Fast Laplace Approximation for Sparse Bayesian Spike and Slab Models
Syed Abbas Zilqurnain Naqvi¹, Shandian Zhe¹, Yuan Qi¹ and Jieping Ye²
¹Purdue University ²University of Michigan, Ann Arbor

Introduction

Goal of the paper: We develop approximate inference technique for the Bayesian spike-and-slab models for very high dimensional feature selection problems.

Challenge: For very high dimensional problems, existing MCMC methods converge slowly; and the variational Bayes (VB) and expectation propagation (EP) approaches, unless they enforce structural constraints on the posterior, are impractical for large data.

Solution: To address the computational issue, we develop the (FLAS) model. The features of our approach include:

- FLAS is a hybrid of frequentist and Bayesian treatment, enjoying the benefits of both worlds. It is computationally as efficient as the frequentist methods.
- It is free of any factorization assumptions on the joint posterior, but still enjoys a linear cost $O(np)$.

Evaluation: Our new method FLAS performs feature selection better than or comparable to the alternative approximate methods with less running time, and provides higher prediction accuracy than various alternative sparse methods.

Model

The hierarchical Bayesian model for regression is:

$$p(t | X, w, \tau) = \prod_{i=1}^{n} \mathcal{N}(t_i | x_i^\top w, \tau^{-1})$$

$$p(w | z) = \prod_{j=1}^{p} \mathcal{N}(w_j | 0, \tau_{ij}^2)$$

$$p(z_j = 1 | s_j) = s_j \quad (1 \leq j \leq p)$$

where $w$ are regression weights, $\tau$ is the precision parameter, and $X = [x_1, \ldots, x_n]^\top$. $z_j$ is a binary selection indicators for the $j$-th feature, and $s_j$ is a selection probability with uninformative prior $p(s_j) = \text{Beta}(a_0, b_0)$, with $a_0 = b_0 = 1$. For classification, $p(t | X, w) = \prod_{i=1}^{n} \mathcal{N}(t_i | x_i^\top w, \tau^{-1})$-$\mathcal{N}(t_i | \mu, \tau^{-1})$ where $t_i \in \{0, 1\}$, $w$ are classifier weights, $\sigma(a) = 1/(1 + \exp(-a))$, and $t = [t_1, \ldots, t_n]^\top$.

MAP estimation for Laplace approximation

We use two scalable nonconvex optimization methods, L-BFGS and GIST. For L-BFGS(FLAS), we marginalize out both $z$ and $s$ and do the following optimization:

$$\min_{w,s} F(w, s) = \min_w L(w) - \sum_{j=1}^{p} \log \left( \frac{1}{2} \mathcal{N}(w_j | 0, \tau_{ij}^2) + \frac{1}{2} \mathcal{N}(w_j | 0, \tau_{ij}^2) \right).$$

For GIST (FLAS*), we only marginalize out $z$ and jointly optimize $w$ and $s$:

$$\min_{w,s} F(w, s) = \min_w L(w) + \min_s R(w, s)$$

where $R(w, s) = \sum_{j=1}^{p} R_j(w_j, s_j)$ and $R_j(w_j, s_j) = - \log (s_j \mathcal{N}(w_j | 0, \tau_{ij}^2) + (1 - s_j) \mathcal{N}(w_j | 0, \tau_{ij}^2))$.

Marginal Posterior variance estimation using Ensemble Nystrom

The inverse of Hessian for regression is approximated using Nystrom method as:

$$\mathbf{H}^{-1} \approx \mathbf{H}^* = \tau \mathbf{X}^\top \mathbf{X} + \text{diag}(v).$$

Since we can choose $k \ll p$, the inversion cost will still be linear in $p$. For classification, $\mathbf{H} = \mathbf{X}^\top \mathbf{X} + \text{diag}(v)$. Applying Woodbury matrix identity we can estimate the diagonal entries in $O(nkp^2)$ time:

$$\text{diag}(\mathbf{H}^{-1}) = \text{diag}(\mathbf{H}^*)^{-1} - \text{diag}(\mathbf{H})^{-1} \mathbf{X} \text{diag}(\mathbf{H})^{-1} \mathbf{X}^\top \text{diag}(\mathbf{H})^{-1} \mathbf{X} \text{diag}(\mathbf{H})^{-1}.$$