

Neural Networks

IKHAN CHOI

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

1. UPDATE OF PARAMETERS

1.1. **Gradient descent method.** ascending stochastic gradient

1.2. **Back propagation.** Consider

$$b = \sigma(wa), \quad c = \sigma(Wb), \quad J = J(c).$$

We have

$$\frac{\partial c^k}{\partial b^j} = [c^k(1 - c^k)] \cdot W_j^k \quad \text{and} \quad \frac{\partial c^k}{\partial W_j^k} = [c^k(1 - c^k)] \cdot b^j.$$

Let J be the objectivity function. Then,

$$dJ = \frac{\partial J}{\partial c^k} \frac{\partial c^k}{\partial W_j^k} dW_j^k = \frac{\partial J}{\partial c^k} [c^k(1 - c^k)] b^j dW_j^k,$$

and

$$dJ = \frac{\partial J}{\partial c^k} \frac{\partial c^k}{\partial b^j} \frac{\partial b^j}{\partial w_i^j} dw_i^j = \frac{\partial J}{\partial c^k} \cdot [c^k(1 - c^k)] W_j^k \cdot [b^j(1 - b^j)] a^i dw_i^j.$$

2. MAXIMUM LIKELIHOOD ESTIMATE

Definition 2.1. Let f be a distribution function on a measure space X . Let $\{f_\theta\}_\theta$ be a parametrized family of distribution functions on X . The *likelihood* $L_n(\theta) : \Omega^n \rightarrow \mathbb{R}_{\geq 0}$ for a fixed parameter θ 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 distribution f .

The objective of the likelihood function is to find θ such that f_θ 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 $n \rightarrow \infty$.

By the Jensen inequality,

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

Exclude the region $f = 0$ from the integration region. In other words, bigger $L_n(\theta)$ is, closer f_θ and f are.

3. GENERATIVE ADVERSARIAL NETWORKS

Let X be the set of all images having a given pixel size. Suppose the data distribution p_{data} on X which embodies learning materials is given. If $x \in X$ is an image that looks like a real human face, then the distribution(mass) function p_{data} has nonnegligible values near the point x . We cannot describe the distribution function p_{data} completely, but only can sample from it.

Let p_g be a distribution on X . The generator $G : \Omega \rightarrow X$ is just an arbitrarily taken random variable satisfying p_g for sampling. The discriminator $D : X \rightarrow [0, 1]$ is a function. Our purpose is to construct a new method for approximating $p_g \rightarrow p_{data}$ by simultaneously updating the discriminator function D .

Let $x_i \sim p_{data}$ and $z \sim p_g$ be random variables $\Omega \rightarrow X$. Let D maximize

$$\log D(x) + \log(1 - D(z))$$

and p_g minimize

$$\log(1 - D(z)).$$

Balancing the convergence rates between p_g and D is important.