http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (6): 1257-1261, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1257.1261

Convergence Rate of Least Square Regressions with Data Dependent Hypothesis

[1]Ye Peixin, [1]Han Yongjie and [2]Duan Liqin
[1]School of Mathematics and LPMC, Nankai University, Tianjin 300071, China
[2]Mathematics and Science College, Shanghai Normal University, Shanghai, 200234, China

Abstract: We estimated the error of least square regression with data dependent hypothesis and coefficient regularization algorithms based on general kernel. When the kernel belongs to some kind of Mercer kernel, under a very mild regularity condition on the regression function, we derive a dimensional-free learning rate $m^{-1/6}$.

Key words: Least square regressions, data dependent hypothesis, coefficient regularization, Mercer kernel, learning rate

INTRODUCTION

Kernel-based least regression square learning is a very popular field in recent years (Vapnik, 1995; Cucker and Smale, 2001, 2002; De Vito *et al*., 2005; Wu *et al*., 2006; Caponnetto and De Vito, 2007; De Mol *et al*., 2009; Gnecco and Sanguineti, 2009; Mairal *et al*., 2010; Schmit, 1907; Koltchinskii 2006; Aronszajn, 1950). Some mathematical foundations of it have been established, (Carmeli *et al*., 2006; Christmann and Steinwart, 2007, 2008; Cucker and Zhou, 2007; Evgeniou *et al*., 2000; Hiriart-Urruty and Lemarechas, 2001; Sheng *et al*., 2012a-c; Sheng and Ye, 2011; Huaisheng and Ye, 2011; Sheng and Xiang, 2011, 2012). In this study, we study some mathematical aspects of least square learning algorithms. We consider some error estimate of least square regression learning with general kernel and coefficient regularization. In particular when the kernel belongs to some kind of Mercer kernel, under a very mild regularity condition on the regression function, we derive a dimensional-free learning rate $m^{-1/6}$.

Let us formulate the problem of learning in a standard way. Let $X\subset\Re^q$, $Y\subset\Re$ be Borel sets and let ρ be a Borel probability measure on $Z = X\times Y$. For function $f: X\to Y$ define the error:

$$E_\rho(f) = \int_Z (y - f(x))^2 d\rho .$$

For each input $x\in X$ and output $y\in Y$, $(f(x)-y)^2$ is the error suffered from the use of f as a model for the process producing y from x. By integrating over $X\times Y$ (w.r.t. ρ, of course) we average out the error over all pairs (x; y). Hence the word "error" for $E_\rho(f)$.

For every $x\in X$, let $\rho(y|x)$ be the conditional (w.r.t. x) probability measure on Y . L*et al*so ρ_x be the marginal probability measure of ρ on X, i.e. the measure on X defined by $\rho_x(S) = \rho(\pi^{-1}(S)$ where π: X is the projection. For every integrable function φ: $X\times Y\to R$ a version of Fubini's Theorem relates ρ, $\rho(y|x)$ and ρ_x as follows:

$$\int_{X\times Y} \varphi(x,y)d\rho = \int_X \left(\int_Y \varphi(x,y)d\rho(y\,|\,x)\right)d\rho_X .$$

This "breaking" of ρ into the measures $\rho(y|x)$ and ρ_x corresponds to looking at $X\times Y$ as a product of an input domain X and an output set Y. In what follows, unless otherwise specified, integrals are to be understood over ρ, $\rho(y|x)$ or ρ_x.

Define f_ρ: $X\to Y$ by:

$$f_\rho(x) = \int_Y y d\rho(y\,|\,x), x\in X .$$

The function f_ρ is called the regression function of ρ. For each $x\in X$, $f_\rho(x)$ is the average of the y coordinate of $\{x\}\times y$ (in topological terms, the average of y on the fiber of x).

It is clear that if:

$$f_\rho \in L_2(\rho_X) = \left\{f : \|f\|_{2,\rho_X} = \left(\int_X |f(x)|^2 \, d\rho_X\right)^{1/2} < +\infty\right\},$$

then it minimizes the error E(f) over all $f\in L_2(\rho_x)$. Thus, in the sense of error E(.) the regression function $f_\rho(x)$ is the best to describe the relation between inputs $x\in X$ and outputs $y\in Y$.

In most cases, the distribution $\rho(x,y)$ is unknown and what one can know is a set of samples $z = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ which are drawn independently and identically distributed according to $\rho(x,y)$. Our goal is to find an

Corresponding Author: Duan Liqin, Mathematics and Science College, Shanghai Normal University, Shanghai, 200234, China

estimator f_z on the base of given data z that approximates f_ρ well with high probability. This is an ill-posed problem and the regularization technique is needed. In many areas of machine learning, the following Tikhonov regularization scheme is commonly used to overcome the ill-posed-ness:

$$f_z^{(H)} := \arg\min_{f\in H}\left\{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2+\lambda\|f\|_H^2\right\} \quad (1)$$

Usually H is taken as a Reproducing Kernel Hilbert Space (RKHS) induced by a Mercer kernel which is continuous, symmetric and positive semi-definite on X×X, (Vapnik, 1995; Cucker and Smale, 2001, 2002; De Vito *et al*., 2005; Wu *et al*., 2006; Caponnetto and De Vito, 2007; De Mol *et al*., 2009; Gnecco and Sanguineti, 2009). The RKHS H_K associated with the kernel K is defined to be the closure of the linear span of the set of functions:

$$\{K_x = K(x,.): x\in X\}$$

with the inner product $\langle K_x, K_y\rangle_{HK} = K(x,y)$. The reproducing property takes the form:

$$\langle F, K_x\rangle_{HK} = f(x), \forall x\in X, f\in H_K. \quad (2)$$

This kind of kernel scheme has been studied due to a lot of literatures, c.f. (Vapnik, 1995; Cucker and Smale, 2001; Cucker and Smale, 2002; De Vito *et al*., 2005; Wu *et al*., 2006; Caponnetto and De Vito, 2007; De Mol *et al*., 2009; Gnecco and Sanguineti, 2009; Mairal *et al*., 2010; Schmit, 1907; Koltchinskii, 2006).

In this study, we consider a different kernel scheme. Let $K:X\times X\to\Re$ be a continuous and bounded function which is called general kernel. For a given data $\bar{Y} := \{y_1, y_2,..., y_m\}\subset X$ the data dependent hypothesis space is given by:

$$H_{K,\bar{Y}} := \left\{f_\alpha(x)=\sum_{j=1}^{m}\alpha_j K(x,y_j): \alpha=(\alpha_1,\alpha_2,\ldots,\alpha_m)\in\Re^m\right\}$$

Every hypothesis function is determined by its coefficients and the penalty is imposed on these coefficients. Then, there comes the general coefficient regularized scheme:

$$\alpha_z = \arg\min_{\alpha\in\Re^m}\left\{\frac{1}{m}\sum_{i=1}^{m}(f_\alpha(x_i)-y_i)^2+\lambda\Omega(\alpha)\right\} \quad (3)$$

where, $\Omega(\alpha)$ is a positive function on $\Re^m$.

Formulation (3) is a data dependent scheme which has been found many applications in the design of support vector machines, micro-array analysis and variable selection (Mairal *et al*., 2010; Schmit, 1907; Koltchinskii, 2006; Aronszajn, 1950). We now study a particular coefficient regularization.

We endow $\Re^m$ with usual inner product, i.e., for any $\alpha = (\alpha_1, \alpha_2,..., \alpha_m)^T$, $b = (b_1, b_2,..., b_m)^T\in\Re^m$, we take:

$$(a,b)_2 = \sum_{i=1}^{m}a_i b_i = a^T b$$

In particular $\|a\|_2^2 = a^T a$.

Set:

$$\Omega(\alpha) = m\|\alpha\|_2^2 = m\sum_{i=1}^{m}|\alpha_i|^2$$

We have the following coefficient regularization with l_2-penalization:

$$\alpha_z := \alpha_{z,\lambda} = \arg\min_{\alpha\in\Re^m}\left\{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - f_\alpha(x_i)\right)^2+\lambda m\|\alpha\|_2^2\right\} \quad (4)$$

Equation 4 is a strict convex optimization problem whose solution may be analyzed with tools from convex analysis (Aronszajn, 1950). Based on this consideration, we shall give the explicit expression for the solution of Eq. 4, with which and a inequality for convex functions show the robustness of the solutions (Lemma 3). Thus, we will use a new approach to estimate the learning rate $\|f_{\alpha z}-f_\rho\|_{2,\rho X}$.

For this purpose, we define the integral regularized risk scheme corresponding to Eq. 4 as:

$$\alpha^{(\rho)} := \alpha_\lambda^{(\rho)} = \arg\min_{\alpha\in\Re^m}\left\{E_\rho(f_\alpha)+\lambda m\|\alpha\|_2^2\right\} \quad (5)$$

Then, we have the following error decomposition:

$$\|f_{\alpha_z}-f_\rho\|_{2,\rho X}\le\|f_{\alpha_z}-f_{\alpha^{(\rho)}}\|_{2,\rho X}+\|f_{\alpha^{(\rho)}}-f_\rho\|_{2,\rho X} \quad (6)$$

where the first term of the right hand-side is called the sample error and the second term is called the approximation error. So the estimate of learning error id reduced to those of sample error and approximation error.

In this study, we assume $|y|<M$ almost surely. So the regression function f_ρ is bounded and square integrable with respect to ρ_x. For the kernel function K, we only assume it is continuous and bounded. We denote:

$$k := \sup_{(x,y)\in X\times X}|K(x,y)| \text{ and } |\rho|_2 := \int_Z y^2 d\rho$$

In study of Gnecco and Sanguineti (2009) the following results have been obtained:

Theorem 1: Let K(x,y) be a general kernel on X×X, α_2 and $\alpha^{(\rho)}$ be defined as in Eq. 4-5, respectively. Then, for any $0<\delta<1$, with confidence 1-δ, there holds:

$$\| f_{\alpha_z} - f_{\alpha^{(\rho)}} \|_{2,\rho_X} \leq \frac{6k^2\sqrt{|\rho|_2}\log\frac{2}{\delta}}{\lambda\sqrt{m}} \quad (7)$$

$$\text{for } m \geq \frac{M^2}{|\rho|_2} \text{ and } \lambda \geq \frac{k^2}{m}$$

Theorem 2: Under the assumption of Theorem 1, for any $0<\delta<1$, with confidence $1<\delta$, there holds:

$$\| f_{\alpha_z} - f_\rho \|_{2,\rho_X} \leq \frac{6k^2\sqrt{|\rho|_2}\log\frac{2}{\delta}}{\lambda\sqrt{m}} + K(f_\rho,\lambda)^{1/2} \quad (8)$$

where:

$$K(f_\rho,\lambda) = \inf_{\alpha\in\mathbb{R}^m}\left(\| f_\alpha - f_\rho \|^2_{2,\rho_X} + \lambda m \| \alpha \|^2_2\right)$$

Define the integral operator corresponding to K(x,y) as:

$$L_K(f,x) = \int_X K(x,t)f(t)d\rho(t), \quad x\in X, f\in L_2(\rho_X)$$

This operator is a compact operator that maps a Hilbert space $L_2(\rho_x)$ to itself.

We assume $f_\rho(x) = L_k(\varphi,x)$ and $\varphi\in L_2(\rho_x)$ which implies f_ρ lies in the range of L_k. Under this assumption, we obtain the following upper estimate for $K(f_\rho,\lambda)$.

Theorem 3 (Sheng *et al.*, 2012a): Let K(x, y) be a general kernel on X×X, $f_\rho(x) = L_k(\varphi,x)$ with $\varphi\in L_2(\rho_x)$. Then:

- There is a discrete set $\Upsilon\subset X$ such that:

$$K(f_\rho,\lambda) \leq \frac{A - \|f\|^2_{2,\rho_X}}{m} + \lambda\|\varphi\|^2_{2,\rho_X}$$

- If $\Upsilon = \{x_1,...,x_m\}\subset X$ is taken from the sample z, then, for any $\delta\in(0,1)$, with confidence 1-δ, there holds:

$$K(f_\rho,\lambda) \leq \frac{1}{\delta}\left(\frac{A - \|f\|^2_{2,\rho_X}}{m} + \lambda\|\varphi\|^2_{2,\rho_X}\right)$$

where:

$$A = \int_X\int_X \varphi(y)^2 K(x,y)^2 d\rho(x)d\rho(y)$$

The above three Theorems are our motivation of our present work. In the next section we provide our main results.

RESULTS

Here, we will show if K(x,y) belongs to some kind of Mercer kernel the convergence rate for the functional $K(f_\rho, \lambda)$ can be derived. Based on this result, we will derive a dimension-free learning rate.

The integral operator:

$$L_K(f,x) = \int_X K(x,t)f(t)d\rho(t), \quad x\in X, f\in L_2(\rho_X)$$

is a compact operator that maps a Hilbert space $L_2(\rho_x)$ to itself. Then, by the Schmidt expansion (Schmit, 1907; Carmeli *et al.*, 2006):

$$K_x(y) := K(x,y) = \sum_{j=0}^{+\infty}\lambda_j\phi_j(x)\phi_j(y), \; x,y\in X \quad (9)$$

where, $\{\lambda_j\}_{j=0}^{+\infty}$ is a non-increasing sequence of eigenvalues of L_K and $\{\phi_j(x)\}_{j=0}^{+\infty}$ forms the corresponding orthonormal eigenfunctions. The convergence is absolute and uniform on X×X. We want to find an approximating sequence of the form:

$$f_{\alpha(f_\rho)}(x) = \sum_{k=1}^{m}\alpha_k(f_\rho)K_{y_k}(x), \quad x\in X$$

where, $\alpha(f_\rho) = (\alpha_1(f_\rho),...., \alpha_m(f_\rho))$ and provide the estimate for:

$$\| f_\rho - f_{\alpha(f_\rho)} \|_{2,\rho_X} + \lambda m \| \alpha(f_\rho) \|_2$$

which can be served as a upper estimate for the functional $K(f_\rho, \lambda)$.

By the Mercer's theorem (Carmeli *et al.*, 2006) we know the eigenvalues $\lambda_i\geq 0$. We assume further that $0<\lambda_i<1$ and the eigenfunctions $\{\phi_j(x)\}_{j=0}^{+\infty}$ forms a complete orthonormal basis of $L_2(\rho_x)$. Moreover, we assume that $|\varphi_i(x)|\leq 1$, for all i and $x\in X$.

Define $\alpha_i(f) = \int_x f(t)\varphi_i(t)d\rho_x(t)$ for $f\in L_2$. Then our main results can be stated as the following theorems:

Theorem 4: If f_ρ satisfies:

$$\sum_{l=0}^{\infty}\frac{|a_l(f_\rho)|}{\lambda_l}<+\infty \tag{10}$$

then, there is an absolute constant C>0 such that:

$$K(f_\rho,\lambda)\le\left(\frac{4k^2C^2(\rho_X(X))h}{m}+\lambda\right)\theta^2 \tag{11}$$

where:

$$\theta=\rho_X(X)\sum_{l=0}^{\infty}\frac{|a_l(f_\rho)|}{\lambda_l}$$

is the VC dimension of the family of real valued functions $\{K_x(t): t\in X\}$.

Theorem 5: If f_ρ satisfies (9) then, for a large probability:

$$\|f_\rho-f_{\alpha_z}\|_{2,\rho_X}\le c\cdot m^{-1/6}$$

To prove Theorem 4, we will exploit the following known result on error bound for sparse approximation. To this end, we first recall the concept of VC dimension. The VC dimension of a family $F=\{F_x\}$ of real valued functions on a set X is the maximum number h of points (t_i) in X that can be separated into two district classes in all 2^h possible ways, by using functions of the form $F_x(t)-\alpha$, where the parameters x and α vary in X and $\Re$, respectively (Vapnik, 1995).

Lemma 1 (Christmann and Steinwart, 2007): Let $X\subset\Re^q$, $H\in L_1(X)$, f be a function on X having the integral representation $f(x)=\int_X K_x(t)H(t)dt$ and h the VC dimension of the family $K_x(t): t\in X\}$. If there exists k>0 such that for all x and t one has $|K_x(t)|\le k$, then, for any m, there exist $y_1, y_2,\ldots, y_m\in X$, $c_1, c_2,\ldots, c_m\in\{-1,1\}$ and an absolute constant C such that:

$$\left\|f(x)-\frac{\|H\|_1}{m}\sum_{i=1}^{m}c_iK_x(y_i)\right\|_{C(X)}\le 2kC\|H\|_1\sqrt{\frac{h}{m}} \tag{12}$$

Now, we are ready to prove Theorem 4.

Proof of theorem 4: The proof of Theorem is based on the Schmidt expansion. In fact, since $0<\lambda_1<1$, (9) implies:

$$\sum_{l=0}^{\infty}|a_l(f_\rho)|<+\infty$$

and hence:

$$f_\rho(x)=\sum_{l=0}^{+\infty}a_l(f_\rho)\phi_l(x)$$

Take:

$$H(y)=\sum_{l=0}^{\infty}\frac{a_l(f_\rho)}{\lambda_l}\phi_l(y)$$

then, by Eq. 9 we know $H\in C(X)$ and:

$$\int_X K_x(y)H(y)d\rho_X(y)=f_\rho(x)$$

By Lemma 1, there are $Y=\{y_1, y_2,\ldots, y_m\}\subset X$ and $c_1, c_2,\ldots, c_m\in\{-1,1\}$ and an absolute constant C>0, such that (12) holds.

Define:

$$\alpha(f_\rho)=\frac{\|H\|_1}{m}(c_1,c_2,\ldots,c_m)$$

Then:

$$\|f_\rho-f_{\alpha(f_\rho)}\|_{2,\rho_X}\le\sqrt{\rho_X(X)}\,\|f_\rho-f_{\alpha(f_\rho)}\|_{C(X)}\le 2kC\sqrt{\rho_X(X)}\,\|H\|_1\sqrt{\frac{h}{m}}$$

and:

$$\|\alpha(f_\rho)\|_2=\frac{\|H\|_1}{\sqrt{m}}\le\frac{\theta}{\sqrt{m}}$$

Therefore:

$$K(f_\rho,\lambda)\le\|f_\rho-f_{\alpha(f)}\|_{2,\rho_X}^2+\lambda m\|\alpha(f_\rho)\|_2^2\le\left(\frac{4k^2C^2\rho_X(X)h}{m}+\lambda\right)\theta^2$$

Thus we complete the proof of Theorem 4.

Theorem 5 can be derived easily from Theorem 2 and 4.

Proof of theorem 5: Combining theorem 2-4, we have:

$$\|f_\rho-f_{\alpha_z}\|_{2,\rho_X}\le\frac{6k^2\sqrt{|\rho|_2}\log\frac{2}{\delta}}{\lambda\sqrt{m}}+\left(\frac{2kC\sqrt{\rho_X(X)h}}{\sqrt{m}}+\sqrt{\lambda}\right)\theta$$

Taking $\lambda=m^{-1/3}$, we get the desired result.

Finally, we want to point out that the similar rate can be derived for classification problem. We will study this problem in the future work.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Foundation of China (Grant No. 11101220, 10971251, 11271199 and 11201104).

REFERENCES

Aronszajn, N., 1950. Theory of reproducing kernels. Trans. Am. Math. Soc., 68: 337-404.

Caponnetto, A. and E. De Vito, 2007. Optimal rates for the regularized least-squares algorithm. Found. Compt. Math., 7: 331-368.

Carmeli, C., E. De Vito and A. Toigo, 2006. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. Anal. Appl., 4: 377-408.

Christmann, A. and I. Steinwart, 2007. Consistency and robustness of kernel based regression methods. Bernoulli, 13: 799-819.

Christmann, A. and I. Steinwart, 2008. Consistency of kernel-based quantile regression. Appl. Stoch. Model. Bus. Ind., 24: 171-183.

Cucker, F. and D.X. Zhou, 2007. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, New York, ISBN: 9780521865593, Pages: 236.

Cucker, F. and S. Smale, 2001. On the mathematical foundations of learning theory. Bull. Am. Math. Soc., 39: 1-49.

Cucker, F. and S. Smale, 2002. Best choices for regularization parameters in learning theory: Onthe bias-variance problem. Found. Compt. Math., 2: 413-428.

De Mol, C., E. De Vito and L. Rosasco, 2009. Elastic-net regularization in learning theory. J. Complexity, 25: 201-230.

De Vito, E., A. Caponnetto and L. Rosasco, 2005. Model selection for regularized least-squares algorithm in learning theory. Found. Compt. Math., 5: 59-85.

Evgeniou, T. M. Pontil and T. Poggio, 2000. Regularization networks and support vector machines. Adv. Comput. Math., 13: 1-50.

Gnecco, G. and M. Sanguineti, 2009. The weight-decay technique in learning from data: An optimization point of view. Comput. Manage. Sci., 6: 53-79.

Hiriart-Urruty, J.B. and C. Lemarechas, 2001. Fundamental of Convex Analysis. Springer, New York, ISBN: 9783540422051, Pages: 259.

Huaisheng, B.A.O. and P.X. Ye, 2011. Error analysis for support vector machine classifiers on unit sphere of Euclidean space. J. Comput. Inform. Syst., 7: 3023-3030.

Koltchinskii, V., 2006. Local Rademacher complexities and oracle inequalities in risk minimization. Ann. Statist., 34: 2593-2656.

Mairal, J., F. Bach, J. Ponce and G. Sapiro, 2010. Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res., 11: 19-60.

Schmit, E., 1907. Zur theorie der linearen und nichtlinearen integralgleichungen. Math. Ann., 63: 433-476.

Sheng, B. and D. Xiang, 2011. The consistency analysis of coefficient regularized classification with convex loss. WSEAS Trans. Math., 10: 291-300.

Sheng, B. and P.X. Ye, 2011. Least square regression learning with data dependent hypothesis and coefficient regularization. J. Comput., 6: 671-675.

Sheng, B. and D. Xiang, 2012. Bound the learning rates with generalized gradients. WSEAS Trans. Signal Process., 8: 1-10.

Sheng, B., P.X. Ye and J.L. Wang, 2012a. Learning rates for least square regressions with coefficient regularization. Acta Math. Sin. English Ser., 28: 2205-2212.

Sheng, B., P.X. Ye and Z.X. Chen, 2012b. The Non-Smooth Analysis Method in Kernel Learning. Science Press, Beijing.

Sheng, B., Z.X. Chen, J.L. Wang and P.X. Ye, 2012c. Learning rates of Tikhonov regularized regressions based on sample dependent RKHS. J. Comput. Anal. Appl., 14: 341-359.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York, USA.

Wu, Q., Y.M. Ying and D.X. Zhou, 2006. Learning rates of least-square regularized regression. Found. Compt. Math., 6: 171-192.