LETTER Communicated by Jurgen Schmidhuber

Solving Nonlinearly Separable Classifications in a Single-Layer Neural Network

Nolan Conaway
nconaway@wisc.edu
Kenneth J. Kurtz
kkurtz@binghamton.edu
Department of Psychology, Binghamton University, Binghamton, NY 13903, U.S.A.

Since the work of Minsky and Papert (1969), it has been understood that single-layer neural networks cannot solve nonlinearly separable classifications (i.e., XOR). We describe and test a novel divergent autoassociative architecture capable of solving nonlinearly separable classifications with a single layer of weights. The proposed network consists of class-specific linear autoassociators. The power of the model comes from treating classification problems as within-class feature prediction rather than directly optimizing a discriminant function. We show unprecedented learning capabilities for a simple, single-layer network (i.e., solving XOR) and demonstrate that the famous limitation in acquiring nonlinearly separable problems is not just about the need for a hidden layer; it is about the choice between directly predicting classes or learning to classify indirectly by predicting features.

1 Introduction

One of the first things anyone learns about artificial neural networks is that a single-layer network can solve only linearly separable classification problems. This essential bit of dogma in the field is attributed to the famous/infamous book *Perceptrons* by Minsky and Papert (1969), which includes the canonical demonstration that a standard single-layer perceptron (Rosenblatt, 1958), directly associating dimensional inputs with class labels, cannot learn the exclusive-OR (XOR) function. This resulted in a reduction in neural network research until the popularization of the backpropagation algorithm for training multilayer networks (Rumelhart, Hinton, & Williams, 1986; see Schmidhuber, 2015, for a full treatment of the discovery of backpropagation). Today, such multilayer perceptrons (MLPs) maintain considerable currency, especially with the rise of deep learning techniques that use multiple hidden layers.

N. C. is now at the University of Wisconsin–Madison.

Neural Computation **29, 861–866** (2017)
doi:10.1162/NECO_a_00931

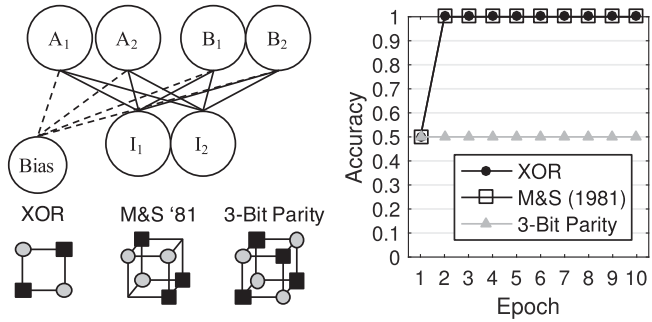

Figure 1: (Top left) Depiction of the network. (Bottom left) Illustration of NLS classifications. (Right) Model performance on various NLS classifications.

In this letter, we show how a simple single-layer network with standard delta rule learning (Widrow & Hoff, 1960) can successfully learn nonlinearly separable (NLS) classification problems. The solution lies with how the classification task is realized in the network architecture. As opposed to the canonical conception of a network trained to predict the class from the features, we train the network to predict the features with respect to each class. Rather than unsuccessfully optimizing a set of weights to discriminate between classes that are not subject to a linear discriminant, the network we present instead learns weights that linearly approximate the internal structure of each class and uses each item's consistency with that structure as the basis for classification.

The divergent autoencoder (DIVA) is a cognitive model of human category learning (Kurtz, 2007, 2015) developed based on this core design principle: a separate channel of autoassociative learning for each class with a shared hidden layer. For each training item, the hidden → output weights are updated only along the correct category channel, and the shared input → hidden weights are updated every time. We present a novel extension in the form of a neural network that uses the divergent autoassociative approach to classification without the hidden layer (see Figure 1). Inputs nodes are fully connected to the autoassociative output channels for each class (note that this is equivalent to separate linear autoassociators for each class) such that network output is calculated as

$$O_{kc} = \sum_{1}^{j} I_j W_{jkc}, \tag{1.1}$$

where I is a vector of inputs and W is a j-by-k-by-c weight matrix connecting input dimensions j to output dimensions k along category channels c. The simplest possible response rule is used to translate the output to a classification outcome by selecting the channel with the most accurate

reconstruction measured as mean-squared error (MSE) between the network input I and class reconstruction O (note that MSE is not evaluated along the bias unit):

$$MSE_c = \frac{1}{k}\sum_{1}^{k}(I_k - O_{kc})^2. \tag{1.2}$$

To be clear, any neural net classifier compares the output activation level relative to the target for each class in order to select a response. In a traditional perceptron architecture, the activation level of one node is shorthand for the performance error of the network on that class; in a divergent autoassociator, the performance error across the nodes dedicated to that class is used, and the shortcut does not apply. The response rule is independent of the network's learning procedure and is used only to obtain class predictions. We thank an anonymous reviewer for noting that support vector machines (Cortes & Vapnik, 1995), echo state, and liquid state machines (Jaeger, 2001; Maass, Natschläger, & Markram, 2002) overcome limitations of single-layer neural networks by a nonlinear prewired preprocessor and that the response rule of our neural network may be viewed as a conceptually similar prewired postprocessor. Weights are updated using delta rule learning to minimize reconstruction error on the channel of each item's correct class,

$$\Delta W_{jkc} = \alpha(I_k - O_{kc})I_j, \tag{1.3}$$

where α is a learning rate parameter.

2 Simulations

We conducted three tests of the proposed network on NLS classification problems using 10 epochs of training (batch update) with zero-valued initial weights and a learning rate of 0.1. All problems were based on binary input features set to values of ± 1. The network was fully connected (including weights between the input and output nodes coding for the same dimension) and used linear output units. Note that the qualitative performance of the network is highly consistent across a variety of design choices, including small random initial weights, learning rate values, online (trial) weight update, output unit activation function (i.e., linear / logistic), and connectivity pattern (i.e., with or without same-dimension connections).

Our primary focus was the XOR problem central to the critique made by Minsky and Papert (1969). In addition, we tested the model on an NLS problem used in an influential psychological study of human category learning (Medin & Schwanenflugel, 1981, experiment 4). This problem consists of six items divided into two categories: [$- - +, - + +, + - -$] and [$- + -, + - +, + + -$] (see Figure 1 for a visualization). We also included the three-bit

parity problem (see Figure 1) that is notable for lacking any second-order statistical regularities within or between classes (also known as type VI from Shepard, Hovland, & Jenkins, 1961). We expected the network to be unconstrained by the linear separability of classes because it is not directly trained to make a class discrimination; instead it is learning within-class feature relationships. The network was not expected to perform well on three-bit parity since there are no such regularities to learn.

3 Results

The single-layer network easily mastered the XOR problem (see the results in Figure 1). The network learns positive weights on the same-dimension connections (since each dimension predicts itself), but this merely promotes copying of any input pattern on both channels. Successful solution of the classification problem lies with the weights connecting different dimensions along each autoassociative channel (e.g., $I_1 \rightarrow O_2$). In order to predict the input activations at the output layer, these weights for category 1 [$--$ and $++$] become positive, while the weights for category 2 [$-+$ and $+-$] become negative. Each category channel therefore learns a representation enabling accurate reconstruction of its own exemplars, while producing poor reconstructions of exemplars from other classes, effectively solving the classification.

In the remaining tests, we sought to extend beyond the XOR problem and establish boundary conditions. We found that the network mastered the Medin and Schwanenflugel (1981) NLS classification but not the three-bit parity problem (see Figure 1). In both cases, the network learns positive weights for the same-dimension connections, but these do not provide the solution. In the Medin and Schwanenflugel (1981) NLS problem, the network relies on a critical within-category regularity: two perfectly opposed dimensions (I_1 and I_3 in category 1 and I_2 and I_3 in category 2). The network learns negative weights connecting these dimensions and a negative bias for the remaining dimension. For the three-bit parity problem, there are no first- or second-order regularities. Accordingly, the reconstructions are equally good for each channel, and the classification problem is not solved. This set of simulations reveals a general principle governing the network's classification performance: the network succeeds when the classes have nonaligned, within-category regularities in the form of correlations on different dimensions (as in the Medin and Schwanenflugel problem) or correlations in opposing directions on the same dimensions (as in XOR).

4 Discussion

These results allow us to draw a fairly precise conclusion: a single-layer network using divergent autoassociative learning solves classification problems without being constrained by linear separability. To be clear, we have

not altered any of the core computational tools commonly put to use in neural network modeling. The advance lies with reformulating classification in terms of feature prediction rather than direct class prediction—a shift that alters a fundamental performance constraint because the network is no longer addressing the classification problem through an explicit discriminant boundary between classes. But while the linear separability of the classes is unimportant to the learning of a class-specific autoassociator, our network is limited by a different type of constraint: the classes must have nonaligned, within-class regularities. Notably, the failure of a single-layer network to distinguish between classes that have matched regularities or no second-order regularities is not a terrible limitation for a simple classifier.

It is possible to interpret the network's performance under a number of theoretical frameworks—for example, as a Bayesian discrimination or gaussian density estimation. Our primary concern lies within the neural network field where there is a canonical conception regarding the limitations of neural networks without hidden layers. Our results show that this view is in fact a product of the typical instantiation of classification tasks in an architecture with features as inputs and classes as outputs. As a generative rather than discriminative method for classification learning (Ng & Jordan, 2002), the divergent autoassociative approach of predicting features on the way to predicting class alters our most basic assumptions about what a standard single-layer network can do.

References

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Jaeger, H. (2001). *The "echo state" approach to analysing and training recurrent neural networks-with an erratum note* (GMD Technical Report 148). Bonn: German National Research Center for Information Technology.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, *14*(11), 2531–2560.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14* (pp. 841–848). Cambridge, MA: MIT Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *Western Electronic Show and Convention: Convention Record* (vol. 4, pp. 96–104). New York: Institute of Radio Engineers.

Received November 20, 2015; accepted October 26, 2016.