Technische Universität München
Institute for Cognitive Systems
Prof. Dr.-Ing. Gordon Cheng

Backpropagation for MDNs

Biologically inspired Humanoid Robotics

Authors: Shengzhi Wang, Fan Wu, Helene Obert, Adon Yazigi
Group Number: Group A

Contents

Chapter 1

Backpropagation for MDNs

Before we calculate the backpropagation, we want to pre-define something in the next section.

1.1 Pre-Definition

Figure 1.1 shows the mdn we designed. We define that the output from input layer as a_{input1} and a_{input2}. The output of hidden layer is a_{h1}, a_{h2} and a_{h3}. And the output of output layer is a_{o1}, a_{o2}, a_{o3}, a_{o4} and a_{o5}. The weight matrix between input layer and hidden layer is $w_{x_i h_j}$, where x_i means the i-th neuron of input layer and h_j indicates the j-th neuron of hidden layer. And similar, the weight matrix between hidden layer and output layer is defined as $w_{h_i o_j}$, where h_i indicates the i-th neuron of hidden layer and o_j represents the j-th neuron of output layer. We define the weighted input in hidden layer as z_{hi} for the i-th neuron, and the weighted input in output layer as $z_{oi}$for the i-th neuron. So there is a relation between the output from each layer with the weighted input:

$$a_{hi} = f(z_{hi}) \tag{1.1}$$

$$a_{oi} = f(z_{oi}) \tag{1.2}$$

Viewing the mixture parameters, π is equal to the softmax of output of output layer. So the calculation of π is expressed as following:

$$\pi_k = \frac{exp(a_k^{\pi})}{\Sigma_{l=1}^{K} exp(a_l^{\pi})} \tag{1.3}$$

However, in our case, the number of mixture (i.e. the gaussian kernel) is one. So the π is always equal to one, which means:

$$\pi = 1 \tag{1.4}$$

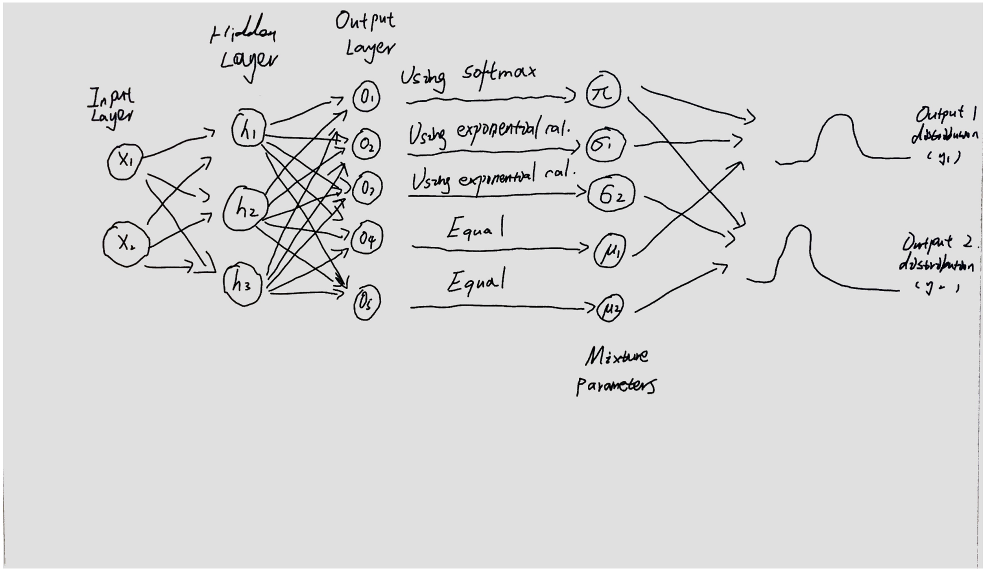

Figure 1.1: Mdn structure

And sigma σ_1 and σ_2 can be calculated as following:

$$\sigma_1 = exp(a_{o2}) \tag{1.5}$$

$$\sigma_2 = exp(a_{o3}) \tag{1.6}$$

The μ is obviously:

$$\mu_1 = a_{o4} \tag{1.7}$$

$$\mu_2 = a_{o5} \tag{1.8}$$

According to [1], we want to maximize the likelihood of gaussian distribution by minimizing the error function. The error function in our case (we have two dimensional output) can be therefore calculated as following:

$$E(\mathbf{w}) = E_1(\mathbf{w}) + E_2(\mathbf{w}) = -\sum_{n=1}^{2} ln\{\pi_n(\mathbf{x}, \mathbf{w})\mathcal{N}(\mathbf{t}|\mu_{\mathbf{n}}(\mathbf{x}, \mathbf{w}), \sigma_n^2(\mathbf{x}, \mathbf{w}))\} \tag{1.9}$$

where $\mathbf{x}$ is the input, $\mathbf{w}$ is the weight matrix, $\mathbf{t}$ is the target (i.e. the real output) and $\mathcal{N}$ is the gaussian density function. The gaussian distribution can be calculated as following:

$$\mathcal{N}(\mathbf{t}|\mu_{\mathbf{n}}(\mathbf{x}, \mathbf{w}), \sigma_n^2(\mathbf{x}, \mathbf{w})) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_n(\mathbf{x}, \mathbf{w})} exp\{-\frac{\|\mathbf{t} - \mu_{\mathbf{n}}(\mathbf{x}, \mathbf{w})\|^2}{2\sigma_n^2(\mathbf{x}, \mathbf{w})}\} \tag{1.10}$$

1.2 Backpropagation Calculation

We start by using the chain rule for E_1 and E_2 w.r.t. the output a_{o1}. Following are the expressions:

$$\frac{\partial E_1}{\partial a_{o1}} = \frac{\partial E_1}{\partial \pi}\frac{\partial \pi}{\partial a_{o1}} \tag{1.11}$$

$$\frac{\partial E_2}{\partial a_{o1}} = \frac{\partial E_2}{\partial \pi}\frac{\partial \pi}{\partial a_{o1}} \tag{1.12}$$

According to (1.4), we know $\frac{\partial \pi}{\partial a_{o1}} = 0$. So the (1.11) and (1.12) are always equal to 0:

$$\frac{\partial E_1}{\partial a_{o1}} = 0 \tag{1.13}$$

$$\frac{\partial E_2}{\partial a_{o1}} = 0 \tag{1.14}$$

So in the future calculation, the error back propagation calculation involving the π is always equal to zero and should not need to be considered anymore.

Then we consider the derivative of error function w.r.t. to μ. We use the same chain rule for E_1 and E_2 w.r.t. the output a_{o4} and a_{o5} respectively. Because of (1.7) and (1.8), we can calculate the derivative of error function w.r.t. to μ_1 and μ_2 according to (1.10)

$$\begin{aligned}
\frac{\partial E_1}{\partial a_{o4}} &= \frac{\partial E_1}{\partial \mu_1}\frac{\partial \mu_1}{\partial a_{o4}} \\
&= \frac{\partial E_1}{\partial \mu_1} \\
&= -\frac{\partial ln\{\pi(\mathbf{x},\mathbf{w})\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))\}}{\partial \mu_1} \\
&= -\frac{\partial ln\{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))\}}{\partial \mu_1} \\
&= -\frac{1}{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w})}\frac{\partial\{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))\}}{\partial \mu_1} \\
&= -\frac{1}{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w})} \cdot \mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w})) \cdot \frac{\mathbf{t}-\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w})}{\sigma_1^2(\mathbf{x},\mathbf{w})} \\
&= \frac{\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w})-\mathbf{t}}{\sigma_1^2(\mathbf{x},\mathbf{w})}
\end{aligned} \tag{1.15}$$

And therefore, the derivative of E_2 on a_{o5} is:

$$\frac{\partial E_2}{\partial a_{o5}} = \frac{\partial E_2}{\partial \mu_2}\frac{\partial \mu_2}{\partial a_{o5}} = \frac{\boldsymbol{\mu_2}(\mathbf{x},\mathbf{w})-\mathbf{t}}{\sigma_2^2(\mathbf{x},\mathbf{w})} \tag{1.16}$$

Similarly, we want to calculate the derivatives of E_1 and E_2 w.r.t. σ. According to (1.5) and (1.6), we have:

$$\begin{aligned}
\frac{\partial E_1}{\partial a_{o2}} &= \frac{\partial E_1}{\partial \sigma_1}\frac{\partial \sigma_1}{\partial a_{o2}} \\
&= \frac{\partial E_1}{\partial \sigma_1}\sigma_1 \\
&= -\frac{\partial ln\{\pi(\mathbf{x},\mathbf{w})\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))\}}{\partial \sigma_1}\sigma_1 \\
&= -\frac{\partial ln\{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))\}}{\partial \sigma_1}\sigma_1 \\
&= -\frac{1}{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))}\cdot \\
&\quad \left(-\frac{\mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))}{\sigma_1} + \mathcal{N}(\mathbf{t}|\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w}),\sigma_1^2(\mathbf{x},\mathbf{w}))\frac{\|\mathbf{t}-\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w})\|^2}{\sigma_1^3}\right)\sigma_1 \\
&= 1 - \frac{\|\mathbf{t}-\boldsymbol{\mu_1}(\mathbf{x},\mathbf{w})\|^2}{\sigma_1^2}
\end{aligned} \tag{1.17}$$

So the derivative of E_2 on σ_2 is therefore expressed as:

$$\frac{\partial E_2}{\partial a_{o3}} = \frac{\partial E_2}{\partial \sigma_2}\frac{\partial \sigma_2}{\partial a_{o3}} = 1 - \frac{\|\mathbf{t}-\boldsymbol{\mu_2}(\mathbf{x},\mathbf{w})\|^2}{\sigma_2^2} \tag{1.18}$$

Until now, the most important part of backpropagation is already done. From now on we just want to briefly show how the derivatives of E_1 and E_2 go through the previous network. The derivative of E_1 and E_2 on the weighted input of output layer is therefore:

$$\frac{\partial E_1}{\partial z_{o2}} = \frac{\partial E_1}{\partial a_{o2}}\frac{\partial a_{o2}}{\partial z_{o2}} \tag{1.19}$$

$$\frac{\partial E_1}{\partial z_{o4}} = \frac{\partial E_1}{\partial a_{o4}}\frac{\partial a_{o4}}{\partial z_{o4}} \tag{1.20}$$

$$\frac{\partial E_2}{\partial z_{o3}} = \frac{\partial E_2}{\partial a_{o3}}\frac{\partial a_{o3}}{\partial z_{o3}} \tag{1.21}$$

$$\frac{\partial E_2}{\partial z_{o5}} = \frac{\partial E_2}{\partial a_{o5}}\frac{\partial a_{o5}}{\partial z_{o5}} \tag{1.22}$$

where the partial derivatives of a on z are depended on the activation function of output layer. Then the derivatives of E_1 and E_2 on the output values of hidden layer are therefore calculated as following:

$$\frac{\partial E_1}{\partial a_{h1}} = w_{h_1o_2}\frac{\partial E_1}{\partial z_{o2}} + w_{h_1o_3}\frac{\partial E_1}{\partial z_{o3}} \tag{1.23}$$

$$\frac{\partial E_2}{\partial a_{h1}} = w_{h_1 o_4} \frac{\partial E_2}{\partial z_{o4}} + w_{h_1 o_5} \frac{\partial E_2}{\partial z_{o5}} \tag{1.24}$$

$$\frac{\partial E_1}{\partial a_{h2}} = w_{h_2 o_2} \frac{\partial E_1}{\partial z_{o2}} + w_{h_2 o_3} \frac{\partial E_1}{\partial z_{o3}} \tag{1.25}$$

$$\frac{\partial E_2}{\partial a_{h2}} = w_{h_2 o_4} \frac{\partial E_2}{\partial z_{o4}} + w_{h_2 o_5} \frac{\partial E_2}{\partial z_{o5}} \tag{1.26}$$

$$\frac{\partial E_1}{\partial a_{h3}} = w_{h_3 o_2} \frac{\partial E_1}{\partial z_{o2}} + w_{h_3 o_3} \frac{\partial E_1}{\partial z_{o3}} \tag{1.27}$$

$$\frac{\partial E_2}{\partial a_{h3}} = w_{h_3 o_4} \frac{\partial E_2}{\partial z_{o4}} + w_{h_3 o_5} \frac{\partial E_2}{\partial z_{o5}} \tag{1.28}$$

Afterwards the derivatives of E_1 and E_2 on the z_{hi} are therefore:

$$\frac{\partial E_1}{\partial z_{h1}} = \frac{\partial E_1}{\partial a_{h1}} \frac{\partial a_{h1}}{\partial z_{h1}} \tag{1.29}$$

$$\frac{\partial E_2}{\partial z_{h1}} = \frac{\partial E_2}{\partial a_{h1}} \frac{\partial a_{h1}}{\partial z_{h1}} \tag{1.30}$$

$$\frac{\partial E_1}{\partial z_{h2}} = \frac{\partial E_1}{\partial a_{h2}} \frac{\partial a_{h2}}{\partial z_{h2}} \tag{1.31}$$

$$\frac{\partial E_2}{\partial z_{h2}} = \frac{\partial E_2}{\partial a_{h2}} \frac{\partial a_{h2}}{\partial z_{h2}} \tag{1.32}$$

$$\frac{\partial E_1}{\partial z_{h3}} = \frac{\partial E_1}{\partial a_{h3}} \frac{\partial a_{h3}}{\partial z_{h3}} \tag{1.33}$$

$$\frac{\partial E_2}{\partial z_{h3}} = \frac{\partial E_2}{\partial a_{h3}} \frac{\partial a_{h3}}{\partial z_{h3}} \tag{1.34}$$

Finally, the derivatives of E_1 and E_2 on the input layer outputs are therefore:

$$\frac{\partial E_1}{\partial a_{input1}} = w_{x_1 h1} \frac{\partial E_1}{\partial z_{h1}} + w_{x_1 h2} \frac{\partial E_1}{\partial z_{h2}} + w_{x_1 h3} \frac{\partial E_1}{\partial z_{h3}} \tag{1.35}$$

$$\frac{\partial E_2}{\partial a_{input1}} = w_{x_1 h1} \frac{\partial E_2}{\partial z_{h1}} + w_{x_1 h2} \frac{\partial E_2}{\partial z_{h2}} + w_{x_1 h3} \frac{\partial E_2}{\partial z_{h3}} \tag{1.36}$$

$$\frac{\partial E_1}{\partial a_{input2}} = w_{x_2 h1} \frac{\partial E_1}{\partial z_{h1}} + w_{x_2 h2} \frac{\partial E_1}{\partial z_{h2}} + w_{x_2 h3} \frac{\partial E_1}{\partial z_{h3}} \tag{1.37}$$

$$\frac{\partial E_2}{\partial a_{input2}} = w_{x_2 h1} \frac{\partial E_2}{\partial z_{h1}} + w_{x_2 h2} \frac{\partial E_2}{\partial z_{h2}} + w_{x_2 h3} \frac{\partial E_2}{\partial z_{h3}} \tag{1.38}$$

And the derivatives of E on each a and z can be calculated as:

$$\frac{\partial E}{\partial a_{h1}} = \frac{\partial E_1}{\partial a_{h1}} + \frac{\partial E_2}{\partial a_{h1}} \tag{1.39}$$

$$\frac{\partial E}{\partial a_{h2}} = \frac{\partial E_1}{\partial a_{h2}} + \frac{\partial E_2}{\partial a_{h2}} \tag{1.40}$$

$$\frac{\partial E}{\partial a_{h3}} = \frac{\partial E_1}{\partial a_{h3}} + \frac{\partial E_2}{\partial a_{h3}} \tag{1.41}$$

$$\frac{\partial E}{\partial z_{h1}} = \frac{\partial E_1}{\partial z_{h1}} + \frac{\partial E_2}{\partial z_{h1}} \tag{1.42}$$

$$\frac{\partial E}{\partial z_{h2}} = \frac{\partial E_1}{\partial z_{h2}} + \frac{\partial E_2}{\partial z_{h2}} \tag{1.43}$$

$$\frac{\partial E}{\partial z_{h3}} = \frac{\partial E_1}{\partial z_{h3}} + \frac{\partial E_2}{\partial z_{h3}} \tag{1.44}$$

$$\frac{\partial E}{\partial a_{input1}} = \frac{\partial E_1}{\partial a_{input1}} + \frac{\partial E_2}{\partial a_{input1}} \tag{1.45}$$

$$\frac{\partial E}{\partial a_{input2}} = \frac{\partial E_1}{\partial a_{input2}} + \frac{\partial E_2}{\partial a_{input2}} \tag{1.46}$$

So above are all the backpropagation process. **Q.E.D.**

Bibliography

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.