EB-Hyp Algorithm

Algorithm 1 Pseudocode for EB-Hyp
1. inputs training data $X$, inference algorithm $A : (X, \eta) \rightarrow p(\theta | X, \eta)$, set of posteriors $p(\theta | X_{\text{train}}, \eta^{(s)})$.
2. output predictive density $p(X_{\text{test}} | X)$, optimal hyper-hyper-parameter $\hat{\lambda}$.
3. while $V$ not converged do
4. draw performance $f^*$ sample from GP posterior $\tilde{p}(f^* | \eta) \sim GP(\cdot | V)$
5. calculate hyperparameter posterior $\tilde{p}(\eta | X) \sim Z^{-1} \tilde{p}(f^* | X)p(\eta)$
6. draw next evaluation point $\eta^{(s)} \leftarrow \arg \max \tilde{p}(f^* | X)$
7. run parameter inference conditioned on hyp. $p(\theta | \eta^{(s)}) \sim A(X, \eta^{(s)})$
8. evaluate performance $f^{(s)} \leftarrow \int p(X|\theta)p(\theta | \eta^{(s)})d\theta$
9. append $(\eta^{(s)}, f^{(s)})$ to history $V$
10. end while
11. find optimal $\hat{\lambda}$
12. return approximation to $p(X_{\text{test}} | X)$ and $\hat{\lambda}$

Results

- Predictive log likelihood for LDA, 20 Newsgp dataset
  - Method Predictive Log Lik. (% Improvement on Baseline)
  - BayesOpt-TS with validation -357648 (0.00%)
    - without validation -361661 (-1.12%)
  - EB-Hyp with validation -357650 (-0.00%)
    - without validation -351911 (+1.60%)
  - Random -266074 (-64.5%)

- Predictive log lik. for deep latent Gaussian model, Labeled Faces in the Wild
  - Method Predictive Log Lik. (% Improvement on Baseline)
  - BayesOpt-TS with validation -17071 (0.00%)
    - without validation -15970 (-6.45%)
  - EB-Hyp with validation -16375 (-4.08%)
    - without validation -15872 (+7.02%)
  - Random -17271 (-1.17%)

Conclusions

- Introduced a general-purpose procedure for dealing with unknown hyperparameters that control the behaviour of machine learning models.
- EB-Hyp is based on approximately marginalizing the hyperparameters by taking a weighted average of posteriors calculated by existing inference algorithms that are time intensive.
- This work points toward a tendency of the standard optimization-based methodologies to overfit hyperparameters. Other things being equal, this tendency is less sensitive to hyperparameters compared to methods that are less sensitive. The result is a bias in the literature towards methods whose generalization performance is less sensitive to hyperparameters. Averaging approaches like EB-Hyp help reduce this bias.

Table of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>the model parameters</td>
</tr>
<tr>
<td>$\eta$</td>
<td>the parameters</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>the hyperparameters</td>
</tr>
<tr>
<td>$\hat{\eta}$</td>
<td>the hyperparameters fit by empirical Bayes</td>
</tr>
<tr>
<td>$\hat{\lambda}$</td>
<td>the hyperparameters fit by empirical Bayes</td>
</tr>
<tr>
<td>$X$</td>
<td>training data</td>
</tr>
<tr>
<td>$X_{\text{test}}$</td>
<td>unseen test data</td>
</tr>
</tbody>
</table>

References