen? Explain your answers.

Problem 5. (10 points) Consider a CNN with two convolutional layers and a fully connected layer. The first layer is a valid convolution with a kernel of size $3 \times 3 \times 3 \times 16$ followed by a 2×2 max pooling. The second layer is a valid convolution with a kernel of size $3 \times 3 \times 16 \times 32$. Then we have a flatten layer followed by a fully connected layer to produce an output of dimension 10. If the input has dimension $64 \times 64 \times 3$, write down the function $y = f(x, \theta)$ as a composition of functions defining hidden variables and identify the trainable variables involved.

Problem 6. (8 points) PCA, Autoencoder, and GAN are three basic methods of unsupervised learning. Describe one similarity and one difference between the following two methods: (1) PCA and Autoencoder; (2) Autoencoder and GAN.