Review for Chapter 4

(1) Title, author and where it was published
    - Title: Machine Learning
    - Author: Tom Mitchell
    - Published: McGraw Hill
(2) Summary of the content

The Artificial neural network (ANN) chapter mainly introduces two types: perceptron training and backpropagation algorithm, with detailed description of a neural network's origin, motivation, derivation, hypothesis space, implication, and a real-world application example. Perceptrons are powerful in classifying linearly separable datasets, and gradient descent method can deal with nonlinearly separable examples by converging toward a best-fit approximation to the target concept. With gradient descent technique, the scientists then develop the multilayer networks and the backpropagation algorithm where a more complex problem setting could be handled. Later, this chapter discusses topics like the strengths and weaknesses of this approach, what common heuristics could be applied to gain a more reliable estimate, and the representational meaning of the hidden layers. The chapter then concludes with an example from face recognition and some insights on other advanced improvements on ANN, which includes more powerful error functions, error minimization procedures, and recurrent networks.

To summarize backpropagation algorithm in short, it basically learns the weights for a multilayer network that minimize the squared error between the predicted and the target output, and it uses gradient descent to update weights. We first need to construct a neural network with a specified number of input nodes, output nodes, and hidden layers, then initialize a random weight vector. We then need to repeat the following steps until some pre-specified termination criteria are met: first compute the output, and then update the weights with the corresponding error terms.

(3) Strengths of the paper

This chapter not only thoroughly introduces the fundamentals of ANN and backpropagation algorithm, but also provides detailed interpretations in the representation of each component in the network. First of all, the author made it clear that even though the initial motivation of a neural network is to mimic the decision making process in a human brain with a faster computing power, its main objective is to find more advanced machine learning algorithms based on a neural network structure. The paper specifies that ANN learning is robust to noise but the problem needs to be representable by many attribute-value pairs. Specifically, to apply gradient descent, we need a hypothesis space that could be continuously parameterized and a loss function is differentiable with respect to those hypothesis parameters.

Moveover, although the paper claims that backpropagation algorithm is a powerful and prevalent technique in ANN, the author chooses to progressively walk the reader through the derivation of this rather complex concept – from a single perceptron to gradient descent, and then to multilayer network, and finally to backpropagation – making this piece reader-friendly and comprehensible. I would say introducing some easier concepts before delving into harder ones is a good technique in writing a paper since it will not intimidate the reader by the volume of mathematical derivation.

In addition, along with theoretical understanding of this algorithm, this paper also introduces some simple applications and visualizations of this algorithm that clearly illustrate the work done by each layer of a neural network. The graph depicting the "hidden unit encoding for input 01000000" demonstrates how the values gradually capture the "hidden" characteristics of the input data as the number of training iterations increases, emphasizing that backpropagation has the ability to invent some new features that are not provided or apparent in the input data. Furthermore, in the face recognition example, the author discusses several design choices made in analyzing a more complicated problem setting and states the corresponding effects that they brought to the result. For example, it is a good practice to reduce the input images from 120 x 128 pixels to 30 x 32 pixels since it requires less computational power while maintaining sufficient information.

Last but not the least, the paper justifies that the major defect of gradient descent is that it is not guaranteed to converge to the global minimum in the error surface, and if the learning rate is too large, it might not be able to converge as well. The author then provides the reader with three common heuristics to alleviate the problem of local minima, which includes adding a momentum term, using stochastic gradient descent approximation, and training the network multiple times with different initialization of the random weights. The last approach is effective since every possible set of weights in fact represents a syntactically distinct hypothesis. Another possible defect is overfitting, but this phenomenon is common to almost all machine learning algorithms. And the common approach is to perform cross-validation on the training data set: that is, separate the training set into training and validation sets so that the algorithms can verify its performance based on the outcome on unseen datasets.

(4) Major critiques

One topic that the paper failed to mention is the exploding gradient problem that might occur under certain contexts when error gradients accumulate to some very large values such that the weight update becomes explosive and eventually causes an overflow to the program. Moveover, this chapter seems to assume that there would always be a global minimum somewhere in the hypothesis space, yet it is not true in the real-world setting, and thus further complications might be added to this algorithm. Besides, based on my understanding, the batch size of the amount of data to be processed per iteration/epoch could also affect the performance of a neural network.

On the other hand, the backpropagation algorithm with gradient descent method is also a little outdated. It might require long training time and computational power on complicated problems to reach a sufficient outcome, yet there exist many other advanced machine learning approaches that could reach a higher accuracy with a shorter amount of time. With this in mind, we should only use this method for simpler problems with smaller data needed to be processed overall, like hand-written digits classification problem or cat-or-dog image classification problem.

Review for Chapter 6

(1) Title, author and where it was published
    (a) Title: Deep Learning
    (b) Author: Ian Goodfellow and Yoshua Bengio and Aaron Courville
    (c) Published: MIT Press
(2) Summary of the content

This chapter introduces the basic concepts of a feedforward network and the relevant design decisions we have to make when constructing this network. A feedforward network is mainly composed of several layers that capture the chain structure of a set of functions, and each unit inside a layer resembles a neuron that takes in a bunch of information and outputs a value determined by its related function and activation function. Later, gradient-based learning is introduced, and the author walks through the critical elements of the network separately, which include cost function, output units, hidden units, and architectural designs. Lastly, the author talks about the backpropagation algorithm and how it is used differently in deep learning and machine learning settings.

(3) Strengths of the paper

First of all, the example on learning XOR function provides the reader with a clear picture of the effect of activation functions – that we are transforming the nonlinear dataset into a space where linear functions could solve. It also demonstrates the representational power that hidden layers and the output layer have.

Moveover, when discussing convoluting concepts, the author first provides the reader with the most common solution, and then maybe talks about other methods, ensuring that the reader gets the gist of this section while not being intimidated by the mathematical complexity behind. For example, in the cost function section, the author emphasizes that maximum likelihood is the most popular choice, and the simple idea is to capture the cross-entropy between the training data and the model distribution. In general, if we define a conditional distribution based on the problem, the principle of maximum likelihood suggests that we should use the negative log of that conditional distribution as the cost function, and it would generally lead to satisfying results. In addition, in the output units section, the author also presents three approaches to three types of output distributions and explains the corresponding mathematical details and implications, from the most comprehensible example (linear units) to a harder one (softmax units).

(4) Major critiques

As noted in the paper, backpropagation needs a large amount of memory as it often involves summation of many tensors, and so the machine costs could be expensive. Besides, it also requires a large amount of data in order to produce a satisfactory outcome since this ANN is heavily based on data fed. Meanwhile, the most common critique for a deep learning approach is the “black box” inside the hidden layers. The author briefly touches on several hidden units like rectified linear units and sigmoid units, yet we still do not know what happens in between units and layers of the complex network. It is clear that further research should be done on this area.