18-661 Introduction to Machine Learning

Neural Networks-I

Fall 2020

ECE – Carnegie Mellon University

Announcements

- Homework 5 has been released and is due on Nov. 6.
- Lecture on Nov. 3 is canceled due to Election Day.

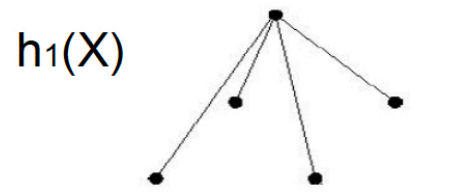
Outline

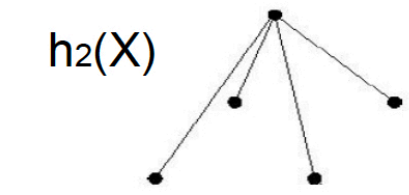
1. Recap: Ensemble Methods
2. Neural Networks: Motivation
3. Perceptrons
4. Neural Network Architectures and Forward Propagation
5. Choosing Activation Units
6. Choosing Neural Network Architectures

Recap: Ensemble Methods

Ensemble methods

- Instead of learning a single (weak) classifier, learn many weak classifiers, preferably those that are good at different parts of the input spaces
- Predicted Class: (Weighted) Average or Majority of output of the weak classifiers
- Strength in Diversity!

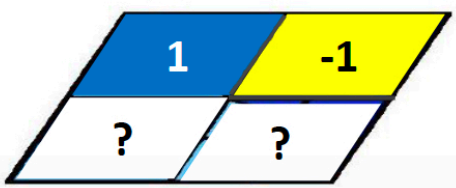


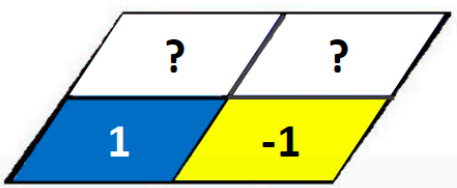


$H: X \rightarrow | Y\ (-1,1)$

$H(X) = h_1(X) + h_2(X)$

$H(X) = \text{sign}(\sum_t \alpha_t h_t(X))$



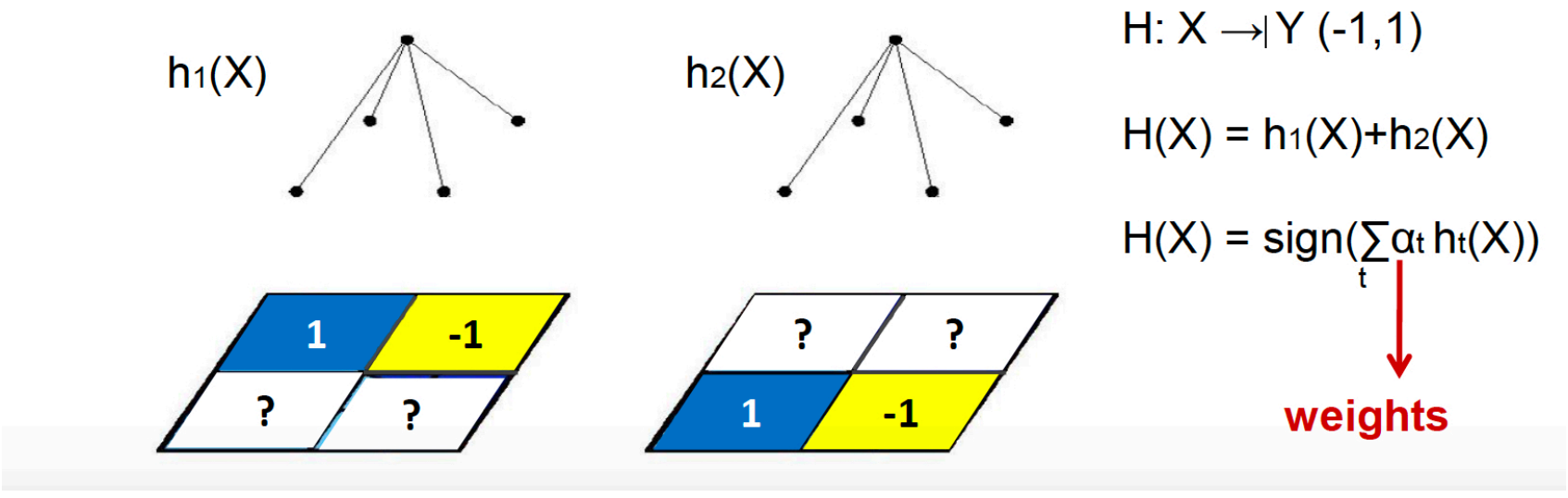


weights

Ensemble methods

In last lecture, we covered the following ensemble methods:

- Bagging or Bootstrap Aggregation
- Random Forests
- AdaBoost



Bagging or Bootstrap Aggregating

To avoid overfitting a decision tree to a given dataset we can average an ensemble of trees learnt on random subsets of the training data.

Bagging Trees (Training Phase)

- For $b = 1, 2, \cdots, B$
  - Choose n training samples $(\mathbf{x}_i, y_i)$ from $\mathcal{D}$ uniformly at random
  - Learn a decision tree h_b on these n samples
- Store the B decision trees $h_1, h_2, \ldots h_B$
- Optimal B (typically in 1000s) chosen using cross-validation

Bagging Trees (Test Phase)

- For a test unlabeled example $\mathbf{x}$
- Find the decision from each of the B trees
- Assign the majority (or most popular) label as the label for $\mathbf{x}$

Random Forests

- Limitation of Bagging: If one or more features are very informative, they will be selected by almost every tree in the bag, reducing the diversity (and potentially increasing the bias).
- Key Idea behind Random Forests: Reduces correlation between trees in the bag without increasing variance too much
  - Same as bagging in terms of sampling training data
  - Before each split, select $m \leq d$ features at random as candidates for splitting $m \sim \sqrt{d}$
  - Take majority vote of B such trees

Limitations of Bagging and Random Forests

- Bagging: Significant correlation between trees that are learnt on different training datasets
- Random Forests try to resolve this by doing "feature bagging" but some correlation still remains
- All B trees are given the same weight when taking the average

Boosting methods: Force classifiers to learn on different parts of the feature space, and take their *weighted* average

Adaboost algorithm

- Given: N samples $\{\boldsymbol{x}_n, y_n\}$, where $y_n \in \{+1, -1\}$, and some way of constructing weak (or base) classifiers
- Initialize weights $w_1(n) = \frac{1}{N}$ for every training sample n
- For $t = 1$ to T
  1. Train a weak classifier $h_t(\boldsymbol{x})$ using current weights $w_t(n)$, by minimizing

$$\epsilon_t = \sum_n w_t(n) \mathbb{I}[y_n \neq h_t(\boldsymbol{x}_n)] \quad \text{(the weighted classification error)}$$

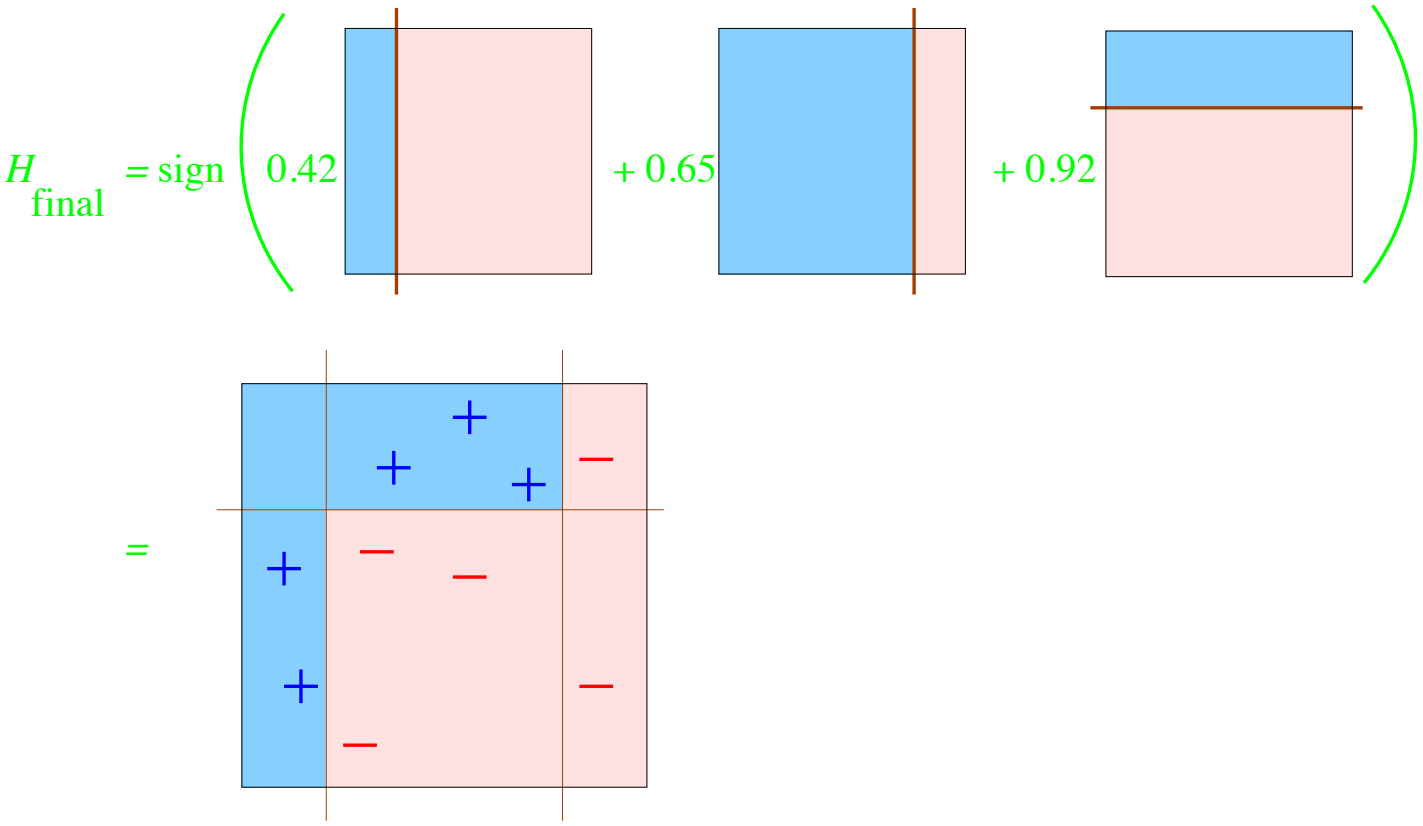
  2. Compute contribution for this classifier: $\beta_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
  3. Update weights on each training sample n

$$w_{t+1}(n) \propto w_t(n) e^{-\beta_t y_n h_t(\boldsymbol{x}_n)}$$

     and normalize them such that $\sum_n w_{t+1}(n) = 1$.
- Output the final classifier

$$h[\boldsymbol{x}] = \text{sign}\left[\sum_{t=1}^{T} \beta_t h_t(\boldsymbol{x})\right]$$

Example: Combining 3 classifiers



- All data points are classified correctly!

Why does AdaBoost work?

It minimizes a loss function related to classification error.

Classification loss

- Suppose we want to have a classifier

$$h(\boldsymbol{x}) = \text{sign}[f(\boldsymbol{x})] = \begin{cases} 1 & \text{if } f(\boldsymbol{x}) > 0 \\ -1 & \text{if } f(\boldsymbol{x}) < 0 \end{cases}$$

- One seemingly natural loss function is 0-1 loss:

$$\ell(h(\boldsymbol{x}), y) = \begin{cases} 0 & \text{if } yf(\boldsymbol{x}) > 0 \\ 1 & \text{if } yf(\boldsymbol{x}) < 0 \end{cases}$$

- AdaBoost uses the surrogate exponential loss:

$$\ell^{\text{EXP}}(h(\boldsymbol{x}), y) = e^{-yf(\boldsymbol{x})}$$

Choosing the t-th classifier

Suppose a classifier $f_{t-1}(\boldsymbol{x})$, and want to add a weak learner $h_t(\boldsymbol{x})$

$$f(\boldsymbol{x}) = f_{t-1}(\boldsymbol{x}) + \beta_t h_t(\boldsymbol{x})$$

Note: $h_t(\cdot)$ outputs -1 or 1, as does sign $[f_{t-1}(\cdot)]$

How can we 'optimally' choose $h_t(\boldsymbol{x})$ and combination coefficient β_t?

Adaboost greedily *minimizes the exponential loss function*!

$$(h_t^*(\boldsymbol{x}), \beta_t^*) = \text{argmin}_{(h_t(\boldsymbol{x}), \beta_t)} \sum_n e^{-y_n f(\boldsymbol{x}_n)}$$

$$h_t^*(\boldsymbol{x}) = \text{argmin}_{h_t(\boldsymbol{x})} \epsilon_t = \sum_n w_t(n) \mathbb{I}[y_n \neq h_t(\boldsymbol{x}_n)]$$

$$\beta_t^* = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t}$$

where we have used $w_t(n)$ as a shorthand for $e^{-y_n f_{t-1}(\boldsymbol{x}_n)}$ up to normalization

Updating the weights

Once we find the optimal weak learner we can update our classifier:

$$f(\boldsymbol{x}) = f_{t-1}(\boldsymbol{x}) + \beta_t^* h_t^*(\boldsymbol{x})$$

We then need to compute the weights for the above classifier as:

$$w_{t+1}(n) \propto e^{-y_n f(\boldsymbol{x}_n)} = e^{-y_n[f_{t-1}(\boldsymbol{x}_n)+\beta_t^* h_t^*(\boldsymbol{x}_n)]}$$

$$= w_t(n) e^{-y_n \beta_t^* h_t^*(\boldsymbol{x}_n)} = \begin{cases} w_t(n) e^{\beta_t^*} & \text{if } y_n \neq h_t^*(\boldsymbol{x}_n) \\ w_t(n) e^{-\beta_t^*} & \text{if } y_n = h_t^*(\boldsymbol{x}_n) \end{cases}$$

Intuition Misclassified data points will get their weights increased, while correctly classified data points will get their weight decreased

Ensemble Methods: Summary

Fight the bias-variance tradeoff by combining (by a weighted sum or majority vote) the outputs of many weak classifiers.

Bagging trains several classifiers on random subsets of the training data. Random forests train several decision trees that are constrained to split on random subsets of data *features*.

- Prevents some correlation between trees due to dominant features.
- All classifiers are weighted equally in making the final decision.

Boosting sequentially adds weak classifiers and optimizes their contributions, increasing the weight of "hard" data points at each step.

- Greedily minimizes a surrogate classification loss
- Commonly uses decision trees as base classifiers

Outline

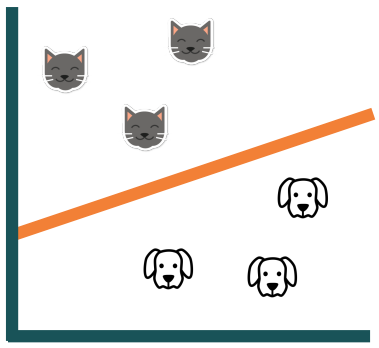
1. Recap: Ensemble Methods
2. Neural Networks: Motivation
3. Perceptrons
4. Neural Network Architectures and Forward Propagation
5. Choosing Activation Units
6. Choosing Neural Network Architectures

Neural Networks: Motivation

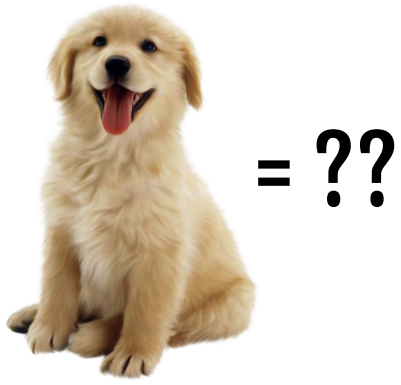
Why a "neural network"?

Many machine learning problems are easy for people but hard for machines.

input: cats and dogs



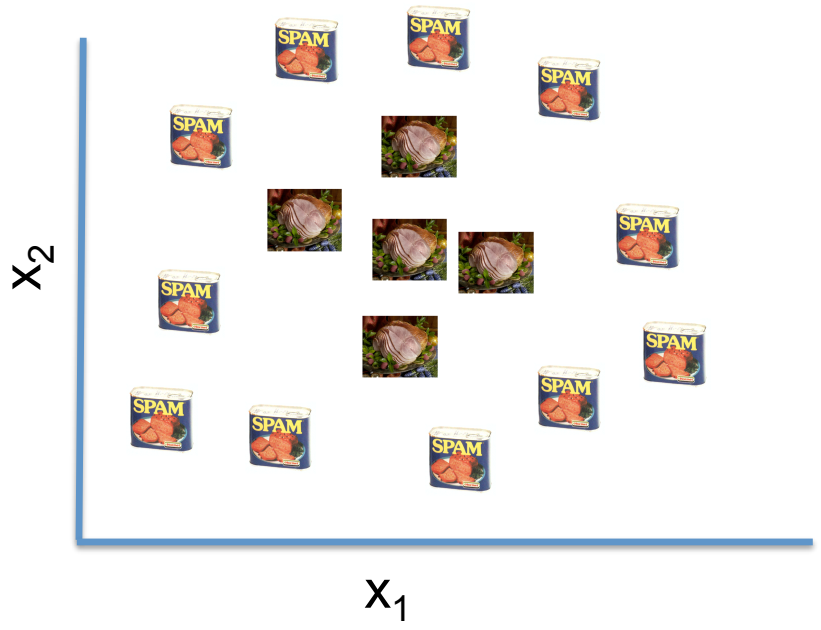
learn: $x \rightarrow y$ relationship



predict: y (*categorical*)

Each "node" in a neural network models a neuron in your brain. With enough neurons, we can learn to do very complex things.

Logistic Regression: How to handle complex boundaries?



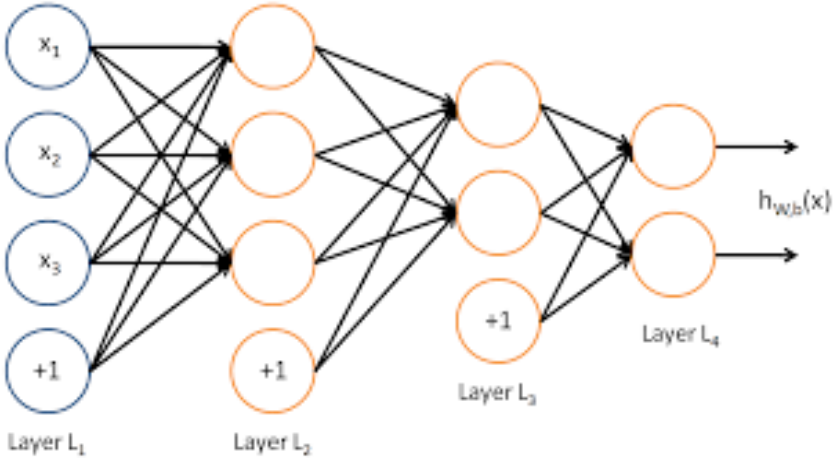
- This data is not linearly separable
- Use non-linear basis functions to add more features...

But what if we had a large number of features?

Each feature x_i is one pixel in an 100×100 input image

- Adding nonlinear (e.g. polynomial) features would result in an enormous $\phi(\mathbf{x})$
- Can we somehow only retain the important features?
- We will need to carefully hand-pick them, which can be hard and tedious...
- Neural networks automate this for us!

Neural Networks compress the set of features



- Start with feature vector **x** containing all pixels in the image
- Layer 1: distill the edges of the image
- Layer 2: distill triangles, circles, etc.
- Layer 3: recognize pointy ears, fur style etc.
- Layer 4: performs logistic regression on the features in layer 3

We cannot directly control what each layer learns; this depends on the training data.

Inspiration from Biology: How does our brain work?

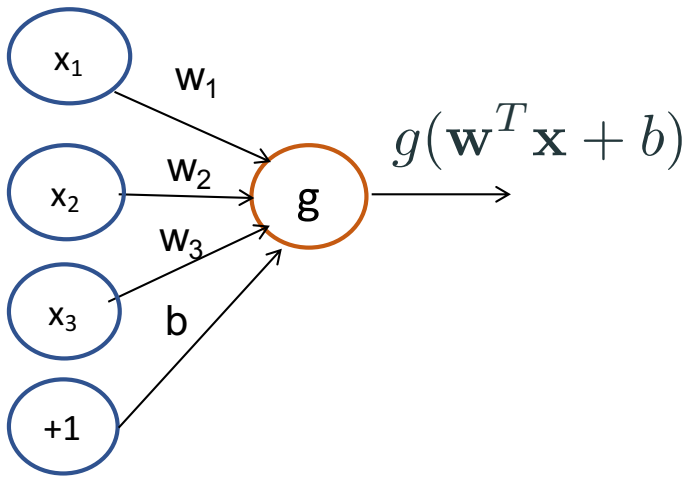
Each feature x_i is one pixel in an 100×100 input image

- Humans easily perform complex image or speech recognition tasks
- We cannot exactly describe a set of rules by which we distinguish cats vs. dogs, but we almost always know the correct answers when a new image is presented to us

Artificial Neuron model

Based on the biological insights, a mathematical model for an 'artificial' neuron was developed

- Each input x_i is multiplied by weight w_i
- Add a $+1$ input neuron, which is multiplied by the bias b
- Apply a non-linear function g to the weighted combination of the inputs, $\mathbf{w}^T\mathbf{x} + b$
- Different candidates for g: heaviside function, sigmoid, tanh, rectified linear unit, etc.

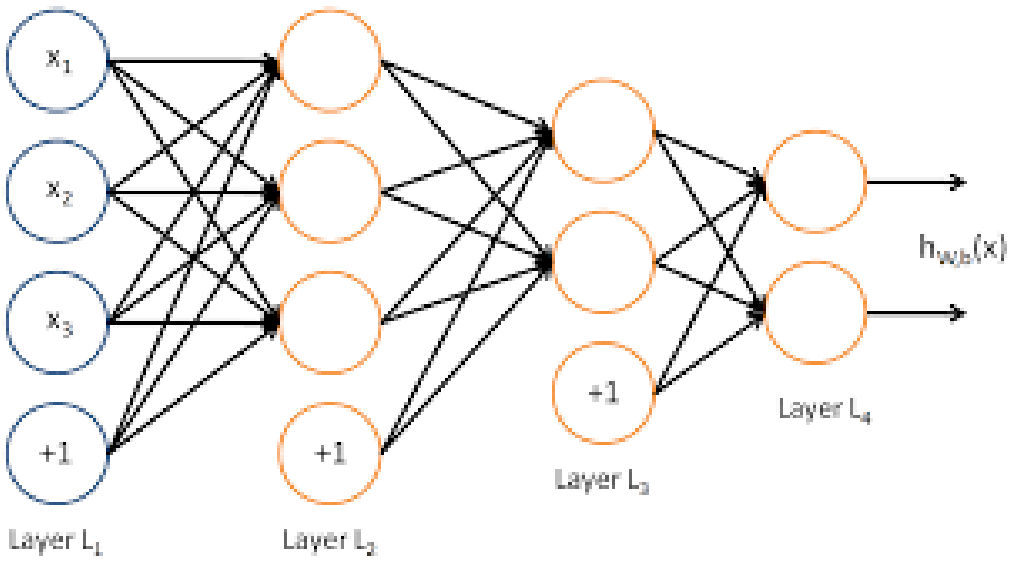


Single Artificial Neuron

Mimicking the human brain

Pass inputs through a "network" of neurons to obtain outputs.

- Neural networks are very good at handling large-scale data.
- They can learn very complex relationships.
- Requires careful configuration: what does this network look like?

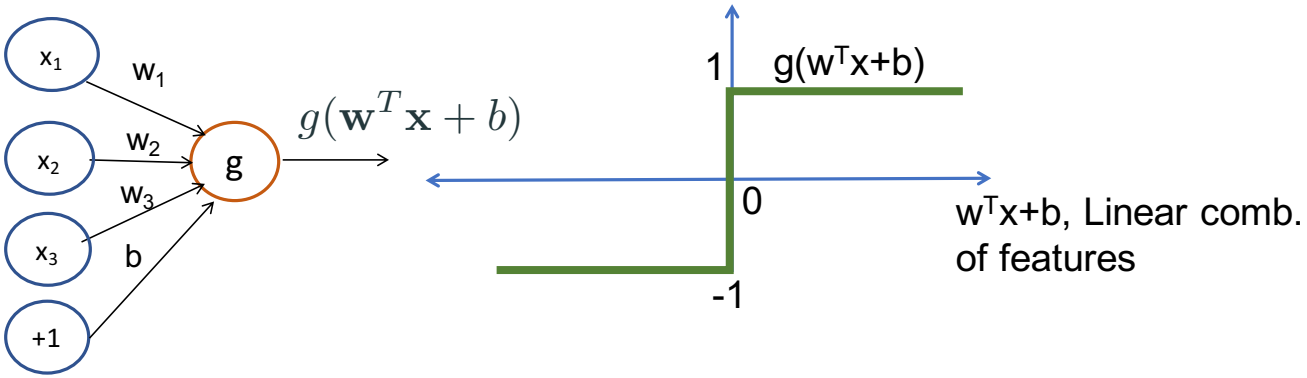


Each neuron is sometimes called a "node" or "unit" in the network. We group functions into "layers" depending on how many neurons their inputs have passed through since the original inputs.

Perceptrons

Perceptron: Rosenblatt (1957)

- The perception is a single-unit neural network with the Heaviside activation function, sign(x).
- It considers a linear binary classification problem to distinguish between two classes $\{-1, +1\}$.



Perceptron

- Assign label sign($\boldsymbol{w}^\top \boldsymbol{x} + b$) to a new sample
- Notation change: Merge b into the vector $\mathbf{w}$ and append 1 to the vector $\mathbf{x}$

How do we learn the weights?

How can we change $\boldsymbol{w}$ such that

$$y_n = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_n)$$

for a given training example $(\boldsymbol{x}_n, y_n)$?

Two cases

- If $y_n = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_n)$, do nothing.
- If $y_n \neq \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_n)$,

$$\boldsymbol{w}^{\text{NEW}} \leftarrow \boldsymbol{w}^{\text{OLD}} + y_n \boldsymbol{x}_n$$

Why would it work?

If $y_n \neq \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_n)$, then

$$y_n(\boldsymbol{w}^\top \boldsymbol{x}_n) < 0$$

What would happen if we change to new $\boldsymbol{w}^{\text{NEW}} = \boldsymbol{w} + y_n \boldsymbol{x}_n$?

$$y_n[(\boldsymbol{w} + y_n \boldsymbol{x}_n)^\top \boldsymbol{x}_n] = y_n \boldsymbol{w}^\top \boldsymbol{x}_n + y_n^2 \boldsymbol{x}_n^\top \boldsymbol{x}_n$$

We are adding a positive number, so it is possible that

$$y_n(\boldsymbol{w}^{\text{NEW}\top} \boldsymbol{x}_n) > 0$$

i.e., we are more likely to classify correctly

Perceptron training algorithm

Iteratively solving one case at a time

- REPEAT
- Pick a data point $\boldsymbol{x}_n$ (can be a fixed order of the training instances)
- Make a prediction $y = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_n)$ using the **current** $\boldsymbol{w}$.
- If $y = y_n$, do nothing.
  Else,
  $$\boldsymbol{w} \leftarrow \boldsymbol{w} + y_n \boldsymbol{x}_n.$$
- UNTIL converged.

Perceptron algorithm properties

Properties

- This is an online algorithm.
- If the training data is linearly separable, the algorithm stops in a finite number of steps.
- The parameter vector is always a linear combination of training instances (requires initialization of $\boldsymbol{w}_0 = 0$).
- No need to set a learning rate (simplifies implementation).

The perceptron algorithm was used (a long time ago) to train w by hand, without a computer.

Convergence under linear separability

- Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{R}^D$ be a sequence of T points processed until convergence,
- Assume $\|\boldsymbol{x}_t\| \leq r$ for all $t \in [1, T]$, for some $r > 0$,
- Assume that there exist $\rho > 0$ and $\boldsymbol{v} \in \mathbb{R}^D$ s.t. for all $t \in [1, T]$, $\rho \leq \frac{y_t(\boldsymbol{v} \cdot \boldsymbol{x}_t)}{\|\boldsymbol{v}\|}$.

Then, the number of updates M made by the Perceptron algorithm when processing $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ is bounded by

$$M \leq r^2/\rho^2.$$

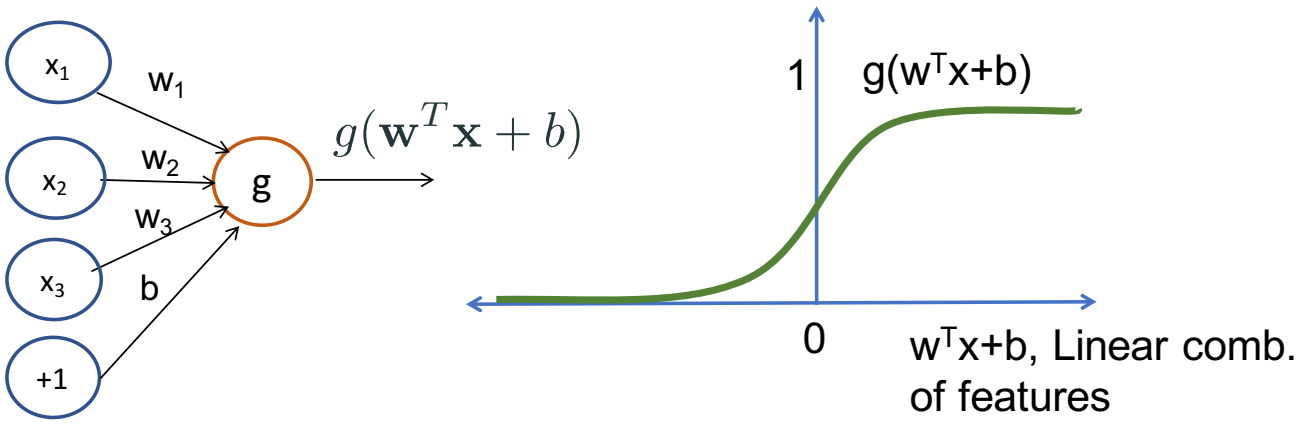
Convergence under linear separability

- Recall that $\rho \leq \frac{y_t(\boldsymbol{v} \cdot \boldsymbol{x}_t)}{\|\boldsymbol{v}\|}$, $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_t \boldsymbol{x}_t$, and $\boldsymbol{w}_0 = 0$.
- Let I be the subset of the T rounds with an update, i.e., $|I| = M$.

$$
\begin{aligned}
M\rho &\leq \frac{\boldsymbol{v} \cdot \sum_{t \in I} y_t \boldsymbol{x}_t}{\|\boldsymbol{v}\|} \leq \left\| \sum_{t \in I} y_t \boldsymbol{x}_t \right\| && \text{(Cauchy-Schwarz inequality)} \\
&= \left\| \sum_{t \in I} (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \right\| && \text{(definition of updates)} \\
&= \|\boldsymbol{w}_{T+1}\| && \text{(telescoping sum, } \boldsymbol{w}_0 = 0\text{)} \\
&= \sqrt{\sum_{t \in I} \|\boldsymbol{w}_{t+1}\|^2 - \|\boldsymbol{w}_t\|^2} && \text{(telescoping sum, } \boldsymbol{w}_0 = 0\text{)} \\
&= \sqrt{\sum_{t \in I} \|\boldsymbol{w}_t + y_t \boldsymbol{x}_t\|^2 - \|\boldsymbol{w}_t\|^2} && \text{(definition of updates)} \\
&= \sqrt{\sum_{t \in I} 2 \underbrace{y_t \boldsymbol{w}_t \cdot \boldsymbol{x}_t}_{\leq 0} + \|\boldsymbol{x}_t\|^2} \\
&\leq \sqrt{\sum_{t \in I} \|\boldsymbol{x}_t\|^2} \leq \sqrt{Mr^2} && \text{(Therefore, } M\rho \leq \sqrt{Mr^2} \rightarrow M \leq \frac{r^2}{\rho^2}\text{)}
\end{aligned}
$$

Perceptrons and Logistic Regression

- Suppose g is the sigmoid function $\sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$
- We can find a linear decision boundary separating two classes. The output is the probability of $\mathbf{x}$ belonging to class 1.



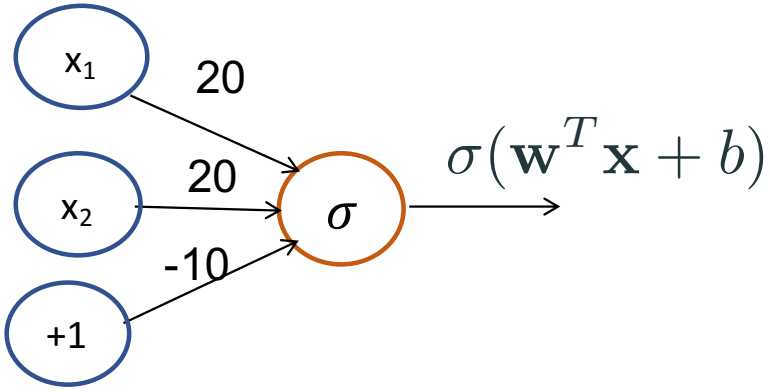
Neuron with Sigmoid activation

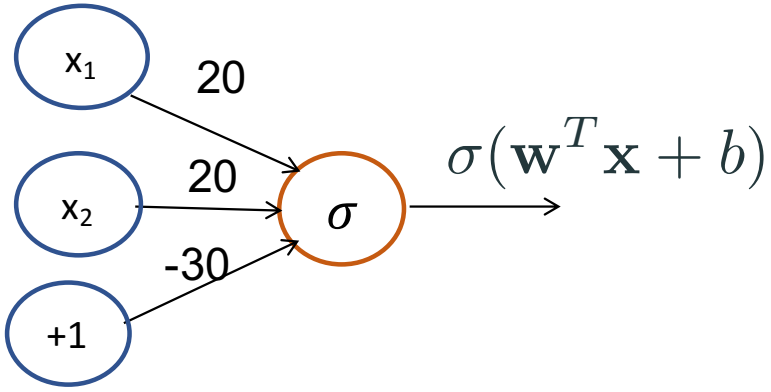
This is binary logistic regression, which we already know.

Neural Network Architectures and Forward Propagation

Example: Logic Gates

We can construct many common functions using just a single neuron





x_1	x_2	output
0	0	0
0	1	1
1	0	1
1	1	1

This is the OR gate

x_1	x_2	output
0	0	0
0	1	0
1	0	0
1	1	1

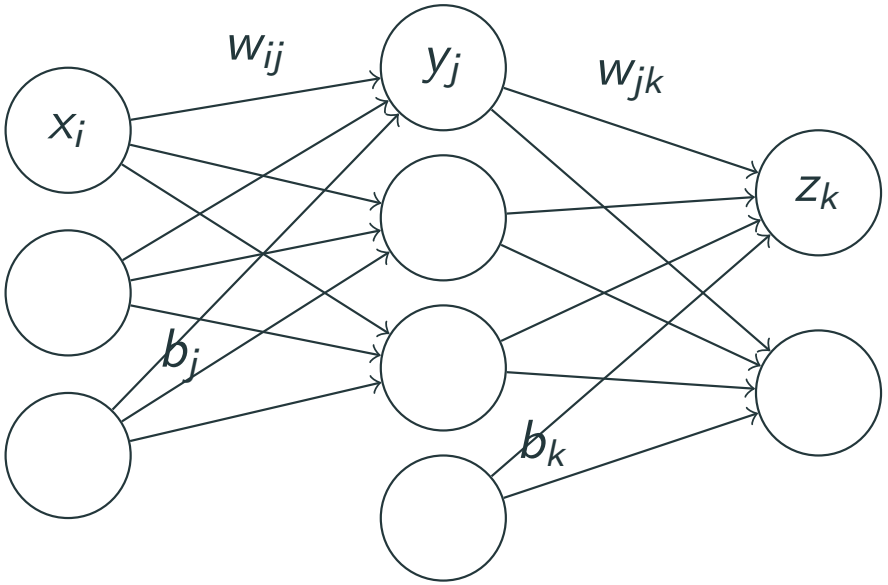
This is the AND gate

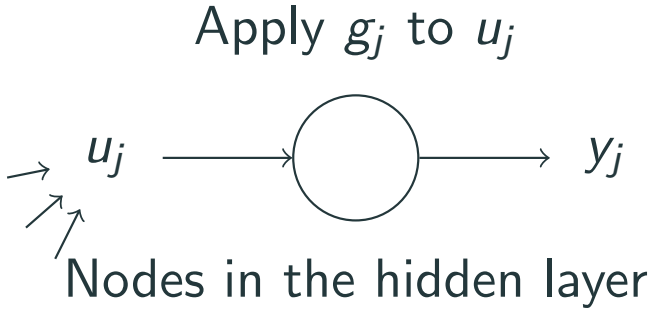
Can we build an XOR Gate?

x_1	x_2	output
0	0	0
0	1	1
1	0	1
1	1	0

- No, because this data is not linearly separable.
- We can create a combination of other logic gates $(x_1 + x_2)(\bar{x_1} + \bar{x_2})$
- Equivalent to creating a multi-layer neural network

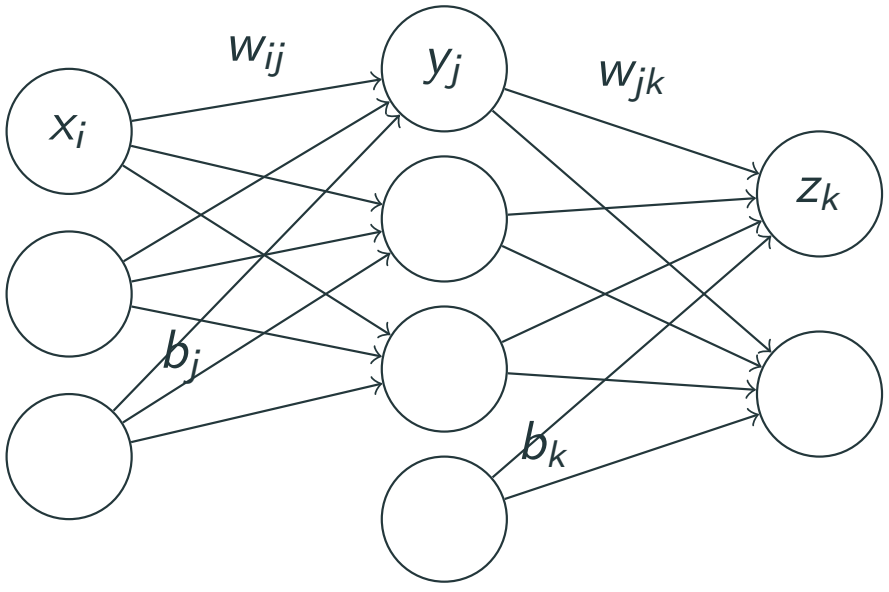
Multi-layer Neural Network

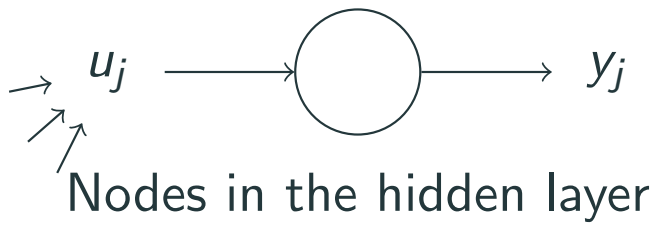




- w_{ij}: weights connecting node i in layer $(\ell - 1)$ to node j in layer ℓ.
- b_j, b_k: bias for nodes j and k.
- u_j, u_k: inputs to nodes j and k (where $u_j = b_j + \sum_i x_i w_{ij}$).
- g_j, g_k: activation function for node j (applied to u_j) and node k.
- $y_j = g_j(u_j)$, $z_k = g_k(u_k)$: output/activation of nodes j and k.
- t_k: target value for node k in the output layer.

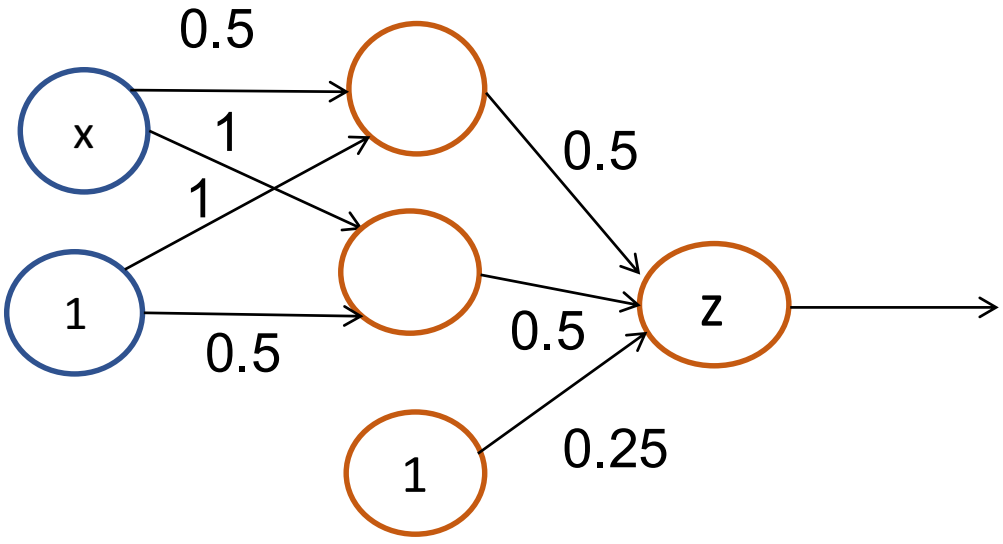
Neural Networks are very powerful





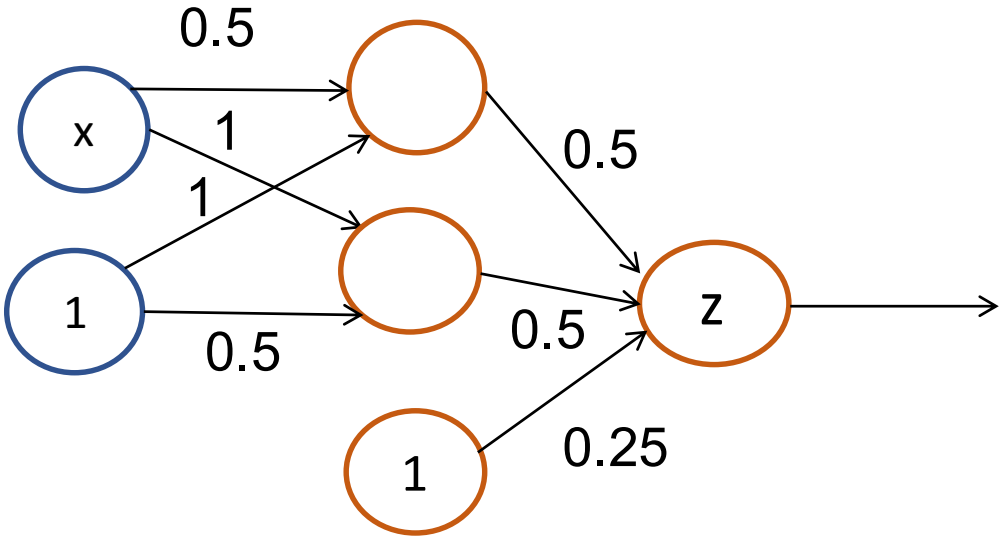
- With enough neurons and layers we can represent very complex input-output relationships
- Can be used for regression, classification, embedding, and many other ML applications

Forward-Propagation in Neural Networks



- Expressing outputs z in terms of inputs x is called forward-propagation.
- We perform forward-propagation when doing inference on a trained neural network.

Exercise: Forward-Propagation



- Outputs of the hidden layer are $\sigma(0.5x + 1)$ and $\sigma(x + 0.5)$
- Input to the last layer is $0.5\sigma(0.5x + 1) + 0.5\sigma(x + 0.5) + 0.25$
- $z = \sigma(0.5\sigma(0.5x + 1) + 0.5\sigma(x + 0.5) + 0.25)$

Choosing Activation Units

Activation choices for each layer

Input layer initially transforms the features.

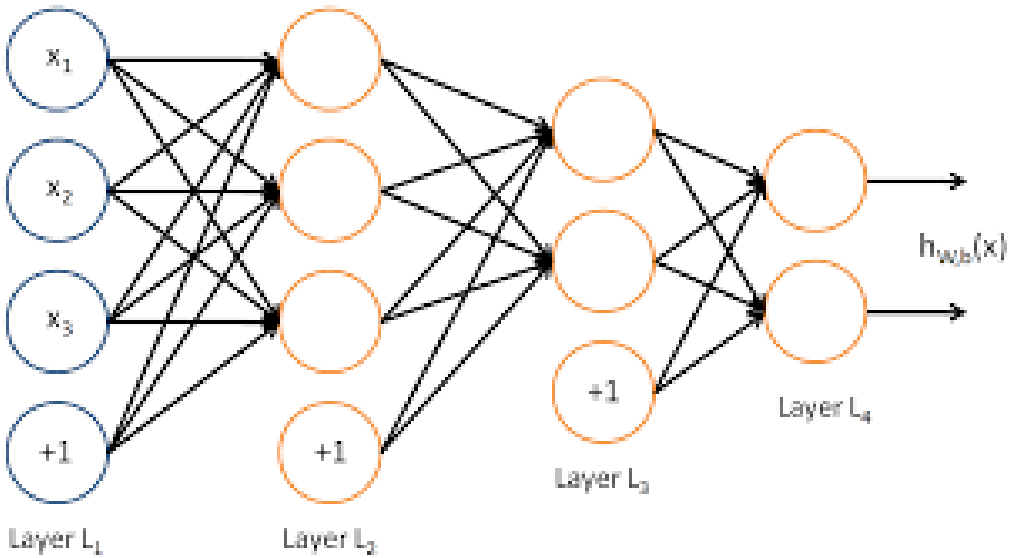
Often uses linear, sigmoid, or tanh activations.

Hidden layers convert activated inputs to classification features.

Highly problem dependent!

Output layer produces a classification decision.

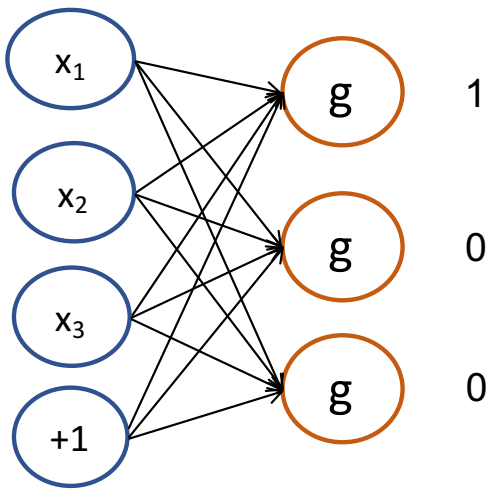
Often, these are the probabilities of the input being in each class.



Softmax for Multi-class Regression

- If the target is takes C possible values
- If y belongs to the first class, the outputs should be $[1, 0, \ldots, 0]$
- Need to produce a vector $\hat{y}$ with $\hat{y}_i = \mathbb{P}\mathrm{r}(y = i|x)$
- Linear output layer ($g(x) = x$) first produces un-normalized log probabilities:

$$z = \mathbf{w}^T \mathbf{x} + b$$

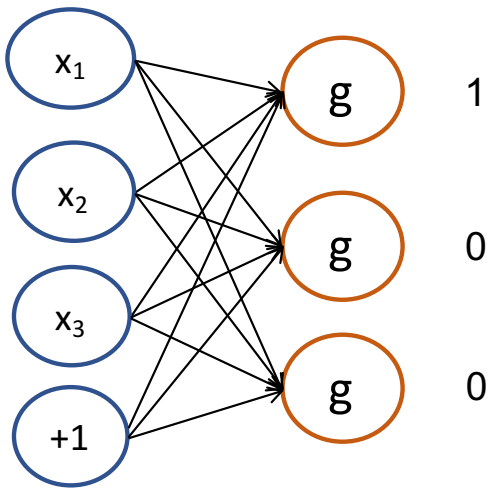


Multiclass Regression for $C = 3$

Softmax for Multi-class Regression

- Softmax:

$$\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

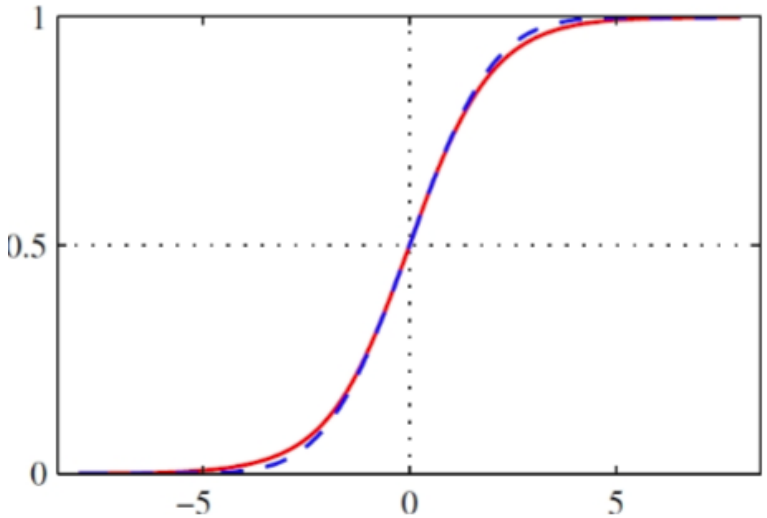


Multiclass Regression for $C = 3$

Sigmoid Units

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

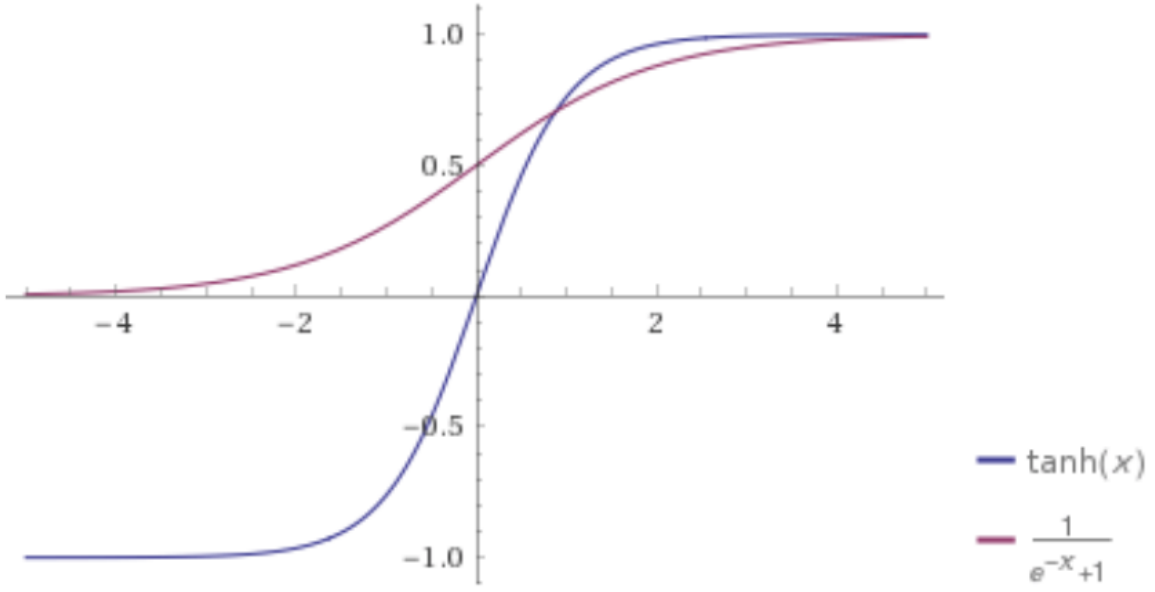
- Squashing type non-linearity: pushes output to range [0,1]



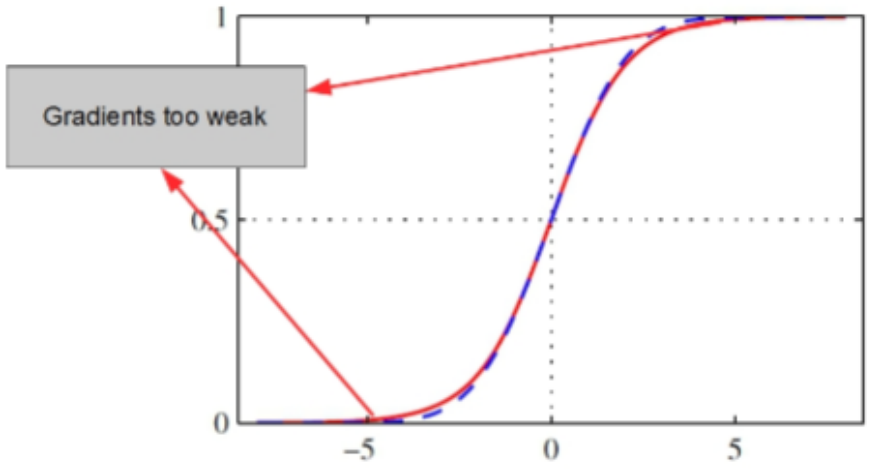
Tanh Units

$$tanh(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}}$$

- Related to sigmoid: $tanh(z) = 2\sigma(2z) - 1$
- Positive: Squashes output to range [-1,1], outputs are zero-centered
- Negative: Both tanh and sigmoid functions *saturate* at very small or large values



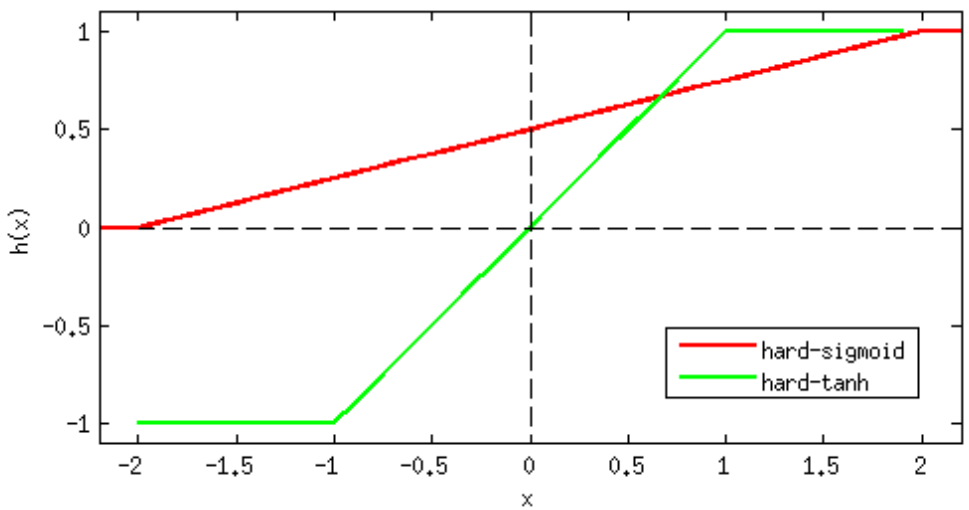
The vanishing gradients problem



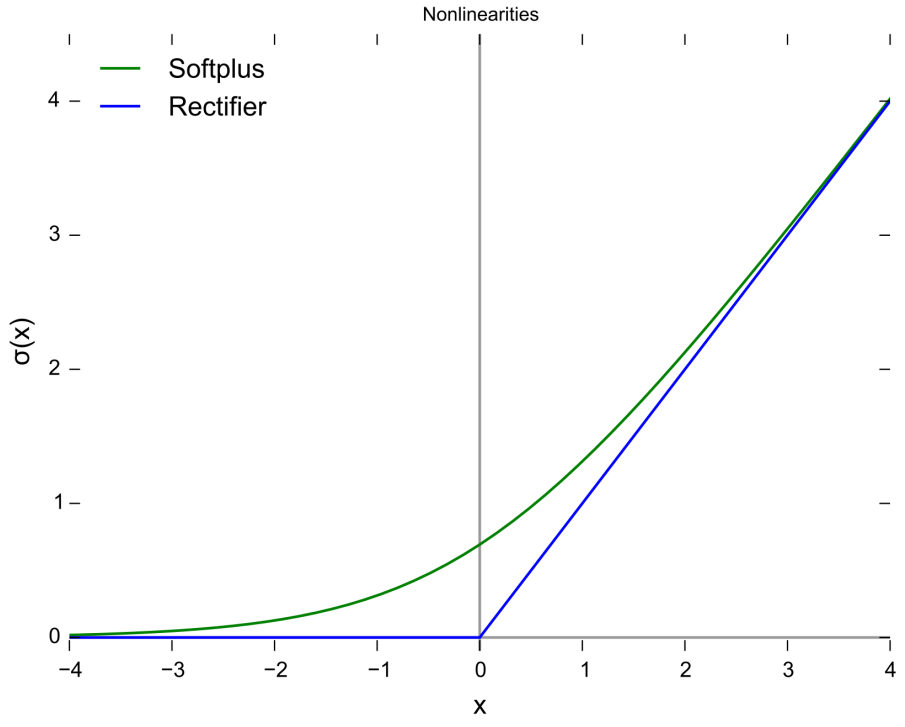
- Problem: Near-constant value across most of their domain, strongly sensitive only when z is closer to zero
- Saturation makes gradient based learning difficult (as we will see next week)

Hard Tanh and Hard Sigmoid

- To avoid the problem of vanishing gradients we can use piece-wise linear approximations to these functions
- This significantly reduces the computation complexity because gradients can take only one a few values

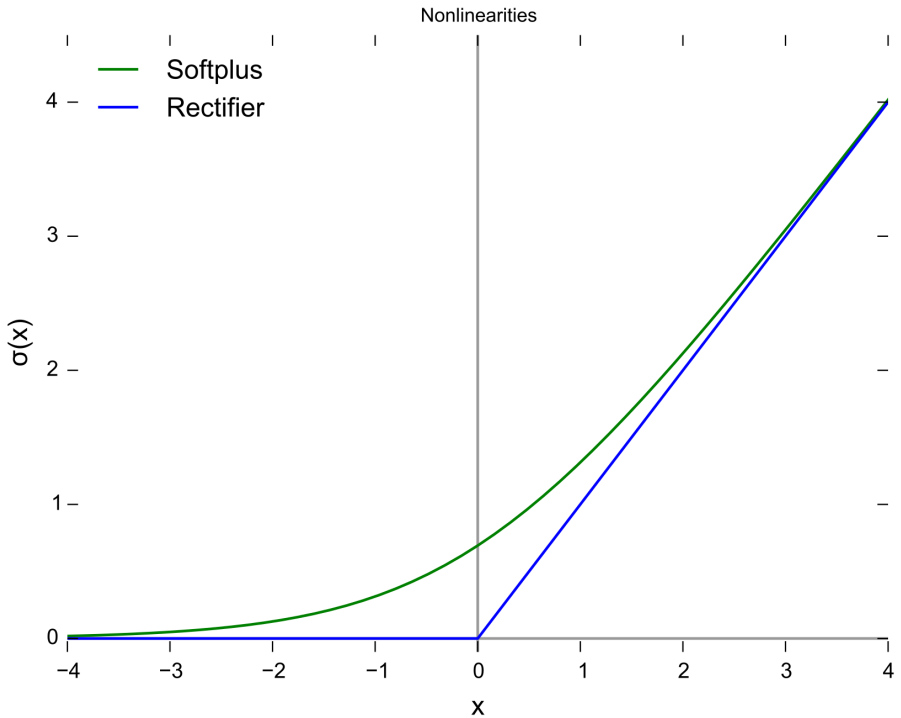


Rectified Linear Units



- Approximates the softplus function which is $\log(1 + e^z)$
- ReLu Activation function is $g(z) = max(0, z)$ with $z \in R$
- Similar to linear units. Easy to optimize!
- Give large and *consistent* gradients when active

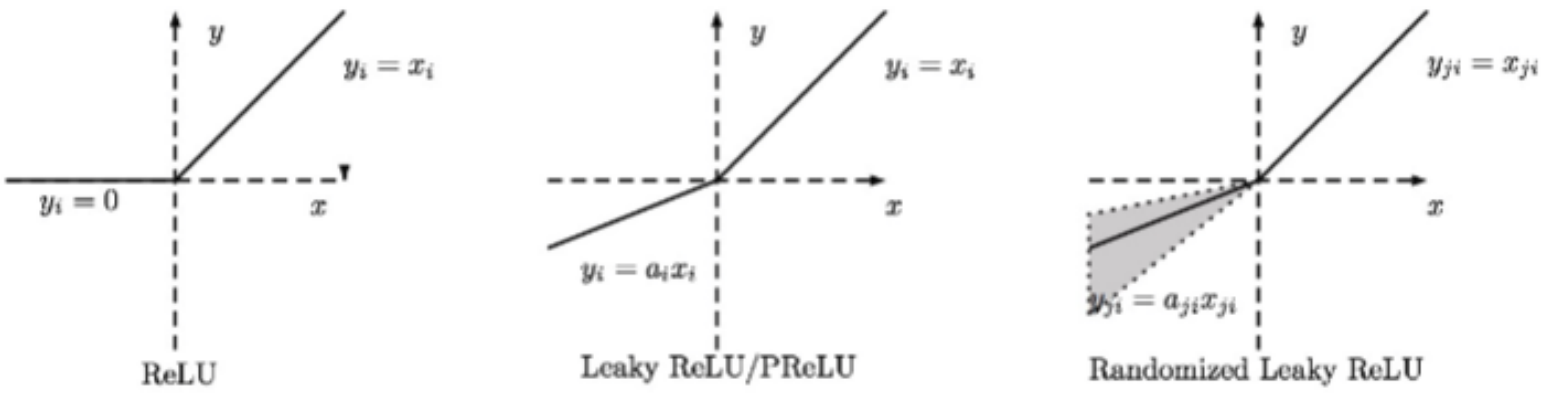
Rectified Linear Units



- Positives:
  - Gives large and *consistent* gradients (does not saturate) when active
  - Efficient to optimize, converges much faster than sigmoid or tanh
- Negatives:
  - Non zero centered output
  - Units "die" i.e when inactive they will never update

Generalized Rectified Linear Units

- Get a non-zero slope when $z_i < 0$
- $g(z, a)_i = max(0, z_i) + a_i min(0, z_i)$
  - **Leaky ReLU:**(Mass et al., 2013) Fix a_i to a small value e.g 0.01
  - **Parametric ReLU** (He et al., 2015) Learn a_i
  - **Randomized ReLU**(Xu et al., 2015) Sample a_i from a fixed range during training, fix during testing



Activation choices for each layer

Output layer produces a classification decision.

- Probabilities of the input being in each class.
- Often uses sigmoid, softmax, or tanh activations.

Hidden layers convert activated inputs to classification features.

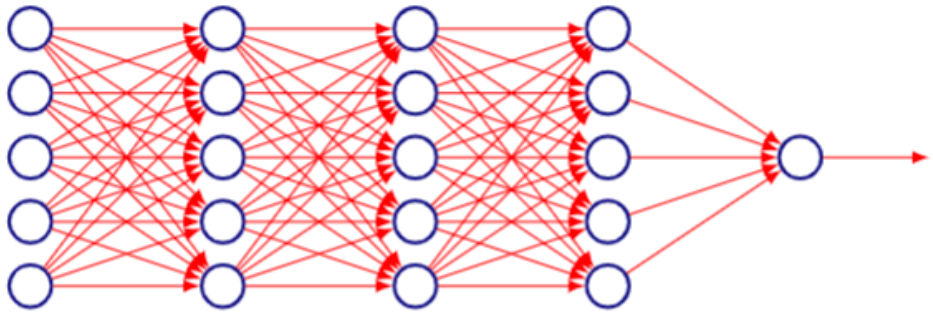
- ReLU, ELU and variants are popular choices.
- Highly problem dependent!

Input layer initially transforms the features.

Often uses linear, sigmoid, or tanh activations.

Choosing Neural Network Architectures

Architecture design



- First layer: $h^{(1)} = g^{(1)}(W^{(1)^T} x + b^{(1)})$
- Second layer: $h^{(2)} = g^{(2)}(W^{(2)^T} h^{(1)} + b^{(2)})$
- How do we decide *depth*, *width?*
- In theory how many layers *suffice?*

Universality

- Theoretical result [Cybenko, 1989]: 2-layer net with linear output with some squashing non-linearity in hidden units can approximate any continuous function over compact domain to arbitrary accuracy (given enough hidden units!)
- Implication: Regardless of function we are trying to learn, a one hidden layer neural network can represent this function.
- But not guaranteed that our training algorithm will be able to learn that function!
- Gives no guidance on how large the network will be (exponential size in worst case)

Advantages of depth

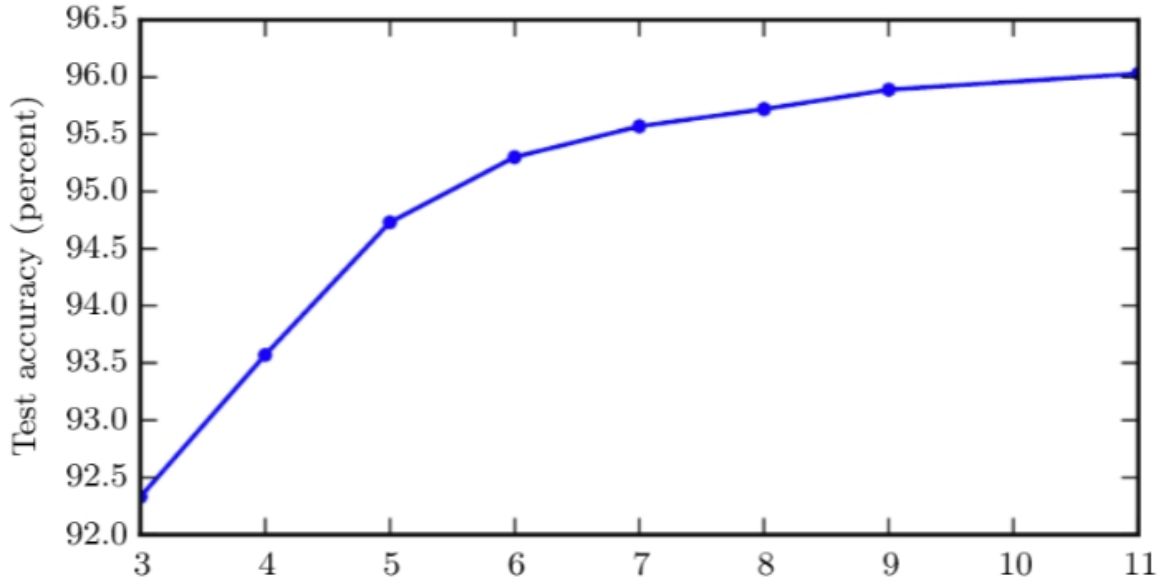
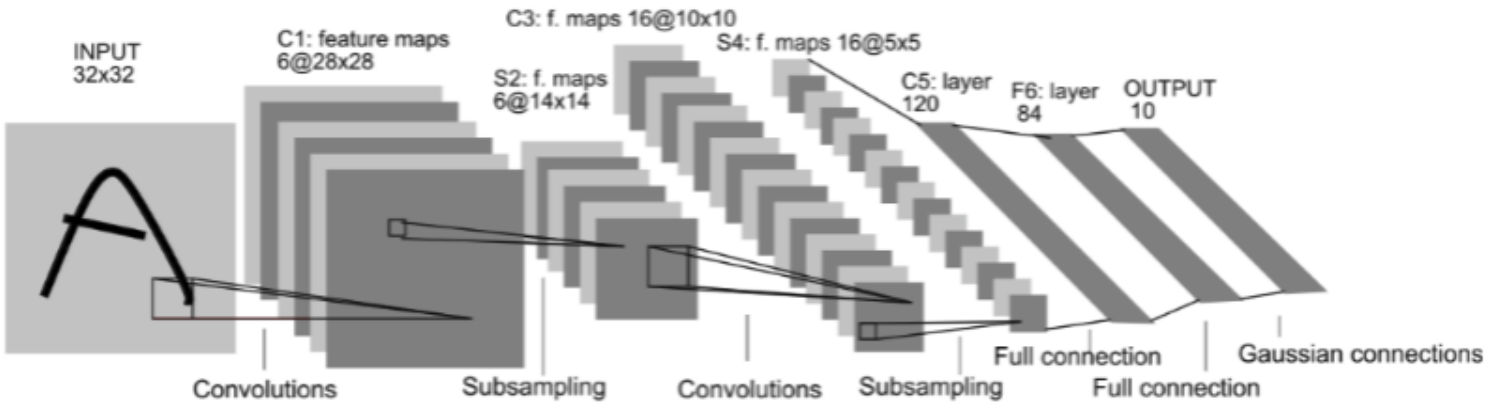


Figure 3: Goodfellow et al., 2014

- Increasing the depth of a neural network generally improves test accuracy.
- Coupled with computation advances, e.g. GPU

Deep convolutional networks

- Deep supervised neural networks are generally too difficult to train
- One notable exception: Convolutional neural networks (CNN)
- Convolutional nets were inspired by the visual system's structure.
- They typically have more than five layers, a number of layers which makes fully-connected neural networks almost impossible to train properly when initialized randomly.



Example: LeNet 5 (LeCun, 1998).

Advantages of deep convolutional networks

- Compared to standard feedforward neural networks with a similarly-sized layer
  - CNNs have much fewer connections and parameters
  - and so they are easier to train
  - while their theoretically-best performance is likely to be only slightly worse.
- Usually applied to image datasets (where convolutions have a long history).

LeNet 5

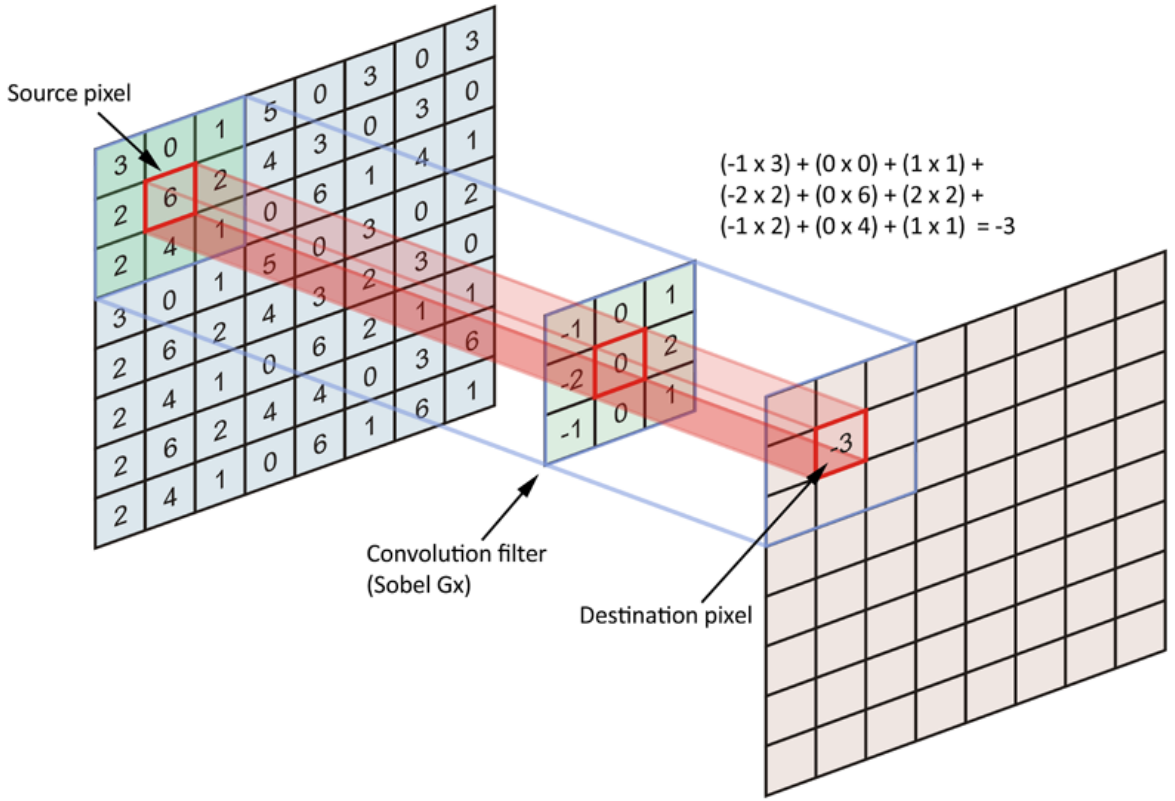
Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: **Gradient-Based Learning Applied to Document Recognition**, *Proceedings of the IEEE, 86(11):2278-2324, November **1998***

Convolutional network layers

Convolve subsets of an image with a small filter.

- Each pixel in the output image is a weighted sum of the filter and a subset of the input.
- Learn the values in the filter (these are your parameters, or weights).

Many fewer parameters (and connections) than a feedforward network.



Summary

You should know:

- Why we call these models "neural" networks.
- How to train a perceptron.
- Basic structure of a neural network.
- How to perform forward-propagation.
- Common choices for neuron activations.
- What we mean by a "deep" or "convolutional" neural network.