Hierarchical Mixture of Experts

Presented by Yongqia Shao and Juan Wu
Academic English for Statistics
12/09/2014
Outline

- Introduction
- Hierarchical mixture of experts
- E-M algorithm
- Experimental results
TREE

- An example
Hierarchical Mixture of Experts

**Soft decision tree:** Takes a weighted (gating) average of all leaves (experts), as opposed to using a single path and a single leaf.

- “soft” partition: tree splits are probabilistic
- Splits can be multiway
- A linear model is fit in each terminal node
Hierarchical Mixture of Experts
Hierarchical Mixture of Experts

**Soft decision tree**: Takes a weighted (gating) average of all leaves (experts), as opposed to using a single path and a single leaf.

- “soft” partition: tree splits are probabilistic
- Splits can be multiway
- A linear model is fit in each terminal node
Hierarchical Mixture of Experts
Expert Network Output

- At the leaves of trees

for each expert:

\[ \mu_{ij} = f(U_{ij}x) \]

output of the expert

Model for response variable
Expert Network Output

For each expert, assume the true output $y$ is chosen from a distribution $P$ with parameters $\theta_{ij}$

$$Y \sim P(y | x, \theta_{ij})$$

Regression: The Gaussian linear regression model is used:

$$Y = \beta_{ij}^T x + \epsilon, \epsilon \sim N(0, \sigma_{ij}^2)$$

Classification: The linear logistic regression model is used:

$$P(Y = 1 | x, \theta_{ij}) = \frac{1}{1 + e^{-\theta_{ij}^T x}}$$
Gating network output

- At the nonterminal of the tree top level:

\[ \xi_i = v_i^T x \]

\[ g_i = \frac{\exp(\xi_i)}{\sum_k \exp(\xi_k)} \]

\[ \sum_i g_i = 1 \]

other level:

\[ \xi_{ij} = v_{ij}^T x \]

\[ g_{j|i} = \frac{\exp(\xi_{ij})}{\sum_k \exp(\xi_{ik})} \]

\[ \sum_j g_{j|i} = 1 \]
Gating Network Output

- At the non-leaves nodes
  
  top node:
  \[ \mu = \sum_i g_i \mu_i \]

  other nodes:
  \[ \mu_i = \sum_j g_{j|i} \mu_{ij} \]
Hierarchical Mixture of Experts

\[ \mu = \sum_i g_i \sum_j g_{j|i} f(U_{ij}x) \]
Therefore, for data set \( X = \{ x^{(t)}, y^{(t)} \}_{1}^{N} \), the total probability of generating \( y \) from \( x \) is given by

\[
P(Y|X, \theta) = \prod_{t} \sum_{i} g_{i}^{(t)}(x, v_{i}) \sum_{j} g_{j|i}^{(t)}(x, v_{ij}) P^{(t)}(y|x, \theta_{ij})
\]

\[
\ln P(Y|X, \theta) = \sum_{t} \ln \left( \sum_{i} g_{i}^{(t)}(x, v_{i}) \sum_{j} g_{j|i}^{(t)}(x, v_{ij}) P^{(t)}(y|x, \theta_{ij}) \right)
\]
E-M algorithm

- Introduce latent variables $z_{ij}$ which have an interpretation as the labels that corresponds to the experts.
- The probability model can be simplified with the knowledge of latent variables

$$P(y^{(t)}, z_{ij}^{(t)} | x^{(t)}, \theta) = g_{i}^{(t)} g_{j|i}^{(t)} P_{ij} (y^{(t)}) = \prod_{i} \prod_{j} \left\{ g_{i}^{(t)} g_{j|i}^{(t)} P_{ij} (y^{(t)}) \right\}^{z_{ij}^{(t)}}$$
E-M algorithm

- Log-likelihood function:

\[
l_c(\theta; y) = \sum_t \sum_i \sum_j z_{ij}^{(t)} \left\{ \ln g_i^{(t)} + \ln g_{ji}^{(t)} + \ln P_{ij}(y^{(t)}) \right\}
\]
E-M algorithm

Define posterior probabilities ... and we get ...

\[ h_{ij} = \frac{g_i g_{j|i} P_{ij}(y)}{\sum_i g_i \sum_j g_{j|i} P_{ij}(y)} \]

\[ E[z_{ij}^{(t)} | \mathcal{X}] = h_{ij}^{(t)} \]

\[ h_i = \frac{g_i \sum_j g_{j|i} P_{ij}(y)}{\sum_i g_i \sum_j g_{j|i} P_{ij}(y)} \]

\[ E[z_i^{(t)} | \mathcal{X}] = h_i^{(t)} \]

\[ h_{j|i} = \frac{g_{j|i} P_{ij}(y)}{\sum_j g_{j|i} P_{ij}(y)} \]

\[ E[z_{j|i}^{(t)} | \mathcal{X}] = h_{j|i}^{(t)} \]
E-M algorithm

- The E-step

\[
Q(\theta, \theta^{(p)}) = E_\mathcal{Z}(I_c(\theta; y)) = \sum_t \sum_i \sum_j h_{ij}^{(t)} \left\{ \ln g^{(t)}_i + \ln g^{(t)}_{j|i} + \ln P_{ij}(y^{(t)}) \right\}
\]

where we have used the fact that:

\[
E[z_{ij}^{(t)}|X] = P(z_{ij}^{(t)} = 1|y^{(t)}, x^{(t)}, \theta^{(p)})
= \frac{P(y^{(t)}|z_{ij}^{(t)} = 1, x^{(t)}, \theta^{(p)})P(z_{ij}^{(t)} = 1|x^{(t)}, \theta^{(p)})}{P(y^{(t)}|x^{(t)}, \theta^{(p)})}
= \frac{P(y^{(t)}|x^{(t)}, \theta_{ij}^{(p)})g^{(t)}_i g^{(t)}_{j|i}}{\sum_i g^{(t)}_i \sum_j g^{(t)}_{j|i} P(y^{(t)}|x^{(t)}, \theta_{ij}^{(p)})}
= h_{ij}^{(t)}.
\]
E-M algorithm

- The M-step

\[
\theta_{ij}^{p+1} = \arg \max_{\theta_{ij}} \sum_t h_{ij}^{(t)} \ln P_{ij}(y^{(t)})
\]

\[
v_{ij}^{p+1} = \arg \max_{v_{ij}} \sum_t \sum_k h_{ij}^{(t)} \ln g_{ij}^{(t)}
\]

\[
v_{ij}^{p+1} = \arg \max_{v_{ij}} \sum_t \sum_k h_{ij}^{(t)} \sum_l h_{ij}^{(t)} \ln g_{ij}^{(t)}
\]
Results

- Simulated data of a four-joint robot arm moving in three-dimensional space

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Relative Error</th>
<th># Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>.31</td>
<td>1</td>
</tr>
<tr>
<td>backprop</td>
<td>.09</td>
<td>5,500</td>
</tr>
<tr>
<td>HME (Algorithm 1)</td>
<td>.10</td>
<td>35</td>
</tr>
<tr>
<td>HME (Algorithm 2)</td>
<td>.12</td>
<td>39</td>
</tr>
<tr>
<td>CART</td>
<td>.17</td>
<td>NA</td>
</tr>
<tr>
<td>CART (linear)</td>
<td>.13</td>
<td>NA</td>
</tr>
<tr>
<td>MARS</td>
<td>.16</td>
<td>NA</td>
</tr>
</tbody>
</table>
Results
Thank you

Hierarchical Mixture of Experts