Machine Learning for Intelligent Systems

Lecture 13: Deep Neural Networks

Reading: UML 20-20.3

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

1

(Stochastic) Gradient Descent

Many learning problems can be written as the following optimization on the sample set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

$$\min_{\vec{w}} \mathcal{L}_S(\vec{w}) \quad for \quad \mathcal{L}_S(\vec{w}) = R(\vec{w}) + C\frac{1}{n}\sum_{i=1}^{n} L(\vec{w}\cdot\vec{x}_i, y_i)$$

Gradient Descent Update: $\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta_t \nabla \mathcal{L}_S(\vec{w}^{(t)})$

$$\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta_t \nabla R(\vec{w}) - \frac{\eta_t C}{n}\sum_{i=1}^{n} \nabla L(\vec{w}^{(t)}\cdot\vec{x}_i, y_i)$$

Stochastic Gradient Descent Update: Take a random $(\vec{x}_i, y_i) \sim S$

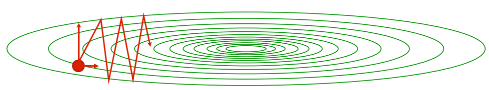
$$\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta_t \nabla R(\vec{w}) - \eta_t C\, \nabla L(\vec{w}^{(t)}\cdot\vec{x}_i, y_i)$$

2

AdaGrad

Adaptive **Grad**ient:

- Adapt the learning rate for each parameter based on the previous gradients.



For all coordinates i:

Derivative $\quad g_{i,t} = \dfrac{\partial \mathcal{L}(\vec{w}^{(t)})}{\partial w_i}$

Update $\quad w_i^{(t+1)} \leftarrow w_i^{(t)} - \eta \dfrac{g_{i,t}}{\sqrt{\underbrace{0.01}_{\text{Stabilizer}} + \underbrace{\sum_{\tau=1}^{t}(g_{i,\tau})^2}_{\text{Sum-square of derivatives}}}}$

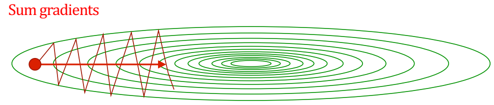
3

Momentum Method

Adaptive Gradient:

- Use previous gradients to encourage movement in important directions.

Sum gradients



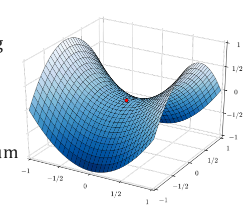
Exp-weighted Average Gradient $\quad \vec{G}^{(t)} \leftarrow (1-\beta)\,\vec{G}^{(t-1)} + \beta \nabla \mathcal{L}(\vec{w}^{(t)})$

Update $\quad \vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta \vec{G}^{(t)}$

4

SGD on Non-Convex

Non convex functions are challenging



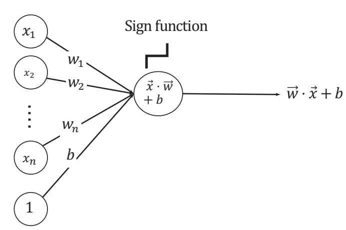
Under specific assumptions SGD provably converges to a local minimum

Neural networks are non-convex and SGD is used for training them.

5

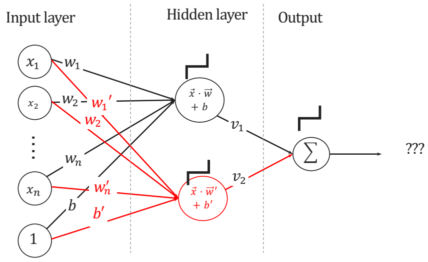
Linear Model

We can represent a linear function as single layer neural network.



6

Naïve Hidden Layers



Linear function of linear functions, is linear.

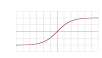
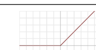
$$v_1(\vec{x}\cdot\vec{w} + b) + v_2(\vec{x}\cdot\vec{w}' + b') = \vec{x}\cdot(v_1\vec{w} + v_2\vec{w}') + (v_1 b + v_2 b')$$

Beyond linearity: We need each layer to transform a linear function to something else.

7

Common Activation Functions

Use a non-linear activation function on nodes of a hidden layer.

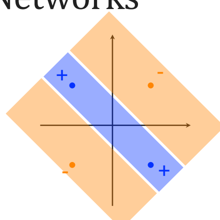
Name	Function	Gradient	Graph
Binary step	$\text{sign}(x)$	$\begin{cases} 0 & x \neq 0 \\ N/A & x = 0 \end{cases}$	
sigmoid	$\sigma(x) = \dfrac{1}{1+\exp(-x)}$	$\sigma(x)(1-\sigma(x))$	
Tanh	$tanh(x) = \dfrac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$	$(1-tanh(x))^2$	
Rectified Linear (ReLu)	$relu(x) = \max(x, 0)$	$\begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$	

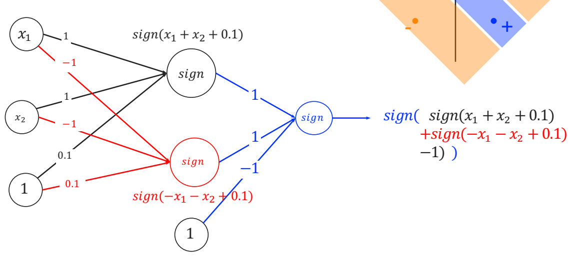
Sometime, $\sigma(x)$ denotes the "generic" notion of activation function, not necessarily sigmoid.

8

Power of Neural Networks

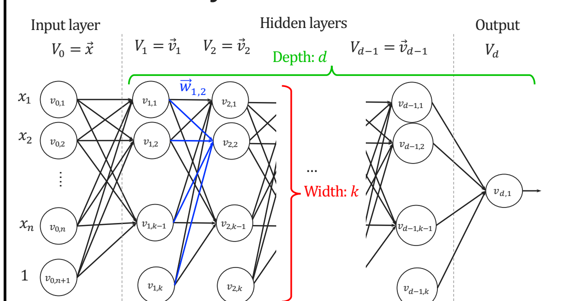
Represent XOR with 1 hidden layer.





9

Multi Layer Neural Network



Vector of weights going from layer i to the j^{th} node of layer $i+1$: $\vec{w}_{i,j}$

Vector of values in layer $i+1$, $\vec{v}_{i+1} = \begin{pmatrix} \sigma(\vec{v}_i \cdot \vec{w}_{i,1}) \\ \vdots \\ \sigma(\vec{v}_i \cdot \vec{w}_{i,k}) \end{pmatrix}$ Output: $\sigma(v_d)$

10

Universal Approximators

If we allow a single hidden layer (depth 2 network) with very large width, we can approximate any continuous function on $\mathbb{R}^n$.

How large?

- For boolean functions, we need at least $\exp(n)$ width.
- Restricting ourselves to polynomial size networks
- → Can't approximate all functions
- → Reduce the chance of overfitting

$\left.\begin{array}{l}\text{Can't approximate all functions}\\ \text{Reduce the chance of overfitting}\end{array}\right\}$ Bias-Variance Tradeoff

11

Other Types of Neural Networks

- Traditional multi-layer networks:

→Layers are fully connected

→Bad for overfitting

Other types

- Convolutional Neural Networks (CNNs)

→Some structured layers to learn features

→1-2 layers of fully connected network at the end

- Recurrent Neural Networks (RNNs)

→ Nodes can feed forward or backward.

12