Learnability of Gaussians with flexible variances

Ding-Xuan Zhou

City University of Hong Kong

E-mail: mazhou@cityu.edu.hk

Supported in part by Research Grants Council of Hong Kong

Start

October 20, 2007

Least-square Regularized Regression

Learn $f : X \to Y$ from random samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$

Take X to be a compact subset of $\mathbf{R}^n$ and $Y = \mathbf{R}$. $y \approx f(x)$
Due to noises or other uncertainty, we assume a (unknown) probability measure ρ on $Z = X \times Y$ governs the sampling.

marginal distribution ρ_X on X: $\{x_i\}_{i=1}^m$ drawn according to ρ_X

conditional distribution $\rho(\cdot|x)$ at $x \in X$

Learning the **regression function**: $f_\rho(x) = \int_Y y d\rho(y|x)$

$y_i \approx f_\rho(x_i)$

Learning with a Fixed Gaussian

$$f_{\mathbf{z},\lambda,\sigma} := \arg\min_{f \in \mathcal{H}_{K_\sigma}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_{K_\sigma}^2 \right\}, \qquad (1)$$

where $\lambda = \lambda(m) > 0$, and $K_\sigma(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}}$ is a Gaussian kernel on X

Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_{K_\sigma}$

completion of span$\{(K_\sigma)_t := K_\sigma(t, \cdot) : t \in X\}$ with the inner product $\langle \quad , \quad \rangle_{K_\sigma}$ satisfying $\langle (K_\sigma)_x, (K_\sigma)_y \rangle_K = K_\sigma(x, y)$.

Theorem 1 *(Smale-Zhou, Constr. Approx. 2007) Assume* $|y| \leq M$ *and that* $f_\rho = \int_X K_\sigma(x,y)g(y)d\rho_X(y)$ *for some* $g \in L^2_{\rho_X}$. *For any* $0 < \delta < 1$, *with confidence* $1-\delta$,

$$\|f_{\mathbf{z},\lambda,\sigma} - f_\rho\|_{L^2_{\rho_X}} \leq 2\log\Big(4/\delta\Big)\Big(12M\Big)^{2/3}\|g\|^{1/3}_{L^2_{\rho_X}}\Big(\frac{1}{m}\Big)^{1/3}$$

where $\lambda = \lambda(m) = \log\Big(4/\delta\Big)\Big(12M/\|g\|_{L^2_{\rho_X}}\Big)^{2/3}\Big(1/m\Big)^{1/3}$.

In Theorem 1, $f_\rho \in C^\infty$

RKHS $\mathcal{H}_{K_\sigma}$ generated by a Gaussian kernel on X

$\mathcal{H}_{K_\sigma} = \mathcal{H}_{K_\sigma}(\mathbf{R}^n)|_X$ where $\mathcal{H}_{K_\sigma}(\mathbf{R}^n)$ is the RKHS generated by K_σ as a Mercer kernel on $\mathbf{R}^n$:

$$\mathcal{H}_{K_\sigma}(\mathbf{R}^n) = \left\{ f \in L^2(\mathbf{R}^n) : \|f\|_{\mathcal{H}_{K_\sigma}(\mathbf{R}^n)} < \infty \right\}$$

where

$$\|f\|_{\mathcal{H}_{K_\sigma}(\mathbf{R}^n)} = \left(\int_{\mathbf{R}^n} \frac{|\widehat{f}(\xi)|^2}{(\sqrt{2\pi}\sigma)^n e^{-\frac{\sigma^2|\xi|^2}{2}}} d\xi \right)^{1/2}.$$

Thus $\mathcal{H}_{K_\sigma}(\mathbf{R}^n) \subset C^\infty(\mathbf{R}^n)$

Steinwart

If X is a domain with piecewise smooth boundary and $d\rho_X(x) \geq c_0 dx$ for some $c_0 > 0$, then for any $\beta > 0$,

$$\mathcal{D}_\sigma(\lambda) := \inf_{f \in \mathcal{H}_{K_\sigma}} \left\{ \|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda \|f\|^2_{K_\sigma} \right\} = O(\lambda^\beta)$$

implies $f_\rho \in C^\infty(X)$.

Note $\|f - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f) - \mathcal{E}(f_\rho)$ where $\mathcal{E}(f) := \int_Z (f(x) - y)^2 d\rho$.

Denote $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 \approx \mathcal{E}(f)$. Then

$$f_{\mathbf{z},\lambda,\sigma} = \arg \min_{f \in \mathcal{H}_{K_\sigma}} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|^2_{K_\sigma} \right\}.$$

If we define

$$f_{\lambda,\sigma} = \arg \min_{f \in \mathcal{H}_{K_\sigma}} \left\{ \mathcal{E}(f) + \lambda \|f\|^2_{K_\sigma} \right\},$$

then $f_{\mathbf{z},\lambda,\sigma} \approx f_{\lambda,\sigma}$ and the error can be estimated in terms of λ by the theory of uniform convergence over the **compact** function set $B_{M/\sqrt{\lambda}} := \{f \in \mathcal{H}_{K_\sigma} : \|f\|_{K_\sigma} \le M/\sqrt{\lambda}\}$ since $f_{\mathbf{z},\lambda,\sigma} \in B_{M/\sqrt{\lambda}}$. But $\|f_{\lambda,\sigma} - f_\rho\|^2_{L^2_{\rho_X}} = O(\lambda^\beta)$ for any $\beta > 0$ implies $f_\rho \in C^\infty(X)$. So the learning ability of a single Gaussian is weak. One may choose less smooth kernel, but we would like radial basis kernels for manifold learning.

One way to increase the learning ability of Gaussian kernels: let σ depend on m and $\sigma = \sigma(m) \to 0$ as $m \to \infty$.

Steinwart-Scovel, Xiang-Zhou, ...

Another way: allow all possible variances $\sigma \in (0, \infty)$

Regularization Schemes with Flexible Gaussians:

Zhou, Wu-Ying-Zhou, Ying-Zhou, Micchelli-Pontil-Wu-Zhou, ...

$$f_{\mathbf{z},\lambda} := \arg \min_{0<\sigma<\infty} \min_{f \in \mathcal{H}_{K_\sigma}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_{K_\sigma}^2 \right\}.$$

Theorem 2 *(Ying-Zhou, J. Mach. Learning Res. 2007) Let ρ_X be the Lebesgue measure on a domain X in $\mathbb{R}^n$ with minimally smooth boundary. If $f_\rho \in H^s(X)$ for some $s \leq 2$ and $\lambda = m^{-\frac{2s+n}{4(4s+n)}}$, then we have*

$$E_{\mathbf{z}\in Z^m} \left(\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2}^2 \right) = O\left(m^{-\frac{s}{2(4s+n)}} \sqrt{\log m} \right).$$

Major difficulty: is the function set

$$\mathcal{H} = \cup_{0<\sigma<\infty} \left\{ f \in \mathcal{H}_{K_\sigma} : \|f\|_{K_\sigma} \leq R \right\}$$

with $R > 0$ learnable? That is, is this function set a uniform Glivenko-Cantelli class? Its closure is not a compact subset of $C(X)$.

Theory of Uniform Convergence for $\sup_{f\in\mathcal{H}} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)|$.

Given a bounded set $\mathcal{H}$ of functions on X, when do we have

$$\lim_{\ell\to\infty} \sup_{\rho} \mathrm{Prob} \left\{ \sup_{m\geq\ell} \sup_{f\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \int_X f(x) d\rho \right| > \epsilon \right\} = 0, \forall \epsilon > 0?$$

Such a set is called a **uniform Glivenko-Cantelli** (UGC) class.

Characterizations: Vapnik-Chervonenkis, and Alon, Ben-David, Cesa-Bianchi, Haussler (1997)

Our quantitative estimates:

If $V : Y \times \mathbf{R} \to \mathbf{R}_+$ is convex with respect to the second variable, $M = \|V(y,0)\|_{L^\infty_\rho(Z)} < \infty$, and

$$C_R = \sup\{\max\{|V'_-(y,t)|, |V'_+(y,t)|\} : y \in Y, |t| \le R\} < \infty,$$

then we have

$$E_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i)) - \int_Z V(y, f(x))d\rho\right|\right\} \le C' C_R R \frac{\log m}{m^{1/4}} + \frac{2M}{\sqrt{m}},$$

where C' is a constant depending on n.

Ideas: reducing the estimates for $\mathcal{H}$ to a much smaller subset $\mathcal{F} = \{(K_\sigma)_x : x \in X, 0 < \sigma < \infty\}$, then bounding empirical covering numbers. The UGC property follows from the characterization of Dudley-Giné-Zinn.

Improve the learning rates when X is a manifold of dimension d with d much smaller than the dimension n of the underlying Euclidean space.

Approximation by Gaussians on Riemannian manifolds

Let X be a d-dimensional connected compact C^∞ submanifold of $\mathbf{R}^n$ without boundary. The approximation scheme is given by a family of linear operators $\{I_\sigma : C(X) \to C(X)\}_{\sigma>0}$ as

$$\begin{aligned} I_\sigma(f)(x) &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_X K_\sigma(x,y) f(y) dV(y) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_X \exp\left\{-\frac{|x-y|^2}{2\sigma^2}\right\} f(y) dV(y), \quad x \in X, \end{aligned}$$

where V is the Riemannian volume measure of X.

Theorem 3 *(Ye-Zhou, Adv. Comput. Math. 2007) If $f_\rho \in$ Lip(s) with $0 < s \leq 1$, then*

$$\|I_\sigma(f_\rho) - f_\rho\|_{C(X)} \leq C_X \|f_\rho\|_{Lip(s)} \sigma^s \qquad \forall \sigma > 0, \qquad (2)$$

where C_X is a positive constant independent of f_ρ or σ. By taking $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{s+d}{8s+4d}}$, we have

$$E_{\mathbf{z}\in Z^m}\left\{\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}}\right\} = O\left(\left(\frac{\log^2 m}{m}\right)^{\frac{s}{8s+4d}}\right).$$

The index $\frac{s}{8s+4d}$ in Theorem 3 is smaller than $\frac{s}{2(4s+n)}$ in Theorem 2 when the manifold dimension d is much smaller than n.

Classification by Gaussians on Riemannian manifolds

Let $\phi(t) = \max\{1 - t, 0\}$ be the hinge loss for the support vector machine classification. Define

$$f_{\mathbf{z},\lambda} = \arg \min_{\sigma \in (0,+\infty)} \min_{f \in \mathcal{H}_{K_\sigma}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \phi\left(y_i f(x_i)\right) + \lambda \|f\|_{K_\sigma}^2 \right\}.$$

By using $I_\sigma : L^p(X) \to L^p(X)$, we obtain learning rates for binary classification to learn the Bayes rule:

$$f_c(x) = \begin{cases} 1, & \text{if } \rho(y = 1|x) \geq \rho(y = -1|x) \\ -1, & \text{if } \rho(y = 1|x) < \rho(y = -1|x). \end{cases}$$

Here $Y = \{1, -1\}$ represents two classes. The misclassification error is defined as $\mathcal{R}(f) : \text{Prob}\{y \neq f(x)\} \geq \mathcal{R}(f_c)$ for any $f : X \to Y$.

The Sobolev space $H_p^k(X)$ is the completion of $\mathcal{C}^\infty(X)$ with respect to the norm

$$\|f\|_{H_p^k(X)} = \sum_{j=0}^{k} \Big(\int_X |\nabla^j f|^p dV \Big)^{1/p},$$

where $\nabla^j f$ denotes the jth covariant derivative of f.

Theorem 4 *If f_c lies in the interpolation space $(L^1(X), H_1^2(X))_\theta$ for some $0 < \theta \le 1$, then by taking $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{2\theta+d}{12\theta+2d}}$, we have*

$$E_{\mathbf{z}\in Z^m}\Big\{\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)\Big\} \le \widetilde{C}\left(\frac{\log^2 m}{m}\right)^{\frac{\theta}{6\theta+d}},$$

where $\widetilde{C}$ is a constant independent of m.

Ongoing topics:

variable selection

dimensionality reduction

graph Laplacian

diffusion map