Gaussian process methods for high dimensional learning

Tokyo Institute of Technology　Taiji Suzuki

In this presentation, we discuss the statistical performances of some Gaussian process methods for high dimensional learning problems. The main part of the talk consists of two parts: the first part is about a Bayes estimator with a Gaussian process prior for the sparse additive model and the second part is about a Gaussian process estimator for high dimensional low rank tensor. Finally, we discuss a combination of these two learning methods.

(1) **Sparse additive model**. The sparse additive model is a useful nonparametric model for high dimensional regression. We suppose that the covariate is composed of M variables $(x^{(1)}, \dots, x^{(M)})$ (they could be mutually dependent, actually, they could be copies of each other). Then the model assumes samples $D_n = \{(x_i, y_i)\}_{i=1}^n$ are generated from the following model:

$$y_i = \sum_{m=1}^{M} f_m^*(x_i^{(m)}) + \xi_i \qquad (i = 1, \dots, n),$$

where $\{\xi_i\}_{i=1}^n$ is an i.i.d. sequence of nose. We assume that the nonzero elements $I_0 := \{m \mid f_m^* \neq 0\}$ of f^* is small (sparse).

We can show that, by employing a Gaussian process prior with a sparsity inducing prior, the minimax optimal convergence rate is attained by the Bayes estimator associated with the prior. The convergence rate is characterized by the covering number of the unit ball of the Reproducing Kernel Hilbert Space (RKHS) corresponding to the Gaussian process prior [1].

(2) **High dimensional low rank tensor estimator**. Low rank tensor estimation is useful in several practical applications such as recommendation system, image-movie processing, multi-task learning, and reduced rank regression. Basically, the method extract higher order relations among multi-modal data. The model is described as follows. The true tensor $A^* \in \mathbb{R}^{M_1 \times \cdots \times M_K}$ is a tensor of degree K, and the samples $D_n = \{(Y_i, X_i)\}_{i=1}^n$ is generated by

$$Y_i = \langle A^*, X_i \rangle + \epsilon_i$$

where $X_i \in \mathbb{R}^{M_1 \times \cdots \times M_K}$, $\langle A^*, X_i \rangle = \sum_{i_1, \dots, i_K = 1}^{M_1, \dots, M_K} A^*_{i_1, \dots, i_K} X_{i,(i_1, \dots, i_K)}$, and ϵ_i is a Gaussian noise with mean 0 and variance σ^2 ($N(0, \sigma^2)$). Here we assume that the true tensor A^* is low rank in terms of *CP-rank*.

To estimate A^*, we employ a Gaussian process prior on the set of low rank tensors (more precisely, Gaussian chaos prior). That is, for the decomposition of a rank d' tensor $A_{i_1, \dots, i_K} = [[U^{(1)}, U^{(2)}, \dots, U^{(K)}]] =: \sum_{r=1}^{d'} U_{r,i_1} U_{r,i_2} \dots U_{r,i_K}$, we put a prior defined as

$$\pi(U^{(1)}, \dots, U^{(K)} | d') \propto \exp\left\{ - \tfrac{d'}{2\sigma_{\mathrm{p}}^2} \textstyle\sum_{k=1}^{K} \mathrm{Tr}[{U^{(k)}}^\top U^{(k)}] \right\}.$$

Moreover, we put a prior on the rank d' as $\pi(d') = \exp(-\zeta d')$ for arbitrary $\zeta > 0$. Then, it is shown that the Bayes estimator corresponding to the prior achieves the minimax optimal rate up to a logarithmic factor [2]. We will show that the Bayes estimator shows nice performances compared with existing convex tensor estimators in numerical experiments.

(3) Finally, we combine these two notions. Suppose that the covariate forms the following composition $x = (x^{(k,m)})_{k=1,m=1}^{K,M}$. We estimate a function $f^*(x) = \sum_{m=1}^{r} \prod_{k=1}^{K} f_m^{(k)*}(x^{(k,m)})$ from the samples generated as

$$y_i = f^*(x_i) + \epsilon_i, \quad (i = 1, \dots, n).$$

This function can also be estimated by employing Gaussian process prior on each $f_m^{(k)}$. The statistical performance of the Bayes estimator is also derived by combining the techniques developed in (1) and (2).

References

[1] T. Suzuki. Pac-bayesian bound for gaussian process regression and multiple kernel additive model. In *Conference on Learning Theory (COLT2012)*, pages 8.1–8.20, 2012.

[2] T. Suzuki. Convergence rate of Bayesian tensor estimator: Optimal rate without restricted strong convexity. In *International Conference on Machine Learning (ICML2015)*, pages 1273–1282, 2015.