COMP9444 Neural Networks
Week 13
Mixture of Experts

The probabilistic generative model used for the Mixture of Experts Model:

1. An input vector $\vec{x}$ is picked at random from some prior distribution.
2. A particular rule is selected according to the conditional probability $P(k|\vec{x}, \vec{a}^{(0)})$, given $\vec{x}$ and $\vec{a}^{(0)}$.
3. for each expert $k \in \{1, ..., K\}$, the model response $d$ is linear in $\vec{x}$ with an additive error $\epsilon_k$ modelled as a Gaussian distributed random variable with zero means and unit variance.

From this the following probability distribution of the random variable $D$ being the response of the mixture of experts model can be derived:

$$P(D = d|\vec{x}, \vec{\theta}^{(0)}) = \sum_{k=1}^{K} P(D = d|\vec{x}, \vec{w}_k^{(0)}) P(k|\vec{x}, \vec{a}^{(0)})$$

where $\vec{\theta}^{(0)}$ is the generative model parameter vector denoting the combination of $\vec{a}^{(0)}$ and $\{\vec{w}_k^{(0)}\}_{k=1}^{K}$.
\[ y_k = \mathbf{w}_k^T \mathbf{x} \]

\[ g_k = \frac{\exp(u_k)}{\sum_{j=1}^{K} \exp(u_j)} \]

where \( u_k = \mathbf{a}_k^T \mathbf{x} \)

The softmax function of \( g_k \) has the following properties for all \( k \in \{1, ..., K\} \):

\[ 0 \leq g_k \leq 1 \]

and

\[ \sum_{k=1}^{K} g_k = 1 \]

Let \( y_k \) denote the output of expert \( k \). Then the overall output of the Mixture of Expert model is given by:

\[ y = \sum_{k=1}^{K} g_k y_k \]
The probability density function of $D$, given the input vector $\vec{x}$ and given that the $k^{th}$ expert of the probabilistic generative model is selected, can be described as:

$$f_D(\vec{x}, k, \vec{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(d - y_k)^2\right)$$

The probability density function of $D$, given $\vec{x}$ is the mixture of the probability density functions

$$\{f_D(\vec{x}, k\vec{\theta})\}_{k=1}^{K}$$

with the mixing parameters being the multinomial probabilities determined by the gating network. We may thus write

$$f_D(d|\vec{x}, \vec{\theta}) = \sum_{k=1}^{K} g_k f_D(d|\vec{x}, k, \vec{\theta})$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{K} g_k \exp\left(-\frac{1}{2}(d - y_k)^2\right)$$

This probability distribution is also called an associative Gaussian mixture model.
Hierarchical Mixture of Experts

\[ f_D(d|\bar{x}, \theta) = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{K} g_k \sum_{j=1}^{2} g_{j|k} \exp \left( -\frac{1}{2} (d - y_{jk})^2 \right) \]

A posteriori probabilities

\[ h_k = \frac{g_k \sum_{j=1}^{2} g_{j|k} \exp \left( -\frac{1}{2} (d - y_{jk})^2 \right)}{\sum_{k=1}^{K} g_k \sum_{j=1}^{2} g_{j|k} \exp \left( -\frac{1}{2} (d - y_{jk})^2 \right)} \]

and

\[ h_{j|k} = \frac{g_{j|k} \exp \left( -\frac{1}{2} (d - y_{jk})^2 \right)}{\sum_{j=1}^{2} g_{j|k} \exp \left( -\frac{1}{2} (d - y_{jk})^2 \right)} \]

The product of \( h_k \) and \( h_{j|k} \) defines the joint a posteriori probability that expert \((j, k)\) produces an output \(y_{jk}\) that matches the desired response \(d\) as given by

\[ h_{jk} = h_k h_{j|k} \]
Hierarchical Mixture of Experts

Let us denote the likelihood function by $l(\tilde{\theta})$ which is given by the probability density function $f_{D}(d|\tilde{x}, \tilde{\theta})$ viewed as a function of $\tilde{\theta}$ only.

$$l(\tilde{\theta}) = f_{d}(d|\tilde{x}, \tilde{\theta})$$

$$L(\tilde{\theta}) = \log[l(\tilde{\theta})] = \log[f_{D}(d|\tilde{x}, \tilde{\theta})]$$

We seek the parameter vector $\theta$ for which $l(\tilde{\theta})$ is maximum. Hence we need to find the solution to

$$\frac{\partial}{\partial \tilde{\theta}} l(\tilde{\theta}) = 0$$

As the logarithm is a monotonic function this is also a solution to

$$\frac{\partial}{\partial \tilde{\theta}} L(\tilde{\theta}) = 0$$
Learning Strategies

\[
\frac{\partial L}{\partial \vec{w}_{jk}} = h_{j|k}(n)h_k(n)(d(n) - y_{jk}(n)) \vec{x}(n)
\]

\[
\frac{\partial L}{\partial \vec{a}_k} = (h_k(n) - g_k(n)) \vec{x}(n)
\]

\[
\frac{\partial L}{\partial \vec{a}_{jk}} = h_k(n)(h_{j|k}(n) - g_{j|k}(n)) \vec{x}(n)
\]
1. Selection of splits: Let $t$ be a node of the tree $T$. Let $\bar{d}(t)$ denote the average of $d_i$ for all cases $(\vec{x}_i, d_i)$ falling into $t$. I.e.

$$\bar{d}(t) = \frac{1}{N(t)} \sum_{x_i \in \{\text{examples in } t\}} d_i$$

where $N(t)$ is the number of examples falling into $t$.

2. Define

$$\varepsilon(t) = \frac{1}{N(t)} \sum_{x_i \in \{\text{examples in } t\}} (d_i - \bar{d}(t))^2$$

and

$$\varepsilon(T) = \sum_{t \in T} \varepsilon(t)$$

Suppose a node $t$ can be split into two subnodes; i.e. the examples associated to $t$ can be split into two subsets by using some criteria. Then,

$$\Delta \varepsilon(s, t) = \varepsilon(t) - \varepsilon(t_L) - \varepsilon(t_R)$$
CART cont’d

The best split $s^*$ is taken for which we have

$$\Delta \varepsilon (s^*, t) = \max_s \Delta \varepsilon (s, t)$$

A node $t$ is decided to be a terminal node, if

$$\max_s \Delta \varepsilon (s, t) > \beta$$

for some use specified threshold $\beta$.

For a terminal node $t$, a linear regression function is determined using the least square criterion. I.e. a weight vector is determined such that the following equation is satisfied:

$$\vec{w} = \arg \min_{s \in t} \sum (\vec{w}^T \vec{x} - d_s)^2$$
The Expectation Maximisation (EM) Algorithm

A parameter vector $\tilde{\theta}$ needs to be found that maximises

$$L(\tilde{\theta}) = \log f_D(d|\tilde{\theta})$$

As our generative model assumes a mixture of potential sources of the data (the various experts) we need to assess from which of the experts the data was generated. This is done by considering indicator functions for each expert which are treated as additional probability variables in the vector $r$ which is the vector $\theta$ extended by the indicator variables.

$$L_c(\tilde{\theta}) = \log f_D(\tilde{r}|\tilde{\theta})$$

$$Q(\tilde{\theta}, \tilde{\theta})(n)) = \mathbb{E}[L_c(\tilde{\theta})]$$

$$\tilde{\theta}(n + 1) = \arg \max_{\tilde{\theta}} Q(\tilde{\theta}, \tilde{\theta}(n))$$
Application of the EM algorithm to HME

\[ f_D(d_i|x_i, \theta) = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{K} g_k^{(i)} g_j^{(i)} \exp\left(-\frac{1}{2}(d_i - y_{jk}^{(i)})^2\right) \]

\[ L(\theta) = \log\left[ \prod_{i=1}^{N} f_D(d_i|x_i, \theta) \right] \]

\[ L(\theta) = \sum_{i=1}^{N} \log\left[ \sum_{k=1}^{K} g_k^{(i)} g_j^{(i)} \exp\left(-\frac{1}{2}(d_i - y_{jk}^{(i)})^2\right) \right] \]
Application of the EM algorithm to the HME

Let $z_k^{(i)}$ denote the indicator functions that indicate whether example $i$ is generated by ‘expert’ $k$ or not and having values ’1’ and ’0’ respectively. Let $z_{j|k}^{(i)}$ denote the indicator functions that indicate whether expert $j$ given that expert $k$ was selected in the HME produced the example.

Then for the expectation step of the EM algorithm, we are interested in

$$E[z_k^{(i)}] = P(z_k^{(i)} = 1 | \vec{x}_i, d_i, \hat{\theta}(n))$$

$$= h_k^{(i)}$$

and

$$E[z_{j|k}^{(i)}] = P(z_{j|k}^{(i)} = 1 | \vec{x}_i, d_i, \hat{\theta}(n))$$

$$= h_{j|k}^{(i)}$$
Application of the EM algorithm to the HME

Let denote $z_{jk}^{(i)}$ the indicator variable saying that expert $jk$ (at the lowest level of the hierarchy), has produced the example.

Then we have

$$E[z_{jk}^{(i)}] = E[z_{jk}^{(i)}z_k^{(i)}] = E[z_{jk}^{(i)}]E[z_k^{(i)}]$$

$$= h_{jk}^{(i)}h_k^{(i)} = h_{jk}^{(i)}$$
Application of the EM algorithm to the HME

\[ L_c(\theta) = \log \left[ \prod_{i=1}^{N} f_c(d_i, z^{(i)}_{jk} | \vec{x}_i, \vec{\theta}) \right] \]

\[ = \log \left[ \prod_{i=1}^{N} \prod_{j=1}^{2} \prod_{k=1}^{2} (g_k^{(i)} g_{jk}^{(i)} f_{jk}(d_i))^z_{jk}^{(i)} \right] \]

\[ = \sum_{i=1}^{N} \sum_{j=1}^{2} \sum_{k=1}^{2} z_{jk}^{(i)} [\log g_k^{(i)} + \log g_{jk}^{(i)} + \log f_{jk}(d_i)] \]

ignoring the constant \((1/2) \log(2\pi)\) we can write

\[ L_c(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{2} \sum_{k=1}^{2} z_{jk}^{(i)} [\log g_k^{(i)} + \log g_{jk}^{(i)} - \frac{1}{2}(d_i - y_{jk}^{(i)})^2] \]
Application of the EM algorithm to the HME

The first expectation step of the EM algorithm is then given by equating

\[ Q(\vec{\theta}, \hat{\vec{\theta}}(n)) = E[L_c(\vec{\theta})] \]

\[ = \sum_{i=1}^{N} \sum_{j=1}^{2} \sum_{k=1}^{2} E[z_{jk}^{(i)}] (\log g_k^{(i)} + \log g_{j|k}^{(i)} - \frac{1}{2}(d_i - y_{jk}^{(i)})^2) \]

Which is equivalent to

\[ = \sum_{i=1}^{N} \sum_{j=1}^{2} \sum_{k=1}^{2} h_{jk}^{(i)} (\log g_k^{(i)} + \log g_{j|k}^{(i)} - \frac{1}{2}(d_i - y_{jk}^{(i)})^2) \]
Application of the EM algorithm to the HME

The maximisation step:

\[
\tilde{w}_{jk}(n + 1) = \arg \max_{w_{jk}} \sum_{i=1}^{N} h_{jk}^{(i)} (d_i - y_{jk}^{(i)})^2
\]

\[
\tilde{a}_j(n + 1) = \arg \max_{a_j} \sum_{i=1}^{N} \sum_{k=1}^{2} h_k^{(i)} \log g_k^{(i)}
\]

\[
\tilde{a}_{jk}(n + 1) = \arg \max_{a_{jk}} \sum_{i=1}^{N} \sum_{l=1}^{2} h_l^{(i)} \sum_{m=1}^{2} h_{m|l}^{(i)} \log g_{m|l}^{(i)}
\]