COMPSCI 689
Lecture 11: Multilayer Perceptrons

Benjamin M. Marlin

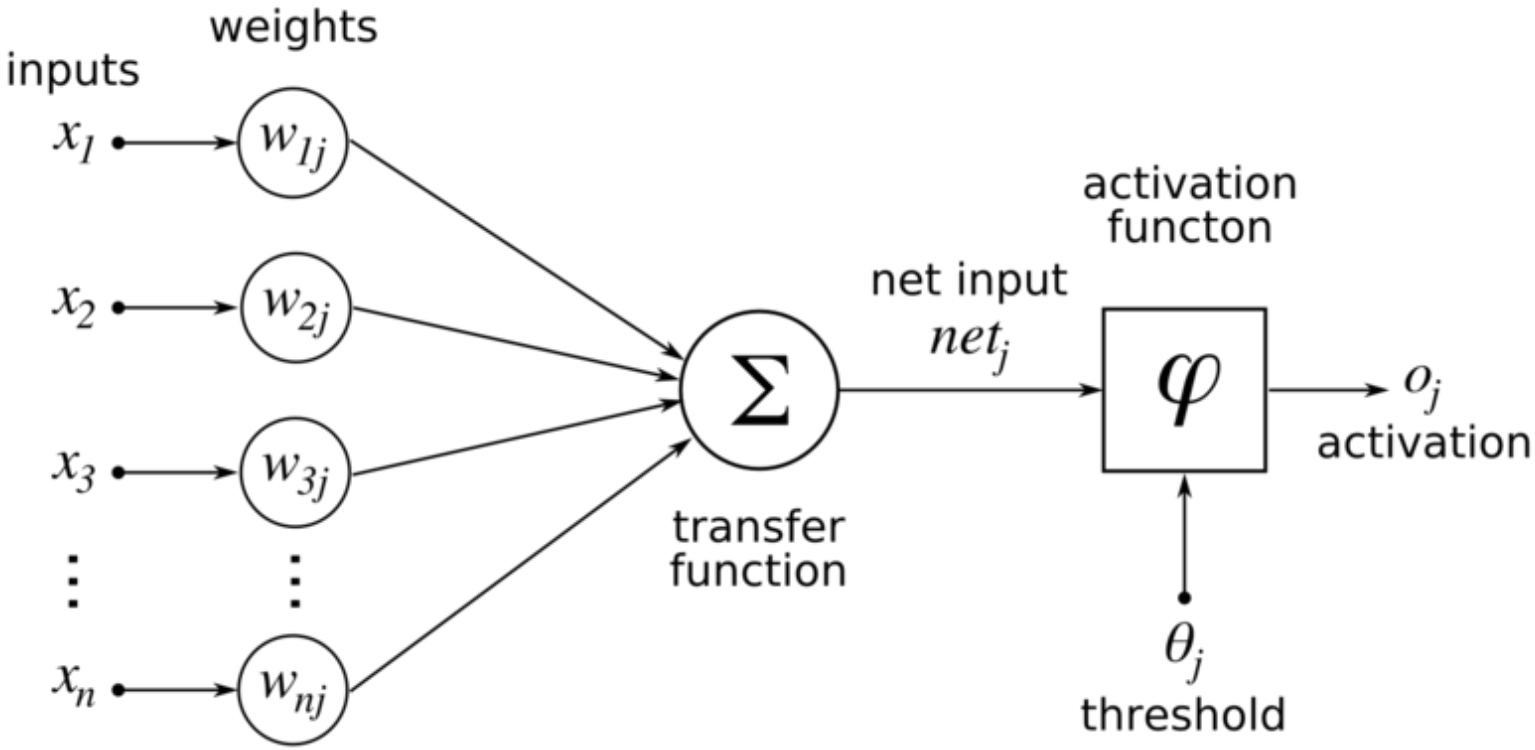
College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

The Trouble with Kernels

- We have previously seen how basis expansions and kernels can be used to increase the capacity of linear models, turning them into models that are non-linear in the features while remaining linear in the parameters.
- **Question:** What is the primary weakness of this approach?
- In this lecture, we'll see how neural networks can learn appropriate feature representations from data at the same time they learn to solve regression and classification problems.
- We'll start with some historical background...

McCulloch and Pitts Neuron (1943)



Assuming $\varphi(x) = \text{sign}(x)$, what model is this? Assuming $\varphi(x) = x$, what model is this?
Assuming $\varphi(x) = 1/(1 + \exp(-x))$, what model is this?

The Perceptron (1950)

The Perceptron is a simple online algorithm for adapting the weights in a McCulloch/Pitts neuron. It was developed in the 1950s by Rosenblatt at Cornell.

Algorithm PERCEPTRONTRAIN(**D**, *MaxIter*)

1: $w_d \leftarrow 0$, for all $d = 1 \ldots D$ // initialize weights
2: $b \leftarrow 0$ // initialize bias
3: **for** $iter = 1 \ldots MaxIter$ **do**
4: **for all** $(x,y) \in \mathbf{D}$ **do**
5: $a \leftarrow \sum_{d=1}^{D} w_d \, x_d + b$ // compute activation for this example
6: **if** $ya \leq 0$ **then**
7: $w_d \leftarrow w_d + yx_d$, for all $d = 1 \ldots D$ // update weights
8: $b \leftarrow b + y$ // update bias
9: **end if**
10: **end for**
11: **end for**
12: **return** $w_0, w_1, \ldots, w_D, b$

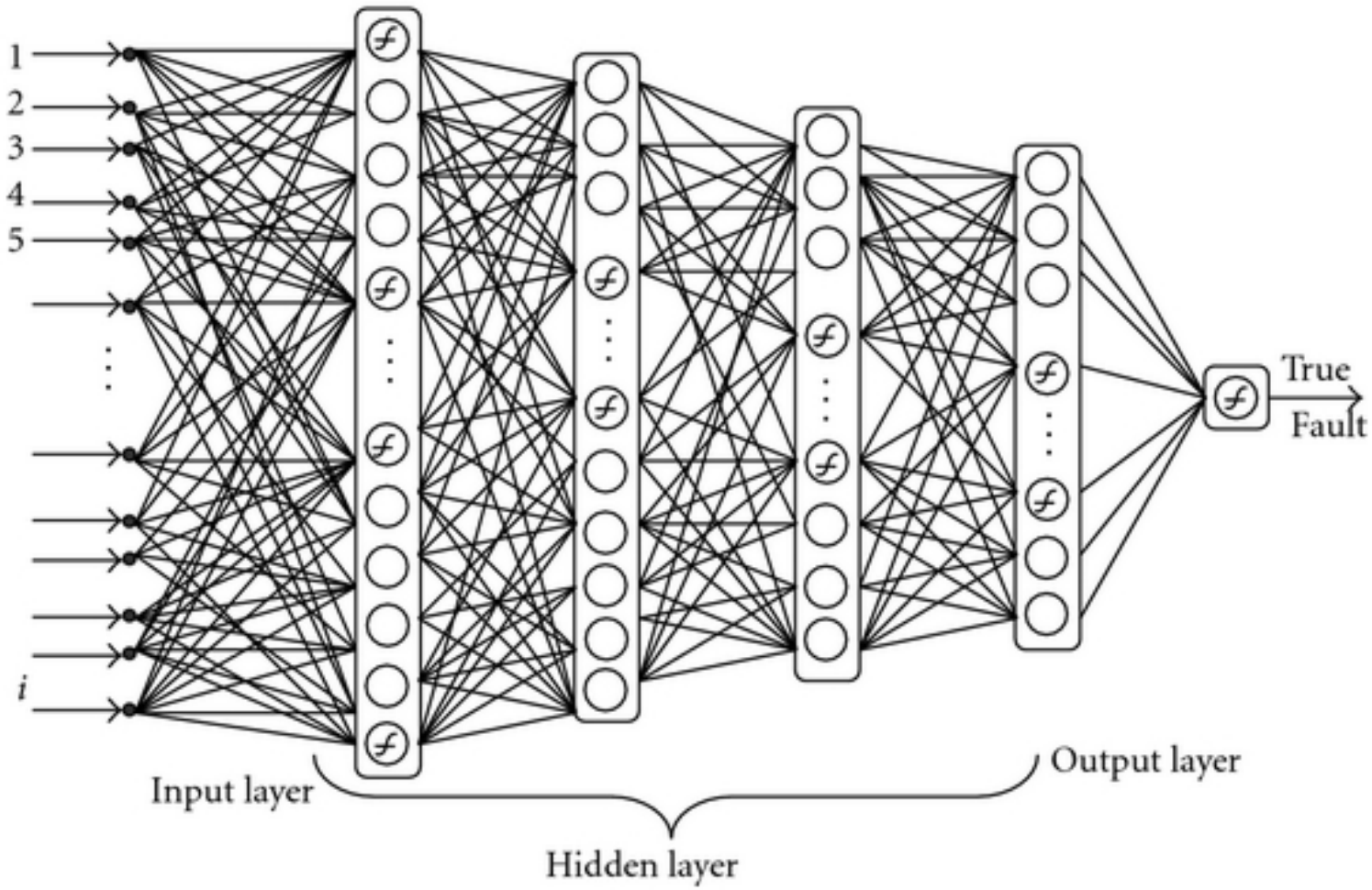
Perceptron Limitations

- This particular algorithm converges to a random separating hyperplane when the data are linearly separable.
- However, the algorithm is not derived from optimizing an objective function and does not converge when data are not linearly separable.
- ERM/RRM for linear models typically gives classification results with improved generalization performance compared to this approach in the separable case, and can also handle the non-separable case correctly.

Perceptron Representational Limitations

- In 1969, Minsky and Papert at MIT popularized a set of arguments showing that the single-layer perceptron could not learn certain classes of functions (including XOR).
- However, Minsky and Papert also showed that more complex functions could be represented using a *multi-layer perceptron* or MLP.
- Unfortunately, at the time, no algorithms were known that could learn such networks from data.

Multi-Layer Perceptron



Sigmoid Neural Networks (1980s)

The eventual solution to MLP learning for classification was to:

1. Make the hidden layer non-linearities smooth (sigmoid/logistic) functions.
2. Make the output layer non-linearity a smooth (sigmoid/logistic/softmax) function.
3. Use a differentiable loss function.
4. Learn using the *backpropagation algorithm*, which was popularized by Rumelhart, Hinton and Williams in the 1980s. Backprop is just vanilla gradient descent on the ERM objective defined by the loss.

Example: 1-Hidden Layer Sigmoid NN Classifier

- The simplest binary classification architecture of this type is a 1-hidden layer sigmoid neural network:

$$h_k = \frac{1}{1 + \exp(-(\mathbf{w}_k^1 \cdot \mathbf{x}^T + b_k^1)}$$

$$\hat{y} = \frac{1}{1 + \exp(-(\mathbf{w}^o \cdot \mathbf{h}^T + b^o)}$$

- We assume K hidden units and D input features. The parameters of the model are a weight vector $\mathbf{w}_k^1 \in \mathbb{R}^D$ and bias $b_k \in \mathbb{R}$ for each hidden unit k, and a weight vector $\mathbf{w}^o \in \mathbb{R}^K$ and a bias b^o for the output.

- h_k is the value of the k^{th} hidden unit. $\mathbf{h} = [h_1, ..., h_K]$. $\hat{y}$ is the output of the network, which can be interpreted as a probability.

Binary Sigmoid NN Learning Problem

- Binary Sigmoid NN Classifiers can be learned in the ERM framework using the binary cross entropy loss (assuming $y \in \{0, 1\}$: $L_{ce}(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$:
- For the 1-hidden layer sigmoid neural network model with parameters $\theta = [\mathbf{w}^1_{1:K}, b^1_{1:K}, \mathbf{w}^o, b^o]$, the learning problem is to minimize $\mathcal{L}(\mathcal{D}, \theta)$ with respect to θ:

$$\mathcal{L}(\mathcal{D}, \theta) = -\sum_{n=1}^{N} (y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n))$$
$$h_{kn} = \sigma(\mathbf{w}^1_k \cdot \mathbf{x}^T_n + b^1_k)$$
$$\hat{y}_n = \sigma(\mathbf{w}^o \cdot \mathbf{h}^T_n + b^o)$$

- Where $\sigma(x) = 1/(1 + \exp(-x))$.

Binary Sigmoid NN Classifier Equivalencies

- This learning approach is equivalent to defining the output of the model to be $\hat{y}' = \mathbf{w}^o \mathbf{h}^T + b^o$ and learning the model using the logistic loss (assuming $y \in \{-1, 1\}$.
- It is also equivalent to learning a conditional Bernoulli model in the MLE framework where the mean function $\mu(\mathbf{x}) = \hat{y}$.
- Finally, the model can be viewed as simultaneously learning a basis expansion and a logistic regression classifier in this new basis.

Gradients for Binary Sigmoid NN Classifiers

- First, $\sigma'(x) = \sigma(x)(1-\sigma(x))$
- $\frac{\partial \mathcal{L}(\mathcal{D},\theta)}{\partial \hat{y}_n} = -y_n/(\hat{y}_n) + (1-y_n)/(1-\hat{y}_n) = \frac{\hat{y}_n - y_n}{\hat{y}_n(1-\hat{y}_n)}$
- $\frac{\partial \hat{y}_n}{\partial w_k^o} = \sigma'(\mathbf{w}^o \mathbf{h}_n^T + b^o)\mathbf{h}_{nk} = \hat{y}_n(1-\hat{y}_n)\mathbf{h}_{nk}$
- $\frac{\partial \hat{y}_n}{\partial h_{kn}} = \sigma'(\mathbf{w}^o \mathbf{h}_n^T + b^o)\mathbf{w}_k^o = \hat{y}_n(1-\hat{y}_n)\mathbf{w}_k^o$
- $\frac{\partial h_{kn}}{\partial w_d^1} = \sigma'(\mathbf{w}_k^1 \mathbf{x}_n^T + b_k^1)\mathbf{x}_{dk} = h_{kn}(1-h_{kn})\mathbf{x}_{dn}$
- Thus $\frac{\partial \mathcal{L}(\mathcal{D},\theta)}{\partial w_k^o} = \sum_{n=1}^N \frac{\partial \mathcal{L}(\mathcal{D},\theta)}{\partial \hat{y}_n}\frac{\partial \hat{y}_n}{\partial w_k^o} = \sum_{n=1}^N (\hat{y}_n - y_n)\mathbf{h}_{nk}$ and,
- $\frac{\partial \mathcal{L}(\mathcal{D},\theta)}{\partial w_d^1} = \sum_{n=1}^N \frac{\partial \mathcal{L}(\mathcal{D},\theta)}{\partial h_{kn}}\frac{\partial h_{kn}}{\partial w_d^1} = \sum_{n=1}^N (\hat{y}_n - y_n)\mathbf{w}_k^o h_{kn}(1-h_{kn})\mathbf{x}_{dn}$
- Nearly identical results hold with respect to the biases.

Learning for Sigmoid NN Regression

- Learning sigmoid neural networks for regression follows the same basic pattern.
- All we need to do is change the output $\hat{y}_n$ to be a linear function of the hidden units, and switch the loss to squared error.
- The derivation of the gradient is nearly identical regardless of the loss used.

Regularization

- Like with linear regression, logistic regression, and SVMs, neural networks typically need to be regularized to give good generalization performance.
- Standard regularizers can be added to the objective function to accomplish this. (i.e. $\|\mathbf{w}\|_2^2$).
- In the neural networks literature, the addition of two-norm regularization is sometimes referred to as *weight decay* because of the form of the resulting gradients.

Optimization Details

- **Optimizers:** Second order optimizers are rarely used to learn neural network models. Instead, first-order methods are typically used (i.e., gradient descent).
- **Stochastic Approximation:** It is also typical to compute gradientss using sub-sets of the data. This typically speeds convergence. In this literature, this is often referred to as *mini-batch* learning or *stochastic gradient descent*.
- **Step Sizes:** Fixed step sizes or fixed step size decay schedules were often used in early work. In this literature, the step size is often referred to as the *learning rate*.
- **Acceleration:** There are a wide array of approaches that either modify the gradient direction or the step sizes to attempt to achieve faster convergence on certain classes of problems (i.e, use of momentum, Adadelta, Adagrad, Adam, RMSprop, etc.).

Sigmoid Neural Network Limitations

- Unlike GLMs and SVMs, multi-layer sigmoid neural network models have many local optima.
- In practice, it can be hard to find a combination of layer sizes and regularization that yields good generalization performance when learning from even moderate amounts of data.
- In addition, the use of sigmoid functions for the hidden units results in a *vanishing gradient* problem that can further slow down learning for deep networks.
- As a result, there were few examples of deep sigmoid neural network models that resulted in perfromance that exceeded hand-crafted features, basis expansions and kernels until relatively recently.

Fixing Deep Sigmoid Neural Network Learning

1. Solve the vanishing gradient problem using rectified linear units: $relu(x) = max(0, x)$.
2. Have access to lots of labeled data (ie: millions of examples).
3. Do the computing on GPUs to achieve significant speedups (i.e.: model training takes 10 days for large computer vision problems instead of 1 year).

Deep Learning Modeling Tools

- As we have seen, deep learning models just require (sub-) gradients of the loss with respect to the parameters to enable learning.
- However, correctly deriving gradients from complex model architectures by hand can be tedious and is error prone.
- As a result, many deep learning tools exist (i.e., Theano, Torch, TensorFlow, Caffe, Keras) that allow models to be specified only in terms of the *forward pass* computation (from inputs to loss).
- This makes it much easier to quickly change the model architecture because only the specification of the forward pass needs to be updated.

Autodiff

- An idea called *automatic differentiation* is used to convert the forward pass specification into the gradient computations needed for learning.
- A computation graph is extracted from the specification of the forward pass.
- The chain rule is then applied to each node in the computation graph to transform it into a computation graph for computing the gradient of the objective.
- Importantly, this procedure yields analytic gradients and learning is identical to classical backprop.
- As a byproduct of the representations used, it's possible to compile both computation graphs against arbitrary numerical libraries for either CPUs or GPUs.