# Neural Networks

IKHAN CHOI

The AI paradigm changes when a new approximating method is discovered.

## 1. Bayesian Networks

**Definition 1.1** (Bayesian network)**.** Let $G$ be a directed acyclic graph.

The graph acts like a parameter space. We want to investigate mutual effects among the paramtrized random variables.

**Theorem 1.1** (Factorization of probablity)**.**

## 2. Neural Networks

### 2.1. Gradient descent method.

2.2. **Back propagation.** Backpropagation refers to algorithms to train the weight matrices for minimizing the cost function $J$, which does not depend explicitly on any variables except the last layer vector $a^{(n)}$. However, since $J$ is a function of the weight matrices implicitly, via $a^{(n)}$, we may find the representation of the gradiant of $J$ as viewing it as a function on the space of weight matrices of each given layer. In other words, we want to find the coefficients of the differential form $dJ$ on the basis $\{dW_{ij}^{(n-1)}\}_{i,j}$, $\{dW_{jk}^{(n-2)}\}_{j,k}$, or $\{dW_{kl}^{(n-3)}\}_{k,l}$, and so on.

Recall the definitions:

$$a_i^{(n)} = \sigma\left(\sum_j W_{ij}^{(n-1)} a_j^{(n-1)}\right).$$

Since the derivative of the sigmoid function is given by $\sigma' = \sigma - \sigma^2$, we can compute the following auxiliary relations

$$\frac{\partial a_i^{(n)}}{\partial a_j^{(n-1)}} = h(a_i^{(n)}) W_{ij}^{(n-1)} \quad \text{and} \quad \frac{\partial a_i^{(n)}}{\partial W_{i'j}^{(n-1)}} = \delta_{ii'} h(a_i^{(n)}) a_j^{(n-1)},$$

where $h(x) = x - x^2$.

Then, we can compute

$$dJ = \sum_i \frac{\partial J}{\partial a_i^{(n)}} \sum_j \frac{\partial a_i^{(n)}}{\partial W_{ij}^{(n-1)}} \, dW_{ij}^{(n-1)} = \sum_{i,j} \frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)}) a_j^{(n-1)} \, dW_{ij}^{(n-1)},$$

*Last Update*: July 15, 2019.

which implies

$$\nabla J(W^{(n-1)}) = \left[\frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)}) a_j^{(n-1)}\right] \frac{\partial}{\partial W_{ij}^{(n-1)}}.$$

Note that it is a function of $a_i$ and $a_j$. The gradient descent method will take

$$W_{ij}^{(n-1)^+} := W_{ij}^{(n-1)} - \alpha \cdot \frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)}) a_j^{(n-1)}$$

with a proper parameter $\alpha > 0$.

By the same reason,

$$\begin{aligned} dJ &= \sum_{i,j,k} \frac{\partial J}{\partial a_i^{(n)}} \frac{\partial a_i^{(n)}}{\partial a_j^{(n-1)}} \frac{\partial a_j^{(n-1)}}{\partial W_{jk}^{(n-2)}} \, dW_{jk}^{(n-2)} \\ &= \sum_{i,j,k} \frac{\partial J}{\partial a_i^{(n)}} \cdot h(a_i^{(n)}) W_{ij}^{(n-1)} \cdot h(a_j^{(n-1)}) a_k^{(n-2)} \, dW_{jk}^{(n-2)}, \end{aligned}$$

which implies

$$\nabla J(W^{(n-2)}) = \left[\sum_i \frac{\partial J}{\partial a_i^{(n)}} \cdot h(a_i^{(n)}) W_{ij}^{(n-1)} \cdot h(a_j^{(n-1)}) a_k^{(n-2)}\right] \frac{\partial}{\partial W_{jk}^{(n-2)}}.$$

Therefore, the gradient descent method will take

$$\begin{aligned} W_{jk}^{(n-2)^+} &: = W_{jk}^{(n-2)} - \alpha \cdot \sum_i \frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)}) W_{ij}^{(n-1)} h(a_j^{(n-1)}) a_k^{(n-2)} \\ &= W_{jk}^{(n-2)} + (1 - a_j^{(n-1)}) a_k^{(n-2)} \sum_i (W_{ij}^{(n-1)^+} - W_{ij}^{(n-1)}) W_{ij}^{(n-1)}. \end{aligned}$$

In similar way,

$$W_{kl}^{(n-3)^+} := W_{kl}^{(n-3)} + (1 - a_k^{(n-2)}) a_l^{(n-3)} \sum_i (W_{jk}^{(n-2)^+} - W_{jk}^{(n-2)}) W_{jk}^{(n-2)} (?)$$

## 3. Maximum likelihood estimate

**Definition 3.1.** Let $f$ be a distribution function on a measure space $X$. Let $\{f_\theta\}_\theta$ be a parametrized family of distrubution functions on $X$. The *likelihood* $L_n(\theta) : \Omega^n \to \mathbb{R}_{\geq 0}$ for a fixed parameter $\theta$ is a random variable defined by

$$L_n(\theta) := \prod_{i=1}^n f_\theta(x_i)$$

where $\{x_i\}_i$ is a family of i.i.d. $X$-valued random variables with a distriburion $f$.

The objective of the likelihood function is to find $\theta$ such that $f_\theta$ approximates the unknown distribution $f$. Write

$$\frac{1}{n} \log L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f_\theta(x_i).$$

By the law of large numbers, $\frac{1}{n}\log L_n(\theta)$ converges to a constant function

$$\mathbb{E}(\log f_\theta(x)) = \int_X f \log f_\theta$$

in measure as $n \to \infty$. This constant function is exactly what we call *cross entropy.*

The *Kullback-Leibler divergence* is a kind of asymmetric distance function defined from the difference with cross entropy

$$D_{KL}(f \| f_\theta) := \int_X f \log f - \int_X f \log f_\theta.$$

It is proved to be always nonnegative by the Jensen inequality:

$$\int_X f \log f_\theta - \int_X f \log f = \int_X f \log \frac{f_\theta}{f} \le \log\left(\int_X f \frac{f_\theta}{f}\right) = 0.$$

Here, we exclude the region $f = 0$ from the integration region. Then, we can say, bigger $L_n(\theta)$ is, closer $f_\theta$ and $f$ are.