Labs
Machine Learning Course
Fall 2016

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
mlo.epfl.ch/page-136795.html
epfmlcourse@gmail.com

Problem Set 13, Dec 15th, 2016 (Neural Networks)

Goals. The goal of this exercise is to

- Better understand neural network
- Implement the feed-forward function and backpropagation in a simple neural net.

Setup, data and sample code. Obtain the folder `labs/ex13` of the course github repository

github.com/epfml/ML_course

In the following problems, we will use a very simple neural network. Let's assume we have a three-layer neural net with one input layer of size $D = 4$, $L = 1$ hidden layers of size $K = 5$, and one output layer of size 1, as shown in Figure 1.

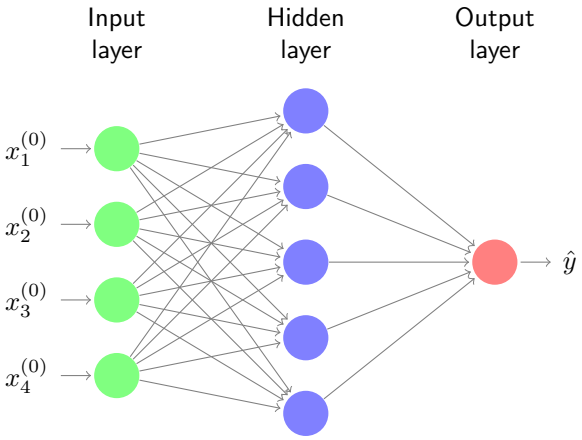

Figure 1: A simple neural network.

Problem 1 (Feed-forward in neural networks):

In our simplified neural network, we have the feed-forward function shown below:

$$x_j^{(1)} = \phi\left(z_j^{(1)}\right) = \phi\left(\sum_{i=1}^{D} w_{i,j}^{(1)} x_i^{(0)} + b_j^{(1)}\right), \tag{1}$$

$$\hat{y} = \phi\left(z_1^{(2)}\right) = \phi\left(\sum_{i=1}^{K} w_{i,1}^{(2)} x_i^{(1)} + b_1^{(2)}\right). \tag{2}$$

Use Equation 1 and Equation 2 to fill in the corresponding template function in the notebook, and pass the test. For simplicity, in the following questions, let the bias term be 0 and use the `Sigmoid` as the activation function $\phi(\cdot)$.

Problem 2 (Backpropagation in neural network):

Assume that we use the squared error as our loss function, as shown in Equation 3:

$$\mathcal{L} = \frac{1}{2}(\hat{y} - y)^2, \tag{3}$$

where we have only one sample in our case, and y is the true value while $\hat{y}$ is the network prediction.

Evaluate the derivative of $\mathcal{L}(\boldsymbol{w})$ with respect to weights $w_{i,1}^{(2)}$ and $w_{i,j}^{(1)}$, and implement the corresponding function in the notebook.

Solution:

First we note that $\mathcal{L}(\boldsymbol{w})$ depends on the weight $w_{i,1}^{(2)}$ only via the summed input $x_i^{(1)}$ to unit i. We can therefore apply the chain rule for partial derivatives to get

$$\frac{\partial \mathcal{L}}{\partial w_{i,1}^{(2)}} = \frac{\partial \mathcal{L}}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial w_{i,1}^{(2)}}. \tag{4}$$

We introduce the useful notation $\delta_i^{(2)} = \frac{\partial \mathcal{L}}{\partial z_i^{(2)}}$, where the δ's are often referred to as *errors*. We then have

$$\frac{\partial z_1^{(2)}}{\partial w_{i,1}^{(2)}} = \frac{\partial \Big(\sum_{i=1}^{K} w_{i,1}^{(2)} x_i^{(1)} \Big)}{\partial w_{i,1}^{(2)}} = x_i^{(1)}. \tag{5}$$

Thus, by substituting the equations, we obtain

$$\frac{\partial \mathcal{L}}{\partial w_{i,1}^{(2)}} = \delta_1^{(2)} x_i^{(1)}. \tag{6}$$

For the output units, we have

$$\delta_1^{(2)} = \frac{\partial \mathcal{L}}{\partial z_1^{(2)}} = \frac{\partial \big(\frac{1}{2}(\hat{y} - y)^2 \big)}{\partial y} \frac{\partial \phi(z_1^{(2)})}{\partial z_1^{(2)}} = (\hat{y} - y)\phi'(z_1^{(2)}) \tag{7}$$

and thus

$$\frac{\partial \mathcal{L}}{\partial w_{i,1}^{(2)}} = (\hat{y} - y)\phi'(z_1^{(2)}) x_i^{(1)}. \tag{8}$$

We re-use the chain rule to get the derivative of $\mathcal{L}(\boldsymbol{w})$ with respect to $w_{i,j}^{(1)}$:

$$\frac{\partial \mathcal{L}}{\partial w_{i,j}^{(1)}} = \frac{\partial \mathcal{L}}{\partial z_j^{(1)}} \frac{\partial z_j^{(1)}}{\partial w_{i,j}^{(1)}} = \delta_j^{(1)} \frac{\partial z_j^{(1)}}{\partial w_{i,j}^{(1)}} = \delta_j^{(1)} \frac{\partial \Big(\sum_{d=1}^{D} w_{d,j}^{(1)} x_d^{(0)} \Big)}{\partial w_{i,j}^{(1)}} = \delta_j^{(1)} x_i^{(0)}, \tag{9}$$

where

$$\delta_j^{(1)} = \frac{\partial \mathcal{L}}{\partial z_j^{(1)}} = \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{(2)}} \frac{\partial z_k^{(2)}}{\partial z_j^{(1)}} = \delta_1^{(2)} \frac{\partial \Big(\sum_i w_{i,1}^{(2)} x_i^{(1)} \Big)}{\partial z_j^{(1)}} = \delta_1^{(2)} \frac{\partial \Big(\sum_i w_{i,1}^{(2)} \phi(z_i^{(1)}) \Big)}{\partial z_j^{(1)}} = \delta_1^{(2)} w_{j,1}^{(2)} \phi'(z_j^{(1)}). \tag{10}$$

Hence,

$$\frac{\partial \mathcal{L}}{\partial w_{i,j}^{(1)}} = \delta_1^{(2)} w_{j,1}^{(2)} \phi'(z_j^{(1)}) x_i^{(0)}. \tag{11}$$

Problem 3 (Effect of regularization):

What is the effect of regularization on the weights? To get some insight, let $\boldsymbol{\Theta}$ be the vector of all weights in the neural network. Recall that we do not penalize the bias terms. Therefore, let us ignore them in the following.

Let $\boldsymbol{\Theta}^*$ be a parameter that minimizes the cost function $\mathcal{L}$ for the given test set (where the cost function does not include the regularization). We would like to study how the optimal weight changes if we include some regularization.

In order to make the problem tractable, assume that $\mathcal{L}(\boldsymbol{\Theta})$ can be locally expanded around the optimal parameter $\boldsymbol{\Theta}^*$ in the form

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{\Theta}^*) + \frac{1}{2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^\top \boldsymbol{H}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*),$$

where $\boldsymbol{H}$ is the Hessian whose components are the entries

$$\frac{\partial^2 \mathcal{L}}{\partial \Theta_i \partial \Theta_j}.$$

Now add a regularization term of the form $\frac{1}{2}\mu|\boldsymbol{\Theta}|_2^2$.

1. Show that the optimum weight vector for the regularized problem is given by

   $$\boldsymbol{Q}(\boldsymbol{\Lambda} + \mu \boldsymbol{I})^{-1}\boldsymbol{\Lambda}\boldsymbol{Q}^\top \boldsymbol{\Theta}^*,$$

   where $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ is the SVD of the symmetric matrix $\boldsymbol{H}$, $\boldsymbol{Q}$ is an orthonormal matrix, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are non-negative and decreasing along the diagonal.

2. Show that $(\boldsymbol{\Lambda} + \mu \boldsymbol{I})^{-1}\boldsymbol{\Lambda}$ is again a diagonal matrix whose i-th entry is now $\lambda_i/(\lambda_i + \mu)$.

3. Argue that along the dimensions of the eigenvectors of $\boldsymbol{H}$ that correspond to large eigenvalues λ_i essentially no changes occur in the weight, but that along the dimensions of eigenvectors of very small eigenvalues the weight is drastically decreased.

Solution:

1. We are minimizing $\mathcal{L} + \frac{1}{2}\mu|\boldsymbol{\Theta}|_2^2$. If we use the specific form of $\mathcal{L}$ and take the derivative, we get the first order condition of optimality

   $$\boldsymbol{H}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) + \mu\boldsymbol{\Theta} = \boldsymbol{0}.$$

   Solving for $\boldsymbol{\Theta}$ gives us $\boldsymbol{\Theta} = (\boldsymbol{H} + \mu\boldsymbol{I})^{-1}\boldsymbol{H}\boldsymbol{\Theta}^*$. If we now insert the SVD $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ this gives the stated solution.

2. Note that $\boldsymbol{\Lambda} + \mu\boldsymbol{I}$ is a diagonal matrix with entries along the diagonal equal to $\lambda_i + \mu$. The inverse of a diagonal matrix is again a diagonal matrix whose entries are the inverses of the entries of the matrix to be inverted. Hence for our case $(\boldsymbol{\Lambda} + \mu\boldsymbol{I})^{-1}$ is a diagonal matrix with entries $1/(\lambda_i + \mu)$. Finally, the multiplication of two diagonal matrices is a diagonal matrix whose entries are just the product of the diagonal entries.

3. Now note that if all λ_i were very large then all entries of the diagonal matrix $(\boldsymbol{\Lambda} + \mu\boldsymbol{I})^{-1}\boldsymbol{\Lambda}$ would be very close to 1 and hence $\boldsymbol{Q}(\boldsymbol{\Lambda} + \mu\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ would be essentially an identity matrix. (We first rotate the vector $\boldsymbol{\Theta}$ according to the unitary matrix $\boldsymbol{Q}^\top$ then pass through essentially an identity then rotate back.) But if some eigenvalues λ_i are close to 0 then $\lambda_i/(\lambda_i + \mu)$ is close to zero. And hence after rotating $\boldsymbol{\Theta}^*$ by multiplying with the unitary matrix $\boldsymbol{Q}^*$, some of the components are essentially set to zero, before we rotate the vector back.