Information Maximization Approach

薛盛可
xueshengke@zju.edu.cn

College of Information Science and Electronic Engineering

Zhejiang University

Spring, 2016

Objective

In terms of training neural networks, the most common way we take is *Gradient Descent Method*. It has been widely used in researches indeed.

An original method Information Maximization Approach [1] that is hypothesized to training a neural network.

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

Outline

Inference

Illustration

Information Maximization

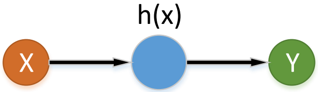


Figure 1: Input X, output Y.

To maximize the mutual information between output Y and input X, which is defined as

$$I(Y,X) = H(Y) - H(Y \mid X), \tag{1}$$

$$\frac{\partial}{\partial w} I(Y,X) = \frac{\partial}{\partial w} H(Y). \tag{2}$$

Maximize the mutual information is equivalent to maximize the entropy of output alone.

Example

We interpret the idea through an example as shown in Fig. 2,

- $f_x(x)$ is the probability density function of input x.
- $f_y(y)$ is the probability density function of output y.
- $h(x) = 1/(1 + e^{-(ax+b)})$ is an adjustable activation function.

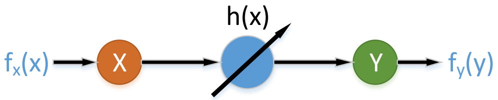


Figure 2: One input and one output.

Derivation

After a certain process of deduce,

$$f_y(y) = \frac{f_x(x)}{|\partial y/\partial x|}, \; H(y) = E[\ln|\partial y/\partial x|] - E[\ln f_x(x)], \tag{3}$$

$$\Delta w \propto \frac{\partial H}{\partial w} = \frac{\partial}{\partial w}\left(\ln\left|\frac{\partial y}{\partial x}\right|\right) = \left(\frac{\partial y}{\partial x}\right)^{-1}\frac{\partial}{\partial w}\left(\frac{\partial y}{\partial x}\right). \tag{4}$$

We conclude the learning rules of information maximization approach,

$$\Delta a \propto \frac{\partial H}{\partial a} = \frac{1}{a} + x(1-2y). \tag{5}$$

$$\Delta b \propto \frac{\partial H}{\partial b} = 1 - 2y. \tag{6}$$

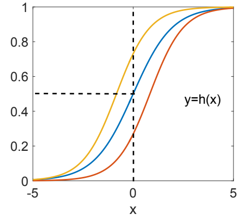
The effect of two rules will be demonstrated intuitively.

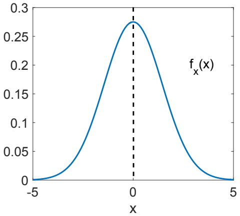
Outline

Inference

Illustration

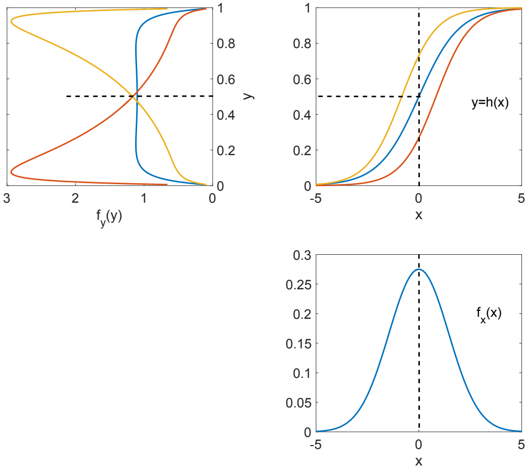
Information Maximization





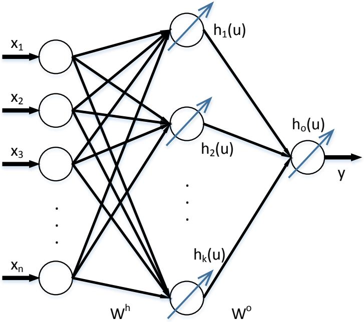
- Δa rule scales the slope of sigmoid curve, to match the variance of $f_x(x)$.
- Δb rule shifts the sigmoid curve horizontally, to align the steepest part of sigmoid curve to the peak of $f_x(x)$.

Conclusion



Eventually, the effect produces an output $f_y(y)$ that is close to a flat unit distribution, i.e., the maximum entropy distribution without assuming any prior knowledge of the input distribution $f_x(x)$.

Training Algorithm



New method to train a neural network that is composed of adjustable neurons:
combine gradient descent method with information maximization approach.

End

Thanks for your listening.