Scalable Approximate Inference for the Bayesian Nonlinear Support Vector Machine

Florian Wenzel, Matthäus Deutsch, Théo Galy-Fajou, Marius Kloft

Department of Computer Science, Humboldt-Universität zu Berlin

{wenzelfl,deutschm,galy,kloft}@hu-berlin.de

Motivation

- There has recently been significant interest in utilizing max-margin based discriminative Bayesian models for various applications
- Most approaches build on a Bayesian formulation of the SVM
- State-of-the-art inference methods are either slow or rely on point estimates
- We propose a *fast inference scheme* based on variational inference for approximating the full posterior
- Our method leads to a fast *auto-tuned SVM* and gives an *uncertainty prediction*

The Bayesian SVM

- Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ be n observations, where $x_i \in \mathbb{R}^d$ is a data point with corresponding label $y_i \in \{-1, 1\}$

The Support Vector Machine (SVM)

- The SVM consists of finding the optimal score function f by solving

$$\arg\min_{f(x)} \sum_{i=1}^{n} \max(1 - y_i f(x_i), 0) + \gamma R(f) \quad (1)$$

- R is a regularizer function and γ a hyperparameter

The Bayesian Linear SVM (Linear BSVM)

- We follow the approach of [1] to develop a Bayesian formulation of the linear SVM
- We introduce latent variables $\lambda := (\lambda_1, \dots, \lambda_n)^\top$ (with improper prior)
- The (proper) *full conditionals* of this model are given by

$$\begin{aligned} \beta|\lambda, \Sigma, \mathcal{D} &\sim \mathcal{N}(BZ(\lambda^{-1} + 1),\, B), \\ \lambda_i|\beta, \mathcal{D}_i &\sim \mathcal{GIG}\left(\frac{1}{2}, 1, (1 - y_i x_i^\top \beta)^2\right) \end{aligned} \quad (2)$$

- where $Z = XY$ and $B^{-1} = Z\Lambda^{-1}Z^\top + \Sigma^{-1}$, $\Lambda = \mathrm{diag}(\lambda)$, $Y = \mathrm{diag}(y)$

The Bayesian Nonlinear SVM (Kernel BSVM)

- [2] developed a kernelized version of the linear model (using ideas of GPs)
- We assume that a continuous decision function $f(x)$ is drawn from a zero-mean GP
- The *full conditionals* of the model are

$$\begin{aligned} f|\lambda, \mathcal{D} &\sim \mathcal{N}(CY(\lambda^{-1}),\, C), \\ \lambda_i|\beta, \mathcal{D}_i &\sim \mathcal{GIG}\left(\frac{1}{2}, 1, (1 - y_i f_i)^2\right) \end{aligned} \quad (3)$$

- where $C^{-1} = \Lambda^{-1} + K^{-1}$ and K is the kernel matrix

Inference

Variational Inference (VI)

VI for the Linear BSVM

- We follow the mean field approach and choose the *variational distributions:*

$$q(\lambda_i) \equiv \mathcal{GIG}(\tfrac{1}{2}, 1, \alpha_i), \qquad q(\beta) \equiv \mathcal{N}(\mu, \zeta) \quad (4)$$

- where $\alpha_i \geq 0$, $\mu \in \mathbb{R}^d$, $\zeta \in \mathbb{R}^{d \times d}$ (positive definite) are free parameters
- The coordinate ascent (*CAVI*) updates are

$$\begin{aligned} \alpha_i &= (1 - z_i^T \mu)^2 + z_i^T \zeta z_i, \\ \zeta &= \left(ZA^{-\frac{1}{2}}Z^T + \Sigma^{-1}\right)^{-1} \\ \mu &= \zeta Z(\alpha^{-\frac{1}{2}} + 1) \end{aligned} \quad (5)$$

- where $A = \mathrm{diag}(\alpha)$ and $\alpha = (\alpha_i)_{1 \leq i \leq n}$

VI for the Kernel BSVM

- We choose the variational distributions $q(\lambda), q(f)$ similar to the linear case to be in the same family as the full conditionals (3)
- The coordinate ascent updates (*CAVI*) are

$$\begin{aligned} \alpha_i &= (1 - y_i \mu_i)^2 + \zeta_{ii} \\ \zeta &= \left(A^{-\frac{1}{2}} + K^{-1}\right)^{-1} \\ \mu &= \zeta Y(\alpha^{-\frac{1}{2}} + 1) \end{aligned} \quad (6)$$

Stochastic Variational Inference (SVI)

- The variational infernce scheme for the *linear BSVM* can be directly extended to an SVI scheme (we use an adaptive learning rate schedule [3])
- This leads to great speed up (see experiments)
- The *kernel BSVM* does not have a set of global variables – therefore, SVI cannot be directly applied
- *Solution:* Use inducing point GP with global sparse prior [4] that would lead to an appropriate model for SVI (this is future work)

Beyond the Standard SVM

Reformulating the SVM as probabilistic models lets us apply attractive Bayesian methods such as:

- Computing class membership probabilities (uncertainty in the prediction)
- Automated hyperparameter search

Class Membership Probabilities

- Integrating over the approximate posterior obtained by our inference method lets us compute the class membership probability

Linear BSVM: $$p(y_* = 1|x_*, \mathcal{D}) \approx \Phi\left(\frac{x_*^\top \mu^*}{x_*^\top \zeta^* x_* + 1}\right) \quad (7)$$

Kernel BSVM: $$p(y_* = 1|x_*, \mathcal{D}) \approx \Phi\left(\frac{k_* K^{-1} \mu^*}{k_{**} + k_*^\top \left(K^{-1}\zeta^* K^{-1} - K^{-1}\right) k_* + 1}\right) \quad (8)$$

Hyperparameter Optimization

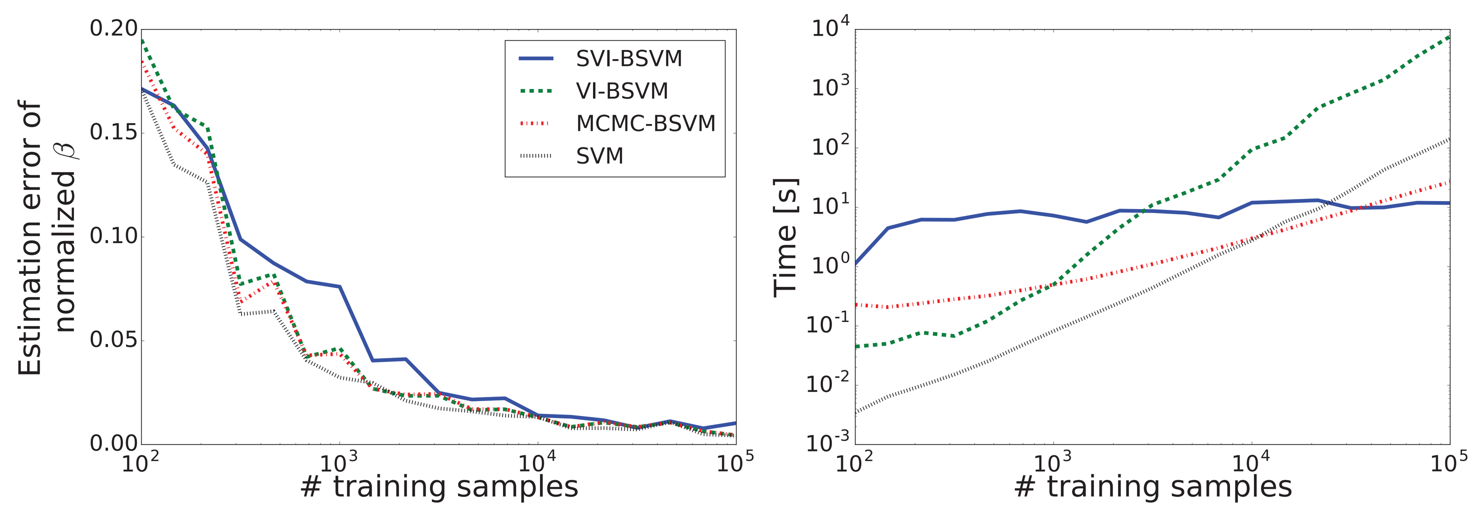
- We estimate the hyperparameters from the data by maximizing the fitted variational lower bound of the marginal likelihood $\mathcal{L}(h) \leq p(y|X, h)$
- We update the hyperparameters simultaneously with the variational parameters and add a hyperparameter optimization step after the variational updates

$$h^{(t)} = h^{(t-1)} + \tilde{\rho}_t \nabla_h \mathcal{L}(\alpha^{(t-1)}, \mu^{(t-1)}, \zeta^{(t-1)}, h) \quad (9)$$

Experiments

Linear BSVM: Prediction Performance and Time

- Synthetic data set with known underlying model parameter β



Kernel BSVM: Prediction Performance and Time

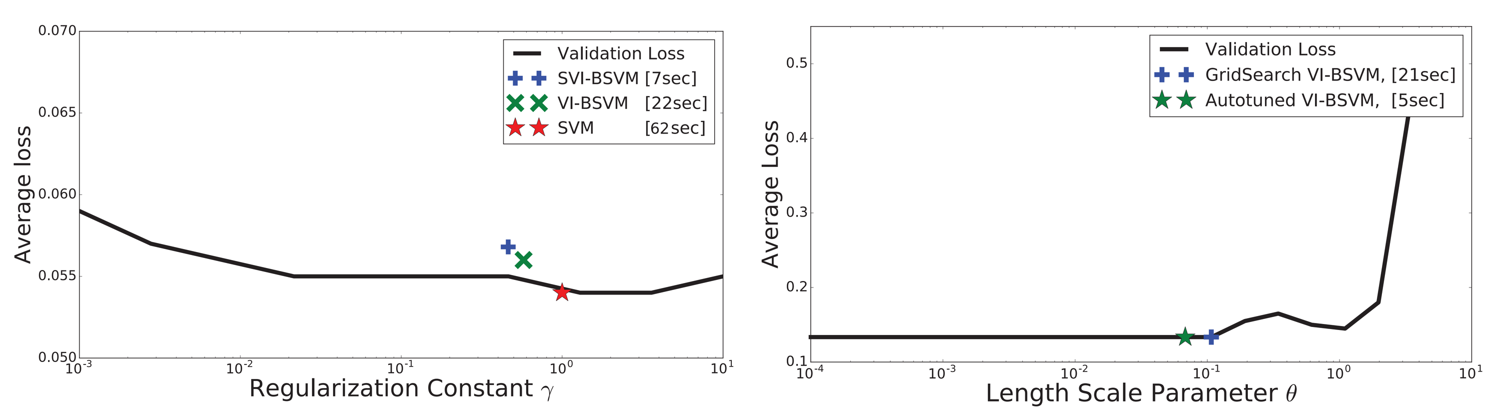
- Average prediction error (in %) from 10-fold cross validation:

Data set	N	d	VI-BSVM	LibSVM	GPC
Sonar	208	60	**12.5**	13.5	19.5
Crabs	200	7	**1.0**	**1.0**	3.1
Pima	768	8	**22.8**	24.7	22.8
USPS 3vs5	1540	256	2.0	**1.6**	2.3

- The state of the art MCMC based inference method for the kernel BSVM in [2] takes **1200 seconds** on the USPS dataset with prediction error 1.49% (reported by the authors)
- Our method only takes **15 seconds** while having only a slightly worse prediction error

Linear and Kernel BSVM: Automated Model Selection

- (Left) We estimate the *regularization constant* of the linear BSM and compare against grid search (grid of 1000 points) for the standard SVM
- (Right) We estimate the *length scale parameter* of the RBF kernel of the kernel BSVM



Conclusion and Forthcoming Research

- We proposed a new inference method for the Bayesian SVM that scales to large datasets and allows for approximating the full posterior
- We can automatically tune the hyperparameters of the SVM and compute the uncertainty in the predictions
- In future work we aim to develop an SVI method for the kernel BSVM applying the concept of GPs for big data [4]

References

[1] N. G. Polson and S. L. Scott, "Data augmentation for support vector machines," *Bayesian Anal.*, 2011.

[2] R. Henao, X. Yuan, and L. Carin, "Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling," in *Proceedings of the 27th International Conference on NIPS*, 2014.

[3] R. Ranganath, C. Wang, D. M. Blei, and E. P. Xing, "An Adaptive Learning Rate for Stochastic Variational Inference," *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[4] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," in *Conference on Uncertainty in Artificial Intellegence*, 2013.