The Bayesian Learning Rule (2022)

M. E. Khan & H. Rue

Aminata Ndiaye & Edimah Songo
M2 MASH

Outline

Motivation

- Create a unified framework to derive established and new algorithms
- Improve existent algorithms (convergence speed)
- Find new algorithms

How we proceed?

- Optimize a Bayesian objective function
- Finding the posterior distribution of the parameters of interest
- Taking more information into account by using the natural gradient

Principle

Bayesian Objective

$$q_*(\theta) = \arg\min_{q(\theta)} \mathbb{E}_q \left[\sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) \right] + \mathbb{D}_{KL}[q(\theta) \parallel p(\theta)]$$

Setting

- $q \in \mathcal{Q}$ a set of regular and minimal exponential family
- $\lambda_{t+1} \leftarrow \lambda_t - \rho_t \tilde{\nabla}_\lambda \left[\mathbb{E}_{q_t}[\bar{\ell}(\theta)] - \mathcal{H}(q_t) \right]$

Steps

- Choice of posterior approximation (here in the exponential family)
- Choice of approximation method for natural gradient (ex: Delta Method, etc.)

BLR and Ridge regression

Quadratic loss for some penalty term $\delta > 0$

$$\bar{\ell}(\theta) = \frac{1}{2}(y - X\theta)^T(y - X\theta) + \frac{1}{2}\delta\theta^T\theta$$
$$\theta^* = (X^TX + \delta I)^{-1}X^Ty$$

Natural gradients in Ridge Regression

- **Candidate posterior:** $\mathcal{N}(m, S)$
- $\mu = E[T(\theta)]$
- **Natural gradients:**
  $\tilde{\nabla}_{\mu^{(1)}} E_q[\bar{\ell}(\theta)] = -X^Ty$ and $\tilde{\nabla}_{\mu^{(2)}} E_q[\bar{\ell}(\theta)] = \frac{1}{2}(X^TX + \delta I)\mu^{(2)}$
- **Solution:** $\theta^* = m^* = (S^*)^{-1}X^Ty = (X^TX + \delta I)^{-1}X^Ty$

Conclusion

Article's takeaways

- **Research possibilities** new for state-of-the-art algorithms
- Only two choices are required (posterior and natural gradient approximations)

Drawbacks

- The article was at times difficult to understand
- Algorithms not necessarily optimal
  - Restriction to exponential families
  - Difficult to compute gradients
  - Topological instability

Potential solution to some issues

New research: *Lie-Group Bayesian Learning Rule (2023)*