Modern Applied Science; Vol. 8, No. 1; 2014
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Minimization of Negative Log Partial Likelihood Function Using Reproducing Kernel Hilbert Space

Nur'azah Abdul Manaf[1,2], Ibragimov Gafurjan[2] & Mohd. Rizam Abu Bakar[2]

[1] Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, KualaLumpur, Malaysia

[2] Department of Mathematics, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia

Correspondence: Nur'azah Abdul Manaf, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, KualaLumpur, Malaysia. Tel: 60-122-906-329. E-mail:azahmanaf@gmail.com

Received: September 22, 2013 Accepted: October 31, 2013 Online Published: January 22, 2014

doi:10.5539/mas.v8n1p140 URL: http://dx.doi.org/10.5539/mas.v8n1p140

Abstract

Reproducing kernel Hilbert space (RKHS) can be used to estimate values of functions, derivatives and integrals of models. The RKHS kernels are useful in finding the optimizer $f(s)=\sum_i a_i K(s,t)$ of the general Cox regression model. The procedure in the minimization of the negative log partial likelihood function is being demonstrated in this paper. Partial differentiation of the loss function is performed to determine the optimal values of $f(s)$.

Keywords: partial derivatives, reproducing kernel method, Cox regression model

1. Introduction

The theory of reproducing kernel is a powerful instrument in many areas of mathematical research. See, for example, Aronszajn (1950), Hille (1972), Burbea (1976), Wahba (1998), Berlinet et al. (2003), Li et al. (2003). Many researchers had shown that there exist strong connection between the problems of applied sciences, mathematical analysis, and many areas of engineering. Many statistical problems can also be solved and data can be analyzed using models related to reproducing kernel Hilbert space (RKHS). RKHS enables us to estimate a variety of mathematical models. A simple linear kernel was used by Li et al. (2003) in Cox regression models to relate expression profiles of censored cancer data sets. N. Abdul Manaf et al. (2011) generated a new kernel to determine the function $f(s)=\sum_i a_i K(s,s_i)$ of the general Cox model and utilized partial derivatives of the negative log partial likelihood to find the optimal values for the HIV patients survival data.

1.1 Reproducing Kernel Hilbert Space (RKHS)

On a domain *Ts*, let *Hs* be a Hilbert space such that there exists an element $\mu_t \in H_S$ for every $t \in T_s$ and the inner product in H_s is $f(t) = <\mu_t, f>$, for every $f \in H_s$. Let $K(s,t) = <\mu_s, \mu_t>$. Then, K is a positive definite kernel on $T_s \times T_s$, meaning that $\sum_{i,j} a_i a_j K(t_i,t_j) \geq 0$ for every $t_1, t_2, ..., t_n \in T_s$ and K is the reproducing kernel for H_s. Inner product $<K(t,\cdot), K(s,\cdot)> \equiv K(s,t)$ gives the originality of "reproducing kernel".

1.2 Moore-Aronszajn Theorem

The Moore-Aronszajn theorem mentioned by Berlinet and Thomas-Aqnan (2003) states that every kernel K which is symmetric and positive definite on a set T_s defines an incomparable RKHS (Berlinet & Thomas-Aqnan, 2003).

The following process shows the construction of space H_s (Berlinet & Thomas-Aqnan, 2003).

Let $K_s = K(s,s_i)$ for all s in T_s. Let H_0 be the linear span of $\{K_s : s \in T_s\}$. Suppose the dot product on H_0 is defined as

$$\left\langle \sum_{i=1}^{m} a_i K_{s_i}, \sum_{j=1}^{n} b_j K_{y_j} \right\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} \overline{a_i} b_j K(y_j, s_i).$$

The symmetry of K generates the symmetry of the dot product.

If we let H_s be the complete set of H_0 with respect to the above dot product, then H_s is the compilation of functions

$$f(s) = \sum_{i=1}^{\infty} a_i K_{s_i}(s)$$

We observed from the Cauchy-Schwarz inequality that this series converges for every x. We obtain the inner product

$$\langle f, K_s \rangle = \left\langle \sum_{i=1}^{\infty} a_i K_{s_i}, K_s \right\rangle = \sum_{i=1}^{\infty} a_i K(s_i, s) = f(s).$$

from the reproducing kernel properties.

1.3 Kernel Method and Its Application

Manaf et al. (2011) used $K(x,y) = < Cx, Dy >$ to obtain $f(x)$ for the general Cox model

$$\lambda_i(t \mid x_i) = \lambda_0(t)\exp(f(x_i))$$

A regularized formulation of Cox regression, $R_{reg}(f)$ was considered as a problem in terms of RKHS, H_K, and

$$\min R_{reg}(f) = \frac{1}{n}\sum_{i=1}^{n} V(t_i, \delta_i, f(x_i)) + \xi \|f\|_{H_K}^2$$

where $V(t_i, \delta_i, f(x_i))$ is the loss function which depends on $f(x)$ at points $\{f(x_i)\}_{i=1}^{n}$. The norm defined in H_K is denoted by $\|f\|^2$ where $f = b + h$, $h \in H_k$, $b \in \Re$, and $\xi > 0$ is a tuning parameter. For the general Cox model, t_i is the survival time when $\delta_i = 1$, and is the censoring time when $\delta_i = 0$.

The use of negative log partial likelihood function in the regression model

$$\lambda_i(t \mid x_i) = \lambda_0(t)\exp(f(x_i))$$

leads us to find the function $f(x)$ that minimizes

$$R_{reg}(f) = -\frac{1}{n}\sum_{i=1}^{n} \delta_i \left[f(x_i) - \log\left\{ \sum_{j \in R_i} \exp(f(x_i) \right\} \right] + \xi \|f\|_{H_k}^2 ,$$

where $R_i = \{j : t_j \geq t_i, j = 1,2,...,n\}$ is the set of HIV patients who were at risk at t_i. The solution of this problem was given by Kimeldorf and Wahba and is known as the *representer theorem* in which the optimizer function $f(s)$ has the form (Kimeldorf & Wahba, 1971)

$$f(s) = c + \sum_{i=1}^{n} a_i K(s, s_i),$$

where K is the reproducing kernel of H_k. Constant c can be omitted in the solution procedure because it can be absorbed into baseline hazard function.

2. Formulas for Partial Derivatives of Loss Function

Our task is to find the function $f(x) = \sum_{i=1}^{n} a_i K(x, x_i)$ when the optimal values of vector $a = (a_1, a_2, \ldots, a_n)$ are applied. We need to minimize

$$R_{reg}(f) = -\frac{1}{n}\sum_{i=1}^{n} \delta_i \left[f(x_i) - \log\left\{ \sum_{j \in R_i} \exp(f(x_j)) \right\} \right] + \xi \|f\|^2_{H_k} \quad (1)$$

which is equivalent to minimizing of

$$-\frac{1}{n}\sum_{i=1}^{n} \delta_i f(x_i) + \frac{1}{n}\sum_{i=1}^{n} \delta_i \log\left\{ \sum_{j \in R_i} \exp(f(x_j)) \right\} + \xi \|f\|^2_{H_k} \quad (2)$$

where $R_i = \{j : t_j \geq t_i, j = 1,2,...,n\}$ is the set of individuals who were at risk at time t_i.

We can state the negative log-likelihood function as follows:

$$R_{reg}(f) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\left[f(x_i) - \ln\left\{\sum_{j\in R_i}\exp(f(x_j))\right\}\right] + \xi\|f\|_{H_k}^2$$
$$= -\frac{1}{n}\sum_{i=1}^{n}a_i\left(\sum_{j=1}^{n}\delta_j K(x_j,x_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\delta_i \ln\left\{B_{1(i)}\right\} + \xi\sum_{i=1}^{n}a_i\left(\sum_{j=1}^{n}a_j K(x_i,x_j)\right).$$

$R_{reg}(f)$ is minimized by using the Newton-Raphson method. We give an illustrative example for R_i. Let $n = 4$ and $t_1 = 1,\ t_2 = 5,\ t_3 = 6,\ t_4 = 8$. Then

$R_1 = \left\{j : t_j \ge t_1\ ,\ j = 1,2,3,4\right\} = \left\{1,2,3,4\right\}$,

$R_2 = \left\{j : t_j \ge t_2\ ,\ j = 1,2,3,4\right\} = \left\{2,3,4\right\}$,

$R_3 = \left\{j : t_j \ge t_3\ ,\ j = 1, 2,3,\ 4\right\} = \left\{\ 3,4\right\}$,

$R_4 = \left\{j : t_j \ge t_4\ ,\ j = 1,2,3,4\right\} = \{4\}$.

Let $A = -\frac{1}{n}\sum_{j=1}^{n}\delta_j f(x_j)$, the first term of (2). Then

$$A = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_j a_i K(x_j,x_i) = -\frac{1}{n}\sum_{i=1}^{n}a_i\left(\sum_{j=1}^{n}\delta_j K(x_j,x_i)\right) \tag{3}$$

We obtain from (3) that

$$\frac{\partial A}{\partial a_p} = -\frac{1}{n}\sum_{j=1}^{n}\delta_j K(x_j,x_p)\ ,\ p = 1,\dots,n. \tag{4}$$

and

$$\frac{\partial^2 A}{\partial a_p \partial a_q} \equiv 0,\ p = 1,\dots,n. \tag{5}$$

Let the second term of (2) be denoted by

$$B = \frac{1}{n}\sum_{i=1}^{n}\delta_i \ln\left\{\sum_{j\in R_i}\exp(f(x_j))\right\}.$$

Then by definition of $f(x_i)$ we get

$$B = \frac{1}{n}\sum_{i=1}^{n}\delta_i \ln\left\{\sum_{j\in R_i}\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right)\right\} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \ln B_{1(i)} \tag{6}$$

where

$$B_{1(i)} = \sum_{j\in R_i}\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right).$$

Find partial derivatives of B. We have from (6) that

$$\frac{\partial B}{\partial a_p} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \frac{1}{B_{1(i)}}\cdot\frac{\partial B_{1(i)}}{\partial a_p} = \frac{1}{n}\sum_{i=1}^{n}\delta_i \frac{1}{B_{1(i)}}\cdot\sum_{j\in R_i}K(x_j,x_p)\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right),\ p = 1,2,\dots,n.$$

Denote

$$B_{2p(i)} = \frac{\partial B_{1(i)}}{\partial a_p} = \sum_{j\in R_i}K(x_j,x_p)\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right).$$

Then, clearly

$$\frac{\partial B}{\partial a_p}=\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{B_{2p(i)}}{B_{1(i)}}, \tag{7}$$

and so

$$\frac{\partial^2 B}{\partial a_p \partial a_q}=\frac{\partial}{\partial a_q}\left(\frac{\partial B}{\partial a_p}\right)=\frac{\partial}{\partial a_q}\sum_{i=1}^{n}\delta_i\frac{B_{2p(i)}}{B_{1(i)}}=\frac{1}{n}\sum_{i=1}^{n}\delta_i\cdot\frac{B_{1(i)}\cdot\frac{\partial B_{2p(i)}}{\partial a_q}-B_{2p(i)}\cdot\frac{\partial B_{1(i)}}{\partial a_q}}{B^2{}_{1(i)}}$$

This gives

$$\frac{\partial^2 B}{\partial a_p \partial a_q}=\frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\frac{1}{B_{1(i)}}\sum_{j\in R_i}K(x_j,x_p)K(x_j,x_q)\exp\left(\sum_{k=1}^{n}a_kK(x_j,x_k)\right)-\frac{B_{2p(i)}B_{2q(i)}}{B_{1(i)}^2}\right),\ p,q=1,2,\ldots,n\ , \tag{8}$$

where $B_{2q(i)}=\dfrac{\partial B_{1(i)}}{\partial a_q}$.

Let $C=\xi\|f\|^2_{H_k}$. Then by definition of f we obtain

$$C=\xi\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jK(x_i,x_j)=\xi\sum_{i=1}^{n}a_i\left(\sum_{j=1}^{n}a_jK(x_i,x_j)\right)=\xi\left(\sum_{\substack{i=1\\ i\neq p}}^{n}a_i\left(\sum_{j=1}^{n}a_jK(x_i,x_j)\right)+a_p\sum_{j=1}^{n}a_jK(x_p,x_j)\right)$$

Hence

$$\frac{\partial C}{\partial a_p}=\xi\left(\sum_{i=1,i\neq p}^{n}a_iK(x_i,x_p)+\sum_{j=1}^{n}a_jK(x_p,x_j)+a_pK(x_p,x_p)\right).$$

This can be written as follows:

$$\frac{\partial C}{\partial a_p}=\xi\left(\sum_{\substack{i=1\\ i\neq p}}^{n}a_iK(x_i,x_p)+\sum_{j=1}^{n}a_jK(x_p,x_j)\right) \tag{9}$$

and hence

$$\frac{\partial^2 C}{\partial a_p\partial a_q}=\xi\left(K(x_q,x_p)+K(x_p,x_q)\right)=2\xi K(x_p,x_q) \tag{10}$$

since $K(x_p,x_q)=K(x_q,x_p)$.

3. Result and Discussion

Let $K(x_j,x_i)=<x_j,x_i>$, $f(x_j)=\sum_{i=1}^{n}a_iK(x_j,x_i)$, x_i = (gender, age, race), and $\lambda_i(t|x_i)=\lambda_0(t)\exp(f(x_i))$, which is the hazard function for Cox model where t_i is lifetime of the i th patient. Then we have to include the following equations to use the Newton-Raphson method:

$$R=-\frac{1}{n}\sum_{i=1}^{n}a_i\left(\sum_{j=1}^{n}\delta_jK(x_j,x_i)\right)+\frac{1}{n}\sum_{i=1}^{n}\delta_i\ln\left\{B_{1(i)}\right\}+\xi\left(\sum_{i=1,i\neq p}^{n}a_i\left(\sum_{j=1}^{n}a_jK(x_i,x_j)\right)+a_p\sum_{j=1}^{n}a_jK(x_p,x_j)\right).$$

Using the partial derivatives obtained above, we have

$$\frac{\partial R}{\partial a_p}=-\frac{1}{n}\sum_{j=1}^{n}\delta_jK(x_j,x_p)+\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{B_{2p(i)}}{B_{1(i)}}+\xi\left(\sum_{i=1}^{n}a_iK(x_i,x_p)+\sum_{j=1}^{n}a_jK(x_p,x_j)\right),$$

and

$$\frac{\partial^2 R}{\partial a_p \partial a_q} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\frac{1}{B_{1(i)}}\sum_{j\in R_i}K(x_j,x_p)K(x_j,x_q)\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right) - \frac{B_{2p(i)}B_{2q(i)}}{B^2_{1(i)}}\right) + 2\xi K(x_p,x_q),$$

where

$$B_{1(i)} = \sum_{j\in R_i}\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right),\quad R_i = \left\{j : t_j \ge t_i\ ,\ j = 1,2,...,n\right\},$$

$$B_{2p(i)} = \sum_{j\in R_i}K(x_j,x_p)\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right),\ \ p = 1, 2,\dots,n\ ,$$

$$B_{2q(i)} = \sum_{j\in R_i}K(x_j,x_q)\exp\left(\sum_{k=1}^{n}a_k K(x_j,x_k)\right),\ \ q = 1,2,\dots,n.$$

We used the kernel $K(x,y) = < Cx, Dy >$, where C and D are diagonal matrices (Manaf et al., 2011). We had verified the positive definiteness and symmetrical properties of kernel *K*. Our research and observations show that the greater the values of function $f(s) = \sum_{i=1}^{n} a_i K(s,s_i)$, the less chance of survival among the HIV patients chosen in random. This is shown in the following Figure 1.

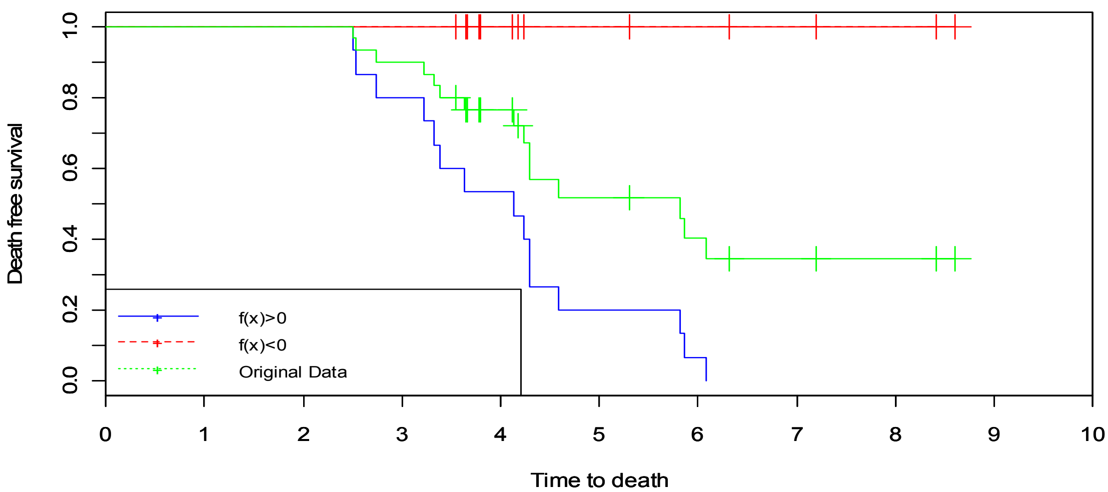

Figure 1. Survival of HIV patients

Using the Newton-Raphson method the optimal values of a_i is obtained by setting the derivatives with respect to a in R to zero. We compared the result with the Gaussian Radial Basis function kernel, $K(x,x_i) = \exp\left(-\frac{1}{2}\left\langle x - x_i, x - x_i\right\rangle\right)$. Our result is shown in Table 1.

Table 1. Results of Two Different Kernels

	$K(x,x_i)=\langle Ax,Bx_i\rangle$			$K(x,x_i)=\exp\left\{-\frac{1}{2}\langle x_i-x,x_i-x\rangle\right\}$		
i	a_i	$f(x_i)$	$\exp(f(x_i))$	a_i	$f(x_i)$	$\exp(f(x_i))$
1	-6.571328e-03	1.0803817	2.9458038	0.129739937	0.18785887	1.2066632
2	3.818778e-04	1.1058087	3.0216670	0.173914696	0.14436110	1.1553012
3	-6.358875e-04	1.3355435	3.8020618	0.218131636	0.23427598	1.2639933
4	-1.701724e-03	1.5978747	4.9425169	0.116957073	0.15375643	1.1662068
5	-3.052682e-03	1.1776466	3.2467242	0.136644639	0.18991145	1.2091425
6	3.269568e-03	1.0421580	2.8353290	0.079175542	0.15908422	1.1724367
7	4.034653e-03	0.9161683	2.4996940	0.116962975	0.16858607	1.1836301
8	2.396903e-03	0.7243128	2.0633128	0.081691220	0.25496977	1.2904226
9	2.878405e-03	0.6907156	1.9951427	0.160807013	0.25418269	1.2894073
10	7.937093e-04	0.4774516	1.6119613	0.174749874	0.26432793	1.3025553
11	2.369007e-03	0.6303122	1.8781969	0.177146810	0.24826214	1.2817959
12	2.997758e-03	0.6639095	1.9423712	0.139188313	0.24145605	1.2731015
13	-1.544913e-03	0.4408361	1.5540060	0.005409531	0.18045094	1.1977574
14	7.542037e-03	0.3090386	1.3621150	0.175677868	0.15535025	1.1680670
15	-3.121462e-04	0.4794866	1.6152450	0.138615666	0.19745615	1.2182996
16	-8.216203e-04	-1.4328849	0.2386195	-0.059706201	-0.22576971	0.7979018
17	-1.165802e-02	-1.2574178	0.2843874	-0.176235680	-0.33967263	0.7120034
18	-8.956521e-03	-0.9429187	0.3894894	-0.123189140	-0.12618218	0.8814543
19	-5.895857e-05	-0.1223129	0.8848715	-0.152238094	-0.04239335	0.9584927
20	-2.303194e-03	-0.8080547	0.4457243	-0.122136320	-0.21545565	0.8061740
21	1.335727e-03	-0.4900623	0.6125882	-0.057326269	-0.17656795	0.8381418
22	-8.834102e-04	-0.2413472	0.7855688	-0.398832658	-0.43254441	0.6488560
23	-3.405090e-03	-1.1368092	0.3208411	-0.403182560	-0.47979802	0.6189084
24	-1.861149e-03	-1.0806055	0.3393900	-0.127591108	-0.32483308	0.7226480
25	2.185279e-04	-0.4989532	0.6071659	-0.045261685	-0.08283722	0.9205010
26	1.289077e-03	-1.1368092	0.3208411	-0.024314446	-0.47979802	0.6189084
27	-4.513436e-03	-0.2686450	0.7644146	-0.110502264	-0.44325517	0.6419434
28	2.688869e-02	-0.5399496	0.5827777	-0.129127364	-0.42241788	0.6554601
29	1.314224e-03	-0.7378419	0.4781447	-0.025176529	-0.07878944	0.9242345
30	4.379507e-03	-0.4496570	0.6378469	-0.080555841	-0.43651604	0.6462841

4. Conclusion

Several other kernels can be generated and then applied to different sets of data. It is important that we are able to verify that the kernels fulfill all the rules and theories of RKHS. The derivatives used in this RKHS method are applicable to all kernels used to determine $f(x)$ in the Cox hazard function models. It should be noted that RKHS kernels can be used in models of several research areas such as in business, engineering and medical sciences because of the connection between data distributions and kernels. In fact, more researches can be performed to show that RKHS method can solve many other problems that involve mathematics and statistics.

References

Aronszajn, N. (1950). Theory of Reproducing Kernels. *Trans. American. Mathematical Soc., 686*, 337-404. http://dx.doi.org/10.1090/S0002-9947-1950-0051437-7

Berlinet, A., & Thomas-Aqnan, C. (2003). Reproducing Kernel Hilbert Spaces in Probability and Statistics. *Kluwer Academic*, U.K.

Burbea, J. (1976). Total Positivity of Certain Reproducing Kernels. *Pacific Journal of Mathematics, 67*(1), 101-130. http://dx.doi.org/10.2140/pjm.1976.67.101

Hille, E. (1972). Introduction to General Theory of Reproducing Kernels. *Rocky Mountain Journal of Mathematics, 2*(3), 321-367. http://dx.doi.org/10.1216/RMJ-1972-2-3-321

Kimeldorf, G., & Wahba (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications, 33*(1), 82-95. http://dx.doi.org/10.1016/0022-247X(71)90184-3

Li, H., & Luan Y. (2003). Kernel Cox Regression Models for Linking Gene Expression Profiles to Censored Survival Data. *Pacific Symposium on Bio-computing, 8*, 65-76.

Manaf, N. A. A., Gafurjan, I., Bakar, A., Rizam, M., & Aljawadi, B. (2011). Exploring Reproducing Kernel Hilbert Space and its Application to Survival Data. *Australian Journal of Basic and Applied Sciences, 5*(11), 1268-1274.

Wahba, G. (1990). SIAM, Philadelphia.

Wahba, G. (1998). Technical Report, UW Madison.

Copyrights