NYU Center for Data Science

DS-GA 1003: Machine Learning (Spring 2020)

Homework 7: Computation Graphs, Backpropagation, and Neural Networks

Due: Monday, May 11, 2020 at 11:59pm

Instructions. You should upload your code and plots to Gradescope. Please map the Gradescope entry on the rubric to your name and NetId. You must follow the policies for submission detailed in Homework 0.

---

1 Introduction

There is no doubt that neural networks are a very important class of machine learning models. Given the sheer number of people who are achieving impressive results with neural networks, one might think that it's relatively easy to get them working. This is a partly an illusion. One reason so many people have success is that, thanks to GitHub, they can copy the exact settings that others have used to achieve success. In fact, in most cases they can start with "pre-trained" models that already work for a similar problem, and "fine-tune" them for their own purposes. It's far easier to tweak and improve a working system than to get one working from scratch. If you create a new model, you're kind of on your own to figure out how to get it working: there's not much theory to guide you and the rules of thumb do not always work. Understanding even the most basic questions, such as the preferred variant of SGD to use for optimization, is still a very active area of research.

One thing is clear, however: If you do need to start from scratch, or debug a neural network model that doesn't seem to be learning, it can be immensely helpful to understand the low-level details of how your neural network works – specifically, back-propagation. With this assignment, you'll have the opportunity to linger on these low-level implementation details. Every major neural network type (RNNs, CNNs, Resnets, etc.) can be implemented using the basic framework we'll develop in this assignment.

To help things along, we[1] have designed a minimalist framework for computation graphs and put together some support code. The intent is for you to read, or at least skim, every line of code provided, so that you'll know you understand all the crucial components and could, in theory, create your own from scratch. In fact, creating your own computation graph framework from scratch is highly encouraged – you'll learn a lot.

2 Computation Graph Framework

To get started, please read the tutorial on the computation graph framework we'll be working with. (Note that it renders better if you view it locally.) The use of computation graphs is not specific to machine learning or neural networks. Computation graphs are just a way to represent a function

[1] Philipp Meerkamp, Pierre Garapon, and David Rosenberg

that facilitates efficient computation of the function's values and its gradients with respect to inputs. The tutorial takes this perspective, and there is very little in it about machine learning, per se.

To see how the framework can be used for machine learning tasks, we've provided a full implementation of linear regression. You should start by working your way through the `__init__` of the `LinearRegression` class in `linear_regression.py`. From there, you'll want to review the node class definitions in `nodes.py`, and finally the class `ComputationGraphFunction` in `graph.py`. `ComputationGraphFunction` is where we repackage a raw computation graph into something that's more friendly to work with for machine learning. The rest of `linear_regression.py` is fairly routine, but it illustrates how to interact with the `ComputationGraphFunction`.

As we've noted earlier in the course, getting gradient calculations correct can be difficult. To help things along, we've provided two functions that can be used to test the backward method of a node and the overall gradient calculation of a `ComputationGraphFunction`. The functions are in `test_utils.py`, and it's recommended that you review the tests provided for the linear regression implementation in `linear_regression.t.py`. (You can run these tests from the command line with `python3 linear_regression.t.py`.) The functions actually doing the testing, `test_node_backward` and `test_ComputationGraphFunction`, may seem a bit intricate, but they're implementging the exact same `gradient_checker` logic we saw in the first homework assignment.

Once you've understood how linear regression works in our framework, you're ready to start implementing your own algorithms...

3 Ridge Regression

When moving to a new system, it's always good to start with something familiar. But that's not the only reason we're doing ridge regression in this homework. As we discussed in class, in ridge regression the parameter vector is "shared", in the sense that it's used twice in the objective function. In the computation graph, this can be seen in the fact that the node for the parameter vector has two outgoing edges. While we don't have this sharing in the multilayer perceptron, we do have it in RNNs and CNNs, which are two of the most important neural network architectures in use today. In the context of RNNs and CNNs, this parameter sharing is also referred to as **parameter tying**. So being able to handle the shared parameters in ridge regression is an important prerequisite to handling more sophisticated models.

We've provided some skeleton code in `ridge_regression.py` and some test code in `ridge_regression.t.py`, which you should eventually be able to pass.

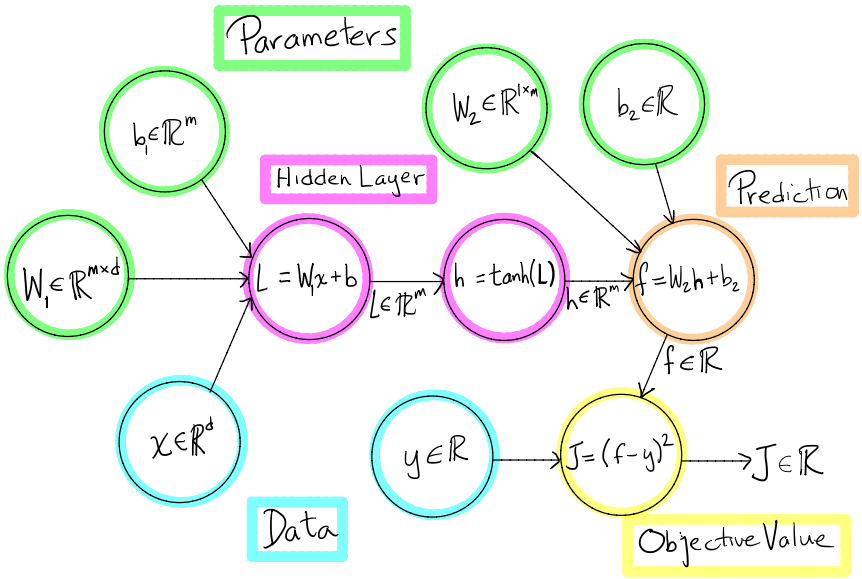
1. Complete the class `L2NormPenaltyNode` in `nodes.py`.
2. Complete the class `SumNode` in `nodes.py`.
3. Implement ridge regression with w regularized and b unregularized. Do this by completing the `__init__` method in `ridge_regression.py`, using the classes created above. When complete, you should be able to pass the tests in `ridge_regression.t.py`. Report the average square error on the **training** set for the parameter settings given in the `main()` function.

4. [Optional] Create a new implementation of ridge regression that supports efficient minibatching. You will replace the `ValueNode x`, which contains a vector, with a `ValueNode X`, which contains a matrix. The convention is that the first dimension indexes examples and the second indexes features, as we have done throughout the course. Many of the nodes will have to be adapted to this use case. Demonstrate the use of minibatching for your ridge regression network, and note the amount of speedup you get.

4 Multilayer Perceptron

In this problem, we'll be implement a multilayer perceptron (MLP) with a single hidden layer and a square loss. We'll implement the computation graph illustrated below:

Multilayer Perceptron, 1 hidden layer, square loss



The crucial new piece here is the nonlinear **hidden layer**, which is what makes the multilayer perceptron a significantly larger hypothesis space than linear prediction functions.

4.1 The standard non-linear layer

The multilayer perceptron consists of a sequence of "layers" implementing the following non-linear function

$$h(x) = \sigma\left(Wx + b\right),$$

where $x \in \mathbf{R}^d$, $W \in \mathbf{R}^{m \times d}$, and $b \in \mathbf{R}^m$, and where m is often referred to as the number of **hidden units** or **hidden nodes**. σ is some non-linear function, typically tanh or ReLU, applied element-wise to the argument of σ. Referring to the computation graph illustration above, we will implement this nonlinear layer with two nodes, one implementing the affine transform $L = W_1 x + b_1$,

and the other implementing the nonlinear function $h = \tanh(L)$. In this problem, we'll work out how to implement the backward method for each of these nodes.

4.1.1 The Affine Transformation

In a general neural network, there may be quite a lot of computation between any given affine transformation $Wx + b$ and the final objective function value J. We will capture all of that in a function $f : \mathbf{R}^m \to \mathbf{R}$, for which $J = f(Wx + b)$. Our goal is to find the partial derivative of J with respect to each element of W, namely $\partial J/\partial W_{ij}$, as well as the partials $\partial J/\partial b_i$, for each element of b. For convenience, let $y = Wx + b$, so we can write $J = f(y)$. Suppose we have already computed the partial derivatives of J with respect to the entries of $y = (y_1, \ldots, y_m)^T$, namely $\frac{\partial J}{\partial y_i}$ for $i = 1, \ldots, m$. Then by the chain rule, we have

$$\frac{\partial J}{\partial W_{ij}} = \sum_{r=1}^{m} \frac{\partial J}{\partial y_r} \frac{\partial y_r}{\partial W_{ij}}.$$

1. Show that $\frac{\partial J}{\partial W_{ij}} = \frac{\partial J}{\partial y_i} x_j$, where $x = (x_1, \ldots, x_d)^T$. [Hint: Although not necessary, you might find it helpful to use the notation $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases}$. So, for examples, $\partial_{x_j} \left(\sum_{i=1}^{n} x_i^2 \right) = 2x_i\delta_{ij} = 2x_j$.]
2. Now let's vectorize this. Let's write $\frac{\partial J}{\partial y} \in \mathbf{R}^{m \times 1}$ for the column vector whose ith entry is $\frac{\partial J}{\partial y_i}$. Let's also define the matrix $\frac{\partial J}{\partial W} \in \mathbf{R}^{m \times d}$, whose ij'th entry is $\frac{\partial J}{\partial W_{ij}}$. Generally speaking, we'll always take $\frac{\partial J}{\partial A}$ to be an array of the same size ("shape" in numpy) as A. Give a vectorized expression for $\frac{\partial J}{\partial W}$ in terms of the column vectors $\frac{\partial J}{\partial y}$ and x. [Hint: Outer product.]
3. Show that $\frac{\partial J}{\partial b} = \frac{\partial J}{\partial y}$, where $\frac{\partial J}{\partial b}$ is defined in the usual way.

4.1.2 Element-wise Transformers

Our nonlinear activation function nodes take an array (e.g. a vector, matrix, higher-order tensor, etc), and apply the same nonlinear transformation $\sigma : \mathbf{R} \to \mathbf{R}$ to every element of the array. Let's abuse notation a bit, as is usually done in this context, and write $\sigma(A)$ for the array that results from applying $\sigma(\cdot)$ to each element of A. If σ is differentiable at $x \in \mathbf{R}$, then we'll write $\sigma'(x)$ for the derivative of σ at x, with $\sigma'(A)$ defined analogously to $\sigma(A)$.

Suppose the objective function value J is written as $J = f(\sigma(A))$, for some function $f : S \mapsto \mathbf{R}$, where S is an array of the same dimensions as $\sigma(A)$ and A. As before, we want to find the array $\frac{\partial J}{\partial A}$ for any A. Suppose for some A we have already computed the array $\frac{\partial J}{\partial S} = \frac{\partial f(S)}{\partial S}$ for $S = \sigma(A)$. At this point, we'll want to use the chain rule to figure out $\frac{\partial J}{\partial A}$. However, because we're dealing with arrays of arbitrary shapes, it can be tricky to write down the chain rule. Appropriately, we'll use a tricky convention: We'll assume all entries of an array A are indexed by a single variable. So, for example, to sum over all entries of an array A, we'll just write $\sum_i A_i$.

1. Show that $\frac{\partial J}{\partial A} = \frac{\partial J}{\partial S} \odot \sigma'(A)$, where we're using $\odot$ to represent the **Hadamard product**. If A and B are arrays of the same shape, then their Hadamard product $A \odot B$ is an array with the same shape as A and B, and for which $(A \odot B)_i = A_i B_i$. That is, it's just the array

formed by multiplying corresponding elements of A and B. Conveniently, in `numpy` if `A` and `B` are arrays of the same shape, then `A*B` is their Hadamard product.

4.2 MLP Implementation

1. Complete the class `AffineNode` in `nodes.py`. Be sure to propagate the gradient with respect to x as well, since when we stack these layers, x will itself be the output of another node that depends on our optimization parameters.
2. Complete the class `TanhNode` in `nodes.py`. As you'll recall, $\frac{d}{dx}\tanh(x) = 1 - \tanh^2 x$. Note that in the forward pass, we'll already have computed tanh of the input and stored it in self.out. So make sure to use `self.out` and not recalculate it in the backward pass.
3. Implement an MLP by completing the skeleton code in `mlp_regression.py` and making use of the nodes above. Your code should pass the tests provided in `mlp_regression.t.py`. Note that to break the symmetry of the problem, we initialize our weights to small random values, rather than all zeros, as we often do for convex optimization problems. Run the MLP for the two settings given in the `main()` function and report the average **training** error. Note that with an MLP, we can take the original scalar as input, in the hopes that it will learn nonlinear features on its own, using the hidden layers. In practice, it is quite challenging to get such a neural network to fit as well as one where we provide features.
4. [Optional] See if you can get a fit on the training set with an MLP that uses just the scalar input that is about as good as the fit using the featurized inputs. You can do that by tweaking model parameters (e.g. the number of hidden nodes or layers) and/or the parameters of optimization. You **may use** any neural network framework (PyTorch, TensorFlow, etc), which can help by providing more advanced optimization techniques (e.g. Adam), variable initialization methods, and/or various normalization approaches (batch norm, etc).

4.3 [OPTIONAL]

1. [Optional] Implement a Softmax node.
2. [Optional] Implement a negative log-likelihood loss node for multiclass classification.
3. [Optional] Use the classes above to apply an MLP to the simple multiclass classification dataset we had on a previous assignment.