Research Article

A Logical Framework of the Evidence Function Approximation Associated with Relevance Vector Machine

Dawei Zou, Liang Tong, Jicheng Wang, Shoufang Fan, and Jinxing Ji

College of Information Engineering, Suihua University, Suihua 152061, China

Correspondence should be addressed to Dawei Zou; 64462929@qq.com

Received 19 November 2019; Revised 15 January 2020; Accepted 23 January 2020; Published 28 March 2020

Academic Editor: Francesc Pozo

Copyright © 2020 Dawei Zou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Relevance vector machine (RVM) is a Bayesian sparse kernel method for regression in statistical learning theory, which can avoid principal limitations of the support vector machine (SVM) and result in faster performance on test data while maintaining comparable generalization error. In this paper, we develop a logical framework of the evidence function approximation associated with RVM based on Taylor expansion instead of traditional technology called “completing the square.” While constructing the term of completing the square, we have to find the term of completing the square by making use of some skill, which in practice increases the difficulty in dealing with the evidence function approximation associated with RVM. The logical framework in this paper based on Taylor expansion shows some advantages compared with the conventional method of completing the square, which is easier to be enforced due to the clear logical framework and avoids the difficulty in looking for the term of completing the square intentionally. From symmetry of covariance in a multivariate Gaussian distribution and algebraic knowledge, we derive approximation and maximization of the evidence function associated with RVM, which is consistent with the previous result using the novel logical framework. Finally, we derive the EM algorithm for RVM, which is also consistent with the previous result except that we use the precision matrix as the covariance.

1. Introduction

When fitting polynomial curve to a given training data, we encounter the problem of choosing the order of the polynomial, which make us discuss an important concept called model selection. Model selection is an intractable problem in machine learning field since we have to face all kinds of complexity of the model according to the complexity of the problem being solved. We try our best to look for suitable approaches to fix the problem. For example, to avoid overfitting problem, we add regularization term [1, 2] to the sum of the squares of the errors between the prediction for each data point and the corresponding target values or partition the available data into a training set which determines the coefficient and a separate validation set, used to optimize the model complexity, also called a hold-out set. However, this is too wasteful of valuable training data and we have to look for more sophisticated methods. Model selection from a Bayesian perspective is an alternative technology [3] whose merit is that the overfitting problem with maximum likelihood can be avoided by marginalizing (summing for discrete case or integrating for continuous case) over the model parameters compared with making point estimates and models can be compared directly on the training data without the need for a validation set, which allows all available data to be used to train, avoids a lot of training operations for each model associated with cross-validation, and multiple complexity parameters can be determined simultaneously as part of the training process.

Bayesian model comparison simply involves the use of probabilities to depict uncertainty in the model choice by a consistent application of the sum and product rules of probability [4]. The core of model selection is to evaluate evidence function, assuming that all models are given equal prior probability. Relevance vector machine (RVM) [5, 6] is a Bayesian sparse kernel method for classification and regression that has many characteristics of the support vector machine (SVM) while avoiding its main limitation. Relevance vector machine (RVM) generally leads to much sparser models than support vector
machine (SVM), which can result in faster performance, which can be applied in automatic emulation and get better performance [7]. In [5], the evidence function approximation associated with relevance vector machine needs some skill to compute and makes use of the logical framework called “completing the square” [8]. However, the method of completing the square has difficulty in looking for the term of completing the square, especially in complex problems in the machine learning field. The paper will develop a novel logical framework based on Taylor expansion, which is different from the classic technique of completing the square over parameter of the model for evaluating the evidence function associated with relevance vector machine shortly. The method presented in this paper is convenient to evaluate the evidence function with relevance vector machine (RVM).

2. Preliminary Mathematical Knowledge

Lemma 1 (Woodbury identity). \((A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}\).

Lemma 2. If \(A\) and \(B\) are matrices of size \(N \times M\), then
\[
\begin{vmatrix} I_N + AB \end{vmatrix} = \begin{vmatrix} I_M + A^TB \end{vmatrix}.
\]

Specially, \(|I_N + ab^T| = |1 + a^Tb|\), where \(a\) and \(b\) are \(N\)-dimensional column vectors.

Lemma 3. \((\partial / \partial x) \ln |A| = \text{Tr} (A^{-1} (\partial A / \partial x))\).

Lemma 4. \((\partial / \partial x) (A^{-1}) = -A^{-1} (\partial A / \partial x) A^{-1}\).

Lemma 5. If \(x\) and \(y\) are \(n\)-dimensional column vectors, then \(\text{Tr} (xy^T) = x^Ty\), where \(\text{Tr} (\cdot)\) denotes the trace of a matrix. The proofs of the abovementioned four lemmas refer to [9–11].

Definition 1. The operator \((h^T \nabla)^k f\) is defined by \((h^T \nabla)^0 f = f\) and \((h^T \nabla)^k f = (h^T \nabla)(h^T \nabla)^{k-1} f\) for \(k \in \mathbb{N}\).

These symbolic powers act as operators on a function of \(n\) variables, where \(V\) is the gradient operator.

For example, we consider a two-dimensional case. Let \(h = (h_1, h_2)^T\) and \(\nabla = (\partial / \partial x_1, \partial / \partial x_2)\), we have \((h^T \nabla) = h_1 (\partial / \partial x_1) + h_2 (\partial / \partial x_2)\) and \((h^T \nabla)^2 f = h_1 (\partial f / \partial x_1) + h_2 (\partial f / \partial x_2)\). After the operator \((h^T \nabla)^2 f\) acts on \(x\), we obtain
\[
((h^T \nabla)^2 f) (x) = h_1 \frac{\partial f (x)}{\partial x_1} + h_2 \frac{\partial f (x)}{\partial x_2} = h^T \nabla^2 f (x).
\]

We also note that
\[
(h^T \nabla)^2 f = (h^T \nabla) \left( h_1 \frac{\partial f (x)}{\partial x_1} + h_2 \frac{\partial f (x)}{\partial x_2} \right) = h_1 \frac{\partial (h^T \nabla f (x))}{\partial x_1} + h_2 \frac{\partial (h^T \nabla f (x))}{\partial x_2} = h^T \nabla^2 f + 2 h_1 h_2 \frac{\partial^2 f (x)}{\partial x_1 \partial x_2} + h_2 \frac{\partial^2 f (x)}{\partial x_2^2}.
\]

The operator \((h^T \nabla)^2 f\) also acts on \(x\), and we obtain
\[
((h^T \nabla)^2 f) (x) = h_1 \frac{\partial^2 f (x)}{\partial x_1^2} + 2 h_1 h_2 \frac{\partial^2 f (x)}{\partial x_1 \partial x_2} + h_2 \frac{\partial^2 f (x)}{\partial x_2^2}.
\]

Furthermore, we obtain
\[
((h^T \nabla)^2 f) (x) = h^T H h,
\]
where \(h = (h_1, h_2)^T\) and \(H = \nabla^2 f (x) = (\nabla f (x)/\partial x_i \partial x_j)\), which is called the Hessian matrix.

We use \((h^T \nabla)^2 f\) = \(h_1 (\partial f / \partial x_1) + h_2 (\partial f / \partial x_2)\) again and obtain
\[
(h^T \nabla)^3 f = (h^T \nabla) \left( h_1 \frac{\partial^2 f (x)}{\partial x_1^2} + 2 h_1 h_2 \frac{\partial^2 f (x)}{\partial x_1 \partial x_2} + h_2 \frac{\partial^2 f (x)}{\partial x_2^2} \right) + h_1 \frac{\partial^3 f (x)}{\partial x_1^3}.
\]

From (6), we obtain
\[
(h^T \nabla)^3 f = h_1 \frac{\partial^3 f (x)}{\partial x_1^3} + 3 h_1 h_2 \frac{\partial^3 f (x)}{\partial x_1^2 \partial x_2} + 3 h_1 h_2 \frac{\partial^3 f (x)}{\partial x_1 \partial x_2^2} + 3 h_2 \frac{\partial^3 f (x)}{\partial x_2^3}.
\]

By making use of the exerting operator \((h^T \nabla)^3 f\) on variable \(x\), we obtain
\[
((h^T \nabla)^3 f) (x) = h_1 \frac{\partial^3 f (x)}{\partial x_1^3} + 3 h_1 h_2 \frac{\partial^3 f (x)}{\partial x_1^2 \partial x_2} + 3 h_1 h_2 \frac{\partial^3 f (x)}{\partial x_1 \partial x_2^2} + 3 h_2 \frac{\partial^3 f (x)}{\partial x_2^3}.
\]

From (8), we can draw a conclusion that \((h^T \nabla)^3 f (x)\) includes all third-order partial derivative of \(f (x)\) to variables.

Lemma 6 (Taylor’s formula). Let \(f : B (x, r) \rightarrow \mathbb{R}\) be a function that belongs to the class \(C (B (x, r))\), where \(B (x, r) \subseteq \mathbb{R}^n\). If \(h \in \mathbb{R}^n\) is such that \(|h| < r\), then there exists \(\theta \in (0, 1)\) such that
\[
f (x + h) = \sum_{m=0}^{n-1} \frac{1}{m!} ((h^T \nabla)^m f) (x) + \frac{1}{n!} ((h^T \nabla)^n f) (x + \theta h).
\]

The proof of Lemma 6 is found in [12].

Corollary 1. If \(f : B (w, r) \rightarrow \mathbb{R}\) is a quadratic form defined by
\[
f (w) = w^T A w + w^T b + c,
\]
where \(A\) is an \(n \times n\) inverse and symmetric matrix, \(b\) and \(w = (w_1, w_2, \ldots, w_n)^T\) are \(n\)-dimensional column vectors, and \(c\) is a constant. Then, the Taylor expansion of \(f (w)\) at the stationary point \(w_0 = (1/2) A^{-1} b\) is given by
\[
f (w) = f (w_0) + \frac{1}{2} (w - w_0)^T H (w - w_0),
\]
where the element of Hessian matrix $H$ is given by $(\partial^2 f / \partial w_i \partial w_j)$.

Proof. The Taylor formula (9) implies

$$f(w) = f(w_0) + ((h^T \nabla f)(w_0) + \frac{1}{2!} ((h^T \nabla f)^2)(w_0)$$

$$+ \frac{1}{3!} ((h^T \nabla f)^3)(w_0 + \theta h),$$

(12)

where $h = w - w_0$.

For obtaining the stationary point for $f(w)$, setting $\nabla f(w) = 2Aw + b = 0$, we obtain the stationary point $w_0 = (1/2)A^{-1}b$ for $f(w)$. Since $f(w)$ is a quadratic function of $w$, we can derive $((h^T \nabla f)^3)(x) = 0$ at an arbitrary variable $x$. From the abovementioned operations, we obtain (11).

The equality of (11) plays a central role in our discussion because it is clearly different from conventional technology called “completing the square.” We will build up a novel logical framework and discuss its merit compared with the traditional method when dealing with evaluating the evidence function with relevance vector machine (RVM) shortly.

3. Evaluation and Maximization of the Evidence Function with RVM

3.1. Linear Basis Function Models and Bayesian Linear Regression. The simplest linear model for regression is the one that includes a linear combination of the input variables:

$$y(x, w) = w_0 + w_1 x_1 + \cdots + w_D x_D.$$

(13)

However, a linear combination of the input variables imposes significant limitations when we use the model. So, we extend the model by using linear combinations of fixed nonlinear functions of the input variables, which take the form

$$y(x, w) = w_0 + \sum_{i=1}^{M-1} w_i \phi_i(x),$$

(14)

where $\phi_i(x)$ is defined as basis function and $w_0$ is called a bias parameter. For convenience, we define an additional dummy “basis function” $\phi_0(x) = 1$ and rewrite

$$y(x, w) = \sum_{i=0}^{M-1} w_i \phi_i(x) = w^T \phi(x),$$

(15)

where $w = (w_0, \ldots, w_{M-1})^T$ and $\phi(x) = (\phi_0(x), \ldots, \phi_{M-1}(x))^T$.

The target variable $t$ is given by a deterministic function $y(x, w)$ with additive Gaussian noise such that

$$t = y(x, w) + \epsilon,$$

(16)

where $\epsilon$ is a zero mean Gaussian random variable with precision (inverse variance) $\beta$. So, we obtain

$$p(t | x, w, \beta) = N(t \mid y(x, w), \beta^{-1}).$$

(17)

Now, we consider a data set of inputs $X = [x_1, \ldots, x_N]$ with corresponding target value $t = [t_1, \ldots, t_N]$. From the assumption that these data points are drawn independently from the Gaussian distribution (17), we obtain the following likelihood function:

$$p(t \mid X, w, \beta) = \prod_{n=1}^{N} N(t_n \mid w^T \phi(x_n), \beta^{-1}).$$

(18)

Since (18) is the exponential of a quadratic function of $w$, we can define a corresponding conjugate prior given by

$$p(w) = N(w \mid m_0, S_0).$$

(19)

From (18) and (19), we obtain the posterior distribution in the form

$$p(w \mid t, X, \beta) = N(w \mid m_N, S_N),$$

(20)

where

$$m_N = S_N^{-1}(S_0^{-1}m_0 + \beta \Phi^T t),$$

(21)

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi,$$

(22)

where $\Phi$ is an $N \times M$ matrix, called the design matrix, whose elements are given by $\Phi_{ij} = \phi_j(x_i)$ so that

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \cdots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_N) & \cdots & \phi_M(x_N) \end{pmatrix}.$$  

(23)

The proofs of (21) and (22) in detail refer to Appendix A.

3.2. Bayesian Model Comparison and Evidence Approximation. Assume that we face a set of $L$ models $\{M_i\}$, where $i = 1, \ldots, L$. Our task is to compare them and choose an optimal model based on some standard from the Bayesian perspective, which can avoid the overfitting phenomenon associated with maximum likelihood. The uncertainty is expressed through a prior probability distribution $p(M_i)$. Given a training set, we suppose that all models are given the equal prior probability, which is feasible because there should be no preference for $L$ given models in practice. We then wish to evaluate the posterior distribution:

$$p(M_i | D) \propto p(M_i) p(D | M_i),$$

(24)

where $p(M_i | D)$ is defined as model evidence, which is also called the marginal likelihood since it can be viewed as a likelihood function over the space of models, in which the parameters have been marginalized out.

Now, we introduce prior distribution over the hyperparameters $\alpha$ and $\beta$ to complete a fully Bayesian treatment of the linear basis function model and we can obtain prediction by marginalizing with respect to the hyperparameters $\alpha$ and $\beta$ and the parameter $w$. However, the complete marginalization over all of these variables is analytically intractable. We then discuss an estimation, in which we determine the hyperparameters $\alpha$
and $\beta$ by maximizing the marginal likelihood function obtained by first integrating over the parameters $w$. The framework is viewed in the statistics field as empirical Bayes [13, 14], or type-2 maximum likelihood [15], or generalized likelihood [16], and is also called the evidence approximation in the machine learning literature [8, 17, 18].

Here, if we introduce hyperprior over the $\alpha$ and $\beta$, then we obtain predictive distribution by marginalizing over the hyperparameters $\alpha$ and $\beta$ and the parameter $w$ as follows:

$$
p(t|t) = \int \int \int p(t|x,w,\beta)p(w|t,\alpha,\beta)p(\alpha,\beta|t)dw d\alpha d\beta,
$$

(25)

where $p(t|x,w,\beta)$ is defined by (17) and $p(w|t,\alpha,\beta)$ is given by (20) with $m_N$ and $S_N$ defined by (21) and (22).

If the posterior distribution $p(\alpha,\beta|t)$ is sharply peaked around $\tilde{\alpha}$ and $\tilde{\beta}$, we can obtain the predictive distribution simply by marginalizing over $w$ in which $\alpha$ and $\beta$ are set to the value $\tilde{\alpha}$ and $\tilde{\beta}$:

$$
p(t|t) = p(t|t,\tilde{\alpha},\tilde{\beta}) = \int p(t|x,w,\beta)p(w|t,\tilde{\alpha},\tilde{\beta})dw.
$$

(26)

The posterior distribution for $\alpha$ and $\beta$ is given by

$$p(\alpha,\beta|t) \propto p(t|\alpha,\beta)p(\alpha,\beta).
$$

(27)

We can obtain the values $\tilde{\alpha}$ and $\tilde{\beta}$ by maximizing the marginal likelihood function $p(t|\alpha,\beta)$ in the evidence framework if the prior is relatively flat. We continue evaluating the marginal likelihood for the linear basis function and obtain its maxima which will allow us to determine the values of these hyperparameters from the training data set only.

### 3.3. Evaluation of the Evidence Function with RVM

Relevance vector machine (RVM) is a Bayesian sparse kernel method for regression. The relevance vector machine for regression is a linear model of the form given by (15). The following likelihood function is given by (18).

The weight prior takes the form

$$p(w|\alpha) = \prod_{n=1}^{N} N\left(w_i|0,\alpha_i^{-1}\right),
$$

(28)

where $\alpha = (\alpha_1, \cdots, \alpha_M)^T$ and $\alpha_i$ is hyperparameter and represents the precision of the corresponding parameter $w_i$.

By making use of (18), we obtain

$$p(t|X,w,\beta) = \prod_{n=1}^{N} N\left(t_n|w^T \phi(x_n),\beta^{-1}\right)
$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2} \sum_{i=1}^{N} \left(t_n - w^T \phi(x_n)\right)^2\right)
$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2} \|t - \phi w\|^2\right).
$$

(29)

Similarly, by making use of (28), we obtain

$$p(w|\alpha) = \prod_{n=1}^{N} N\left(w_i|0,\alpha_i^{-1}\right)
$$

$$= \frac{\sqrt{\prod_{n=1}^{M} \alpha_n}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2} \sum_{n=1}^{M} \alpha_n w_i^2\right)
$$

$$= |A|^{1/2} \left(\frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2} w^T A w\right)\right).
$$

(30)

We obtain the posterior distribution for the weights, which is also Gaussian and takes the form

$$p(w|t,X,\alpha,\beta) = N\left(w|m,\Sigma\right),
$$

(31)

where the covariance and mean are given by

$$m = \beta \Sigma \Phi^T t,
$$

(32)

$$\Sigma = \left(A + \beta \Phi^T \Phi\right)^{-1},
$$

(33)

where $A = \text{diag}(\alpha_i)$. The proofs of (32) and (33) in detail refer to Appendix B.

We can use evidence approximation, which is also known as type-2 maximum likelihood, to determine the values of hyperparameters $\alpha$ and $\beta$. We first obtain the marginal likelihood function by integrating out the weight parameters:

$$p(t|X,\alpha,\beta) = \int p(t|X,w,\beta)p(w|\alpha)dw,
$$

(34)

and then maximize the marginal likelihood function.

We substitute (32) and (33) into (34) and then obtain the following result:

$$p(t|X,\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} |A|^{1/2} \left(\frac{1}{(2\pi)^{M/2}} \int \exp\left(-E(w)\right)dw\right),
$$

(35)

where we have defined

$$E(w) = \frac{\beta}{2} \|t - \phi w\|^2 + \frac{1}{2} w^T A w
$$

$$= \frac{\beta}{2} (t - \phi w)^T (t - \phi w) + \frac{1}{2} w^T A w
$$

$$= \frac{\beta}{2} (t^T - 2w^T \phi^T t + w^T \phi^T \phi w) + \frac{1}{2} w^T A w
$$

(36)

For the stationary point of $E(w)$, we first derive the gradient of $\nabla E(w)$ by using Corollary 1:

$$\nabla E(w) = (\beta \phi^T \phi + A)w - \beta \phi^T t,
$$

(37)

$$\nabla \nabla E(w) = A + \beta \phi^T \phi = \Sigma^{-1}.
$$

(38)

Let

$$\nabla E(w) = (\beta \phi^T \phi + A)w - \beta \phi^T t = 0.
$$

(39)

From (39), we obtain
\[ w = \beta (\Phi^T \Phi + A)^{-1} \Phi^T t = \beta \Sigma \Phi^T t = m. \]  \hspace{1cm} (40)

From (11), (38), and (40), it follows that
\[ E(w) = E(m) + \frac{1}{2} (w - m)^T \Sigma^{-1} (w - m), \]  \hspace{1cm} (41)

We substitute (41) into (35) and then get the following form:

\[
p(t \mid X, \alpha, \beta) = \left( \frac{\beta}{2\pi} \right)^{N/2} \left[ \frac{\beta}{2\pi} \right]^{1/2} \left( \frac{\beta}{2\pi} \right)^{M/2} \exp \left( -\frac{1}{2} (w - m)^T \Sigma^{-1} (w - m) \right) \]  \hspace{1cm} (43)

By making use of (48) and (49), we obtain
\[
p(t \mid X, \alpha, \beta) = \left( \frac{\beta}{2\pi} \right)^{N/2} \left[ \frac{\beta}{2\pi} \right]^{1/2} \left( \frac{\beta}{2\pi} \right)^{M/2} \exp \left( -\frac{1}{2} (\beta^3 I_N + \Phi A^{-1} \Phi^T)^{-1} t^T (\beta^3 I_N + \Phi A^{-1} \Phi^T)^{-1} t \right) \]  \hspace{1cm} (50)

where we have used the following result according to Lemma 2 and (33):
\[
|\beta^3 I_N + \Phi A^{-1} \Phi^T| = \frac{1}{\beta^N} |\beta^N | I_N + \beta \Phi A^{-1} \Phi^T| \]  \hspace{1cm} (51)

where we make use of (32) and (47).

According to Lemma 1, we obtain
\[
(\beta I_N - \beta^2 \Phi \Sigma \Phi^T)^{-1} = (\beta I_N - \beta \Phi (A + \beta \Phi^T \Phi)^{-1} \beta \Phi^T)^{-1}
\]
\[= \beta^{-1} I_N + \Phi A^{-1} \Phi^T. \]  \hspace{1cm} (49)

From (50), we obtain
\[ p(t \mid X, \alpha, \beta) = N(t \mid 0, C), \] 
(52)
together with \( C = \beta^{-1}I + \Phi A^{-1}\Phi^T \).

From (43), we can then write the log of the marginal likelihood in the form:

\[ \ln p(t \mid X, \alpha, \beta) = \frac{N}{2} \ln \beta - E(m) \]
\[ + \frac{1}{2} \sum_{i=1}^{M} \ln \alpha_i + \frac{1}{2} \ln |\Sigma| - \frac{N}{2} \ln (2\pi). \]
(53)

3.4. Maximization of the Evidence Function with RVM. Firstly, we consider the maximizing \( p(t \mid X, \alpha, \beta) \) with respect to \( \alpha_i \). By making use of (33) and Lemma 3, we obtain

\[ \frac{d \ln |\Sigma|}{d \alpha_i} = Tr \left( \Sigma^{-1} \frac{d \Sigma}{d \alpha_i} \right) = Tr \left( (A + \beta \Phi^T \Phi)^{-1} \frac{d (A + \beta \Phi^T \Phi)}{d \alpha_i} \right). \]
(54)

From (54) and Lemma 4, we obtain

\[ \frac{d \ln |\Sigma|}{d \alpha_i} = -Tr \left( \frac{d (A + \beta \Phi^T \Phi)}{d \alpha_i} \Sigma \right), \]
(55)

where we have used

\[ \frac{d (A + \beta \Phi^T \Phi)^{-1}}{d \alpha_i} = -(A + \beta \Phi^T \Phi)^{-1} \frac{d (A + \beta \Phi^T \Phi)}{d \alpha_i} (A + \beta \Phi^T \Phi)^{-1}. \]
(56)

and (33).

From (55), we obtain

\[ \frac{d \ln |\Sigma|}{d \alpha_i} = -\Sigma_{ii}, \]
(57)

where \( \Sigma_{ii} \) is the \( i \)th component of the posterior covariance \( \Sigma \) defined by (33).

From (42), we obtain

\[ \frac{\partial E(m)}{\partial \alpha_i} = \frac{1}{2} m_i^2. \]
(58)

From (57) and (58), we easily obtain the stationary point of (53) with respect to \( \alpha_i \) such that

\[ \frac{1}{2\alpha_i} - \frac{1}{2} \Sigma_{ii} - \frac{1}{2} m_i^2 = 0, \]
(59)

where \( m_i \) is the \( i \)th component of the posterior mean \( m \) defined by (32).

After multiplying both sides of (59) by \( 2\alpha_i \) and arranging, we have the following result:

\[ \alpha_i m_i^2 = 1 - \alpha_i \Sigma_{ii}, \]
(60)

\[ \alpha_i = \frac{\gamma_i}{m_i^2}, \]
(61)

where \( \gamma_i = 1 - \alpha_i \Sigma_{ii} \).

Similarly, we can maximize the log marginal likelihood (53) with respect to \( \beta \). Firstly, we obtain the derivative of \( \ln |\Sigma| \) with respect to \( \beta \) using (33) and Lemma 3:

\[ \frac{d \ln |\Sigma|}{d \beta} = Tr \left( \Sigma^{-1} \frac{d \Sigma}{d \beta} \right) = Tr \left( (A + \beta \Phi^T \Phi)^{-1} \frac{d (A + \beta \Phi^T \Phi)}{d \beta} \right). \]
(62)

By making use of (33) and Lemma 4, we obtain

\[ \frac{d \ln |\Sigma|}{d \beta} = -Tr \left( \Phi^T \Phi \Sigma \right). \]
(63)

Since \((A + \beta \Phi^T \Phi)^{-1} = I_M\), we rearrange and get the following result:

\[ \Phi^T \Sigma = \beta^{-1} (I_M - A \Sigma). \]
(64)

By using definition of matrix \( A \), (63) and (64), we obtain

\[ \frac{d \ln |\Sigma|}{d \beta} = -\frac{1}{\beta} \sum_i (1 - \alpha_i \Sigma_{ii}). \]
(65)

From (42), we obtain

\[ \frac{\partial E(m)}{\partial \beta} = \frac{1}{2} \| t - \Phi m \|^2. \]
(66)

From (65) and (66), we obtain the stationary point of (53) with respect to \( \beta \) such that

\[ \frac{N}{\beta} - \| t - \Phi m \|^2 - \frac{1}{\beta} \sum_i \gamma_i = 0. \]
(67)

By multiplying both sides of (67) by \( \beta \) and rearranging, we obtain

\[ \frac{1}{\beta} = \frac{\| t - \Phi m \|^2}{N - \sum \gamma_i}. \]
(68)

The result of (61) and (68) is consistent with [5, 8]. We note that (61) and (68) are implicit solutions for \( \alpha \) and \( \beta \) since both \( \gamma_i \) and \( 1/\beta \) depend on \( \alpha \) and \( \beta \). To address the problem, an iterative procedure is devised. We use (32) and (33) to evaluate the mean and covariance of the posterior distribution by choosing initial value \( \alpha \) and \( \beta \), respectively, and then alternately re-evaluate the hyperparameters using (61) and (68). Again, learning continues to use (32) and (33) to re-evaluate the mean and covariance of the posterior distribution, until a suitable convergence standard is satisfied.

4. EM Algorithm for RVM

A powerful and elegant technique for finding likelihood solutions for probabilistic models with latent variables is called the expectation-maximization algorithm or EM algorithm which also forms the basis for derivation of the variational framework [19, 20].

The integration with respect to the weights to obtain the marginal likelihood can make use of the EM algorithm for maximization with the weights as latent variables very
naturally. Since it is most common to use the EM for maximization, $L$ will now become the (positive) log evidence. To derive the EM algorithm for the RVM, we firstly obtain the log marginal likelihood by marginalizing the weights from the joint distribution over $t$ and $w$:

$$L = \ln p(t \mid X, \alpha, \beta) = \ln \int p(t, w \mid X, \alpha, \beta) \, dw,$$

where the joint distribution over $t$ and $w$ is both equal to the marginal likelihood times the posterior on $w$ and to the likelihood multiplied by the prior on $w$:

$$p(t, w \mid X, \alpha, \beta) = p(t \mid X, w, \beta) p(w \mid \alpha) = p(w \mid t, X, \alpha, \beta) p(t \mid X, \alpha, \beta).$$

By defining a variational probability distribution $q(w)$ over the weights and using Jensen’s inequality, we obtain a lower bound on $L$:

$$L = \ln \int q(w) \frac{p(t, w \mid X, \alpha, \beta)}{q(w)} \, dw \geq \int q(w) \ln \frac{p(t, w \mid X, \alpha, \beta)}{q(w)} \, dw \quad (71)$$

$$= F(q, \alpha, \beta).$$

The EM algorithm achieves the maximization of the log marginal likelihood $L$ by iteratively maximizing the lower bound. In the Expectation step (E-step), $F$ is maximized with respect to the variational distribution $q(w)$ for fixed parameters $\alpha$ and $\beta$, and in the Maximization step (M-step), $F$ is maximized with respect to the hyperparameters $\alpha$ and $\beta$ for fixed $q(w)$. Insight into how to perform the E-step can be gained by rewriting the lower bound $F$ as

$$F = L + \text{KL}(q(w) \parallel p(w \mid t, X, \alpha, \beta)),$$

where $\text{KL}(q(w) \parallel p(w \mid t, X, \alpha, \beta))$ is the Kullback–Leibler (KL) divergence between the variational distribution $q(w)$ and the posterior distribution of the weights. The KL divergence is always a positive number and is only equal to zero if the two distributions are identical. The E-step corresponds thus to making $q(w)$ equal to the posterior on $w$, which implies $F = L$ and $q(w) = p(w \mid t, X, \alpha, \beta)$. Since the posterior is Gaussian, the E-step reduces to computing its mean $m$ and covariance $\Sigma$ defined by (32) and (33).

To perform the M-step, we rewrite $F$ in a different way:

$$F = \int q(w) \ln p(t, w \mid X, \alpha, \beta) \, dw - H(q(w)), \quad (73)$$

where $H(q(w))$ is entropy of $q(w)$, which is an irrelevant constant because it does not depend on $\alpha$ or $\beta$, which implies $\partial H(q(w)) / \partial \alpha = 0$ and $\partial H(q(w)) / \partial \beta = 0$. Therefore, the M-step is to maximize the average of the log joint distribution of $t$ and $w$ over $q(w)$ with respect to the parameters $\alpha$ and $\beta$:

$$F = \int q(w) \ln p(t, w \mid X, \alpha, \beta) \, dw - H(q(w))$$

$$= \int q(w) (\ln p(w \mid \alpha) + \ln p(t \mid X, w) - H(q(w))) \, dw$$

$$= \int q(w) \ln p(w \mid \alpha) \, dw + \int q(w) \ln p(t \mid X, w, \beta) \, dw - H(q(w)).$$

From the first term of (74), we obtain

$$\int q(w) \ln p(w \mid \alpha) \, dw$$

$$= \int q(w) \left( \frac{1}{2} \ln |A| - \frac{M}{2} \ln (2\pi) - \frac{1}{2} w^T A w \right) \, dw$$

$$= \frac{1}{2} \ln |A| - \frac{M}{2} \ln (2\pi) - \frac{1}{2} \int q(w) w^T A w \, dw,$$

where we have used $\int q(w) w = 1$.

From the third term of (75) and Lemma 5, we obtain

$$\int q(w) w^T A w \, dw$$

$$= \text{Tr} \left( \int (w^T A w) \, dw \right)$$

$$= \text{Tr} \left( A \left( \text{E}_w (w^w) \right) \right)$$

$$= \text{Tr} \left( A \left( \text{mm}^T + \Sigma \right) \right)$$

$$= \text{Tr} \left( A \text{Am} + \text{Tr} (A \Sigma) \right)$$

$$= m^T A m + \text{Tr} (A \Sigma),$$

where $E_w(\cdot)$ denotes the expectation of a random variable over the distribution $q(w)$ and we have used $E_w(ww^T) = \text{mm}^T + \Sigma$, where the mean $m$ and covariance $\Sigma$ is defined by (32) and (33).

From (75) and (76), we obtain

$$\int q(w) \ln p(w \mid \alpha) \, dw$$

$$= \frac{1}{2} \ln |A| - \frac{M}{2} \ln (2\pi) - \frac{1}{2} \int q(w) w^T A w \, dw$$

$$= \frac{1}{2} \ln |A| - \frac{M}{2} \ln (2\pi) - \frac{1}{2} m^T A m - \frac{1}{2} \text{Tr} (A \Sigma).$$

We start to calculate the second term of (74) and obtain
where we have used Lemma 5 and \( E_q(ww^T) = \text{mm}^T + \Sigma \).

From (78) and (79), we obtain

\[
\int q(w) \ln p(t \mid X, w, \beta) \, dw \\
= \int q(w) \left( \frac{N}{2} \ln \beta - \frac{N}{4} \ln 2(2 \pi) - \frac{\beta}{2} \left( \|t - \Phi w\|^2 \right) \right) \, dw \\
= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2(2 \pi) - \frac{\beta}{2} \|t - \Phi w\|^2 \\
= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2(2 \pi) - \frac{\beta}{2} \|t - \Phi w\|^2 \\
= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2(2 \pi) \\
= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2(2 \pi) \\
- \frac{\beta}{2} \left( \|t - \Phi w\|^2 + \text{Tr}(\Phi^T \Phi \Sigma) \right). \\
(80)
\]

From (77) and (80), we obtain

\[
F = \int q(w) \ln p(t, w \mid X, \alpha, \beta) \, dw - H(q(w)) \\
= \frac{1}{2} \ln |A| - \frac{M + N}{2} \ln 2(2 \pi) - \frac{1}{2} \text{Tr}(A \Sigma) - \frac{N}{2} \ln \beta \\
- \frac{\beta}{2} \left( \|t - \Phi m\|^2 + \text{Tr}(\Phi^T \Phi \Sigma) \right) - H(q(w)). \\
(81)
\]

We easily obtain the stationary point of (81) with respect to \( \alpha \) and \( \beta \) such that

\[
\frac{\partial F}{\partial \alpha_i} = 0, \\
\frac{\partial F}{\partial \beta} = 0. \\
(82, 83)
\]

So, we obtain the update rules as follows:

\[
\alpha_i^{\text{new}} = \frac{1}{m_i + \Sigma_{ii}}, \\
\beta^{\text{new}} = \frac{N}{\|t - \Phi m\|^2 + \text{Tr}(\Phi^T \Phi \Sigma)}. \\
(84, 85)
\]

where \( m_i \) is the \( i \)-th component of the posterior mean \( m \) defined by (32) and \( \Sigma_{ii} \) is the \( i \)-th component of the posterior covariance \( \Sigma \) defined by (33).

By making use of (64) and (65), we further obtain the following result for (85):

\[
\beta^{\text{new}} = \frac{N}{\|t - \Phi m\|^2 + \left( \beta^{\text{old}} \right)^{-1} \sum_i y_i}, \\
(86)
\]

where \( y_i = 1 - \alpha_i \Sigma_{ii} \).

From the abovementioned discussion, we obtain the update rules by using the EM algorithm to maximize the log marginal likelihood, which is also consistent with Quiñonero Candela’s result, except that we use the precision matrix as the covariance [21].

5. Conclusions

From the abovementioned discussion, we can draw the conclusion that the novel logical framework presented in this paper based on Taylor expansion shows some advantages compared with the conventional method of completing the square, which is easier to be implemented due to the clear logical framework, and avoids the limitation in seeking the term of completing the square intentionally. The approximation and maximization of
the evidence function associated with RVM is consistent with Tipping’s result of (61) and (68) using the novel logical framework. We obtain the update rules of (84) and (85) by using the EM algorithm for maximization of the log marginal likelihood, which is also consistent with Quiñonero Candela’s result, except that we use the precision matrix as the covariance.

Appendix

A. The Proofs of (21) and (22)

A quadratic form that belongs to the exponent term in a normal distribution is given for determining the corresponding mean and covariance of the distribution. The problem can be solved simply by noting that the negative exponent in a normal distribution $N(x \mid \mu, \Sigma)$ can be written as

$$f(w) = \frac{1}{2} \left( w^T \left( \Sigma^{-1} + \beta \Phi^T \Phi \right) w - 2 w^T \left( \Sigma^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t} \right) + \beta \mathbf{t}^T \mathbf{t} + \mathbf{m}_0^T \mathbf{m}_0 \right).$$  \hspace{1cm} (A.4)

By setting $\nabla f(w) = (\Sigma^{-1} + \beta \Phi^T \Phi) \mathbf{w} - (\Sigma^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) = 0$, we obtain

$$m_N = (\Sigma^{-1} + \beta \Phi^T \Phi)^{-1} (\Sigma^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}),$$

$$S_N^{-1} = \nabla \nabla f(w) = \Sigma^{-1} + \beta \Phi^T \Phi. \hspace{1cm} (A.5)$$

We then choose suitable normalization term for (A.2) and obtain the result of (21) and (22).

A. The Proofs of (32) and (33)

From (29), we obtain

$$p(t \mid X, w, \beta) = N(t \mid \Phi w, \beta^{-1}). \hspace{1cm} (B.1)$$

From (30), we obtain

$$p(w \mid \alpha) = N(w \mid 0, A^{-1}). \hspace{1cm} (B.2)$$

By using $\mathbf{m}_0 = 0$ and $\mathbf{S}_0 = A^{-1}$, we obtain from (21) and (22) $f(x) = \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu).$ \hspace{1cm} (A.1)

By setting $\nabla f(x) = \Sigma^{-1} (x - \mu) = 0$, we obtain $x = \mu$, which implies the stationary point of $f(x)$ equals to the mean $\mu$ of the normal distribution $N(x \mid \mu, \Sigma)$. In addition, we obtain $\Sigma^{-1} = \nabla \nabla f(x)$. Next, we will use the rules to solve the mean and covariance of a normal distribution.

From (18) and (19), we use the product rule of probability and obtain

$$p(w \mid t, X, \beta) \propto p(w) p(t \mid X, w, \beta). \hspace{1cm} (A.2)$$

We substitute (18) and (19) into (A.2) and obtain the following form:

$$\Sigma^{-1} = S_N^{-1} = \Sigma^{-1} + \beta \Phi^T \Phi = A + \beta \Phi^T \Phi,$$

$$m = S_N^{-1} (S^{-1}_N \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) = \beta (A + \beta \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}. \hspace{1cm} (B.3)$$

So, we prove the result of (32) and (33).

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Fundamental Research Funds on Basic Research Project of Heilongjiang Provincial Colleges and Universities (nos. KYYWF10236180209, KYYWF10236190103, and KYYWF10236190203) and Innovation Project for University Students of Heilongjiang Province (no. 201610236016).

References

[1] A. E. Hoerl and R. W. Kennard, “Ridge regression: biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
10 Mathematical Problems in Engineering

[2] B. M. G. Kibria, “Performance of some new ridge regression estimators,” *Communications in Statistics-Simulation and Computation*, vol. 32, no. 2, pp. 419–435, 2003.

[3] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] R. E. Kass and A. E. Raftery, “Bayes factors,” *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 377–395, 1995.

[5] E. T. Michael, “Sparse Bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[6] C. E. Rasmussen and J. Quiñonero-Candela, “Healing the relevance vector machine by augmentation,” in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 689–696, University of Bonn, Bonn, Germany, 2005.

[7] L. Martino, J. Vicent, and G. Camps-Valls, “Automatic emulation by adaptive relevance vector machines,” in *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, Tromsø, Norway, June 2017.

[8] M. B. Christopher, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 2006.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computation*, John Hopkins University Press, Baltimore, MD, USA, 1996.

[10] H. Lütkepohl, *Handbook of Matrices*, Wiley, Hoboken, NJ, USA, 1996.

[11] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Economics*, Wiley, Hoboken, NJ, USA, 1999.

[12] S. Dan, *Mathematical Analysis for Machine Learning and Data Mining*, pp. 652–655, World Scientific, Singapore, 2018.

[13] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley, Hoboken, NJ, USA, 1994.

[14] A. Gelman, J. B. Carlin, and H. S. Stern, *Bayesian Data Analysis*, Chapman and Hall, London, UK, 2004.

[15] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, New York, NY, USA, 1985.

[16] G. Wahba, “A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem,” *The Annals of Statistics*, vol. 13, no. 4, pp. 1378–1402, 1985.

[17] S. F. Gull, *Maximum Entropy and Bayesian Methods*, Kluwer, Amsterdam, Netherlands, 1989.

[18] D. J. C. MacKay, “Bayesian interpolation,” *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM Algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[20] G. J. McLachlan and T. Krishnan, *The EM Algorithm and its Extensions*, Wiley, Hoboken, NJ, USA, 1997.

[21] J. Quiñonero Candela, “Sparse probabilistic linear models and the RVM,” *Learning with Uncertainty-Gaussian Processes and Relevance Vector Machines (Ph.D.)*, Technical University of Denmark, Lyngby, Denmark, 2004.