Lecture Notes on Relevance Vector Machines

Man-Wai MAK

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
enmwmak@polyu.edu.hk

Abstract

This document provides full derivations of the relevance vector machines (RVM). It consolidates the equations from Tipping's original paper and Bishop's book to give a more coherent description and explanation of RVM.

Please cite this document as: M.W. Mak, "Lecture Notes on Relevance Vector Machines", *Technical Report and Lecture Note Series, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University*, Auguest 2015.

1. Support Vector Machines

Support vector machine (SVMs) are well-known supervised learning method used for classification and regression. Assume that we are given N training vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with labels $y_n \in \{+1, -1\}, n = 1, \ldots, N$. Using the pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$, an SVM can be trained. Given a test vector $\mathbf{x}_t$, the SVM's output is written as

$$f(\mathbf{x}_t; \mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x}_t, \mathbf{x}_i) + w_0 \tag{1}$$

where $\mathbf{w} = [w_0, \ldots, w_N]$ are the weights determined by maximizing the margin of separation between the two classes, w_0 is a bias term, and $K(\mathbf{x}_t, \mathbf{x}_i)$ is a kernel function. Note that there is no assumption on the probability distribution of $\mathbf{x}_t$, w_i, and $f(\mathbf{x}_t; \mathbf{w})$.

2. Relevance Vector Machines

RVM is a Bayesian treatment of Eq. 1. When an RVM is applied to regression, the target y's are assumed to follow a Gaussian distribution with

mean $f(\mathbf{x};\mathbf{w})$ and variance σ^2; when it is applied to classification, the target conditional distribution $p(y|\mathbf{x})$ is assumed to follow a Bernoulli distribution.

2.1. RVM Regression

Assume that we have a training set comprising input-target pairs:

$$\mathcal{T} = \{(\mathbf{x}_n, y_n); n = 1, \ldots, N\}.$$

Denote $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_n\}$ as the input vectors and target outputs, respectively. When an RVM is applied to regression, the targets y_n's are assumed to be sampled from the following model:

$$y_n = f(\mathbf{x}_n;\mathbf{w}) + \epsilon_n, \quad n = 1, \ldots, N,$$

where $f(\mathbf{x}_n;\mathbf{w})$ is given by Eq. 1, and ϵ_n follows a Gaussian distribution with zero mean and variance σ^2. This is equivalent to say that

$$p(y_n|\mathbf{x}_n) = \mathcal{N}(y_n|f(\mathbf{x}_n;\mathbf{w}), \sigma^2).$$

Assume also that y_n's $(n = 1, \ldots, N)$ are independent, the likelihood of the training data set can be written as:

$$\begin{aligned} p(\mathbf{y}|\mathbf{w},\sigma^2) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{\Phi}\mathbf{w}\|^2\right\} \\ &= \mathcal{N}\left(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I}\right) \end{aligned} \tag{2}$$

where

$$\begin{aligned} \mathbf{y} &= [y_1, \ldots, y_N]^\mathsf{T}; \quad \mathbf{w} = [w_0, \ldots, w_N]^\mathsf{T}; \\ \mathbf{\Phi} &= [\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)]^\mathsf{T} \\ \boldsymbol{\phi}(\mathbf{x}_i) &= [1, K(\mathbf{x}_i,\mathbf{x}_1), K(\mathbf{x}_i,\mathbf{x}_2), \ldots, K(\mathbf{x}_i,\mathbf{x}_N)]^\mathsf{T}. \end{aligned} \tag{3}$$

To avoid over-fitting, RVM defines a zero-mean Gaussian prior distribution over $\mathbf{w}$:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} \mathcal{N}(w_i|0,\alpha_i^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1}) \tag{4}$$

where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_N]^\mathsf{T}$, α_i is the hyperparameter associated with weight w_i and $\mathbf{A} = \mathrm{diag}(\alpha_0, \alpha_1, \ldots, \alpha_N)$.

Given the distribution of $\mathbf{y}$ in Eq. 2 and the prior distribution of $\mathbf{w}$ in Eq. 4, we can use the formula of conditional Gaussians (Eq. 2.116 in [1]) to

obtain the posterior distribution over the weights as follows:[1]

$$p(\mathbf{w}|\mathbf{y}, \mathcal{X}, \boldsymbol{\alpha}, \sigma^2) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5}$$

where

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{y} \quad \text{and} \quad \boldsymbol{\Sigma} = (\sigma^{-2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \mathbf{A})^{-1}. \tag{6}$$

The optimal value of $\boldsymbol{\alpha}$ and σ^2 can be obtained by maximizing the following marginal likelihood with respect to $\boldsymbol{\alpha}$ and σ^2:

$$\begin{aligned} p(\mathbf{y}|\mathcal{X}, \boldsymbol{\alpha}, \sigma^2) &= \int p(\mathbf{y}|\mathcal{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\ &= \int \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})d\mathbf{w} \\ &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}}). \end{aligned} \tag{7}$$

Setting

$$\frac{\partial \log p(\mathbf{y}|\mathcal{X}, \boldsymbol{\alpha}, \sigma^2)}{\partial \log \alpha_i} = 0 \quad \text{and} \quad \frac{\partial \log p(\mathbf{y}|\mathcal{X}, \boldsymbol{\alpha}, \sigma^2)}{\partial \log \sigma^{-2}} = 0,$$

we obtain the following update formulae for α_i and σ^2 (see Appendix for the derivations):

$$\alpha_i^{\text{new}} = \frac{\gamma_i}{\mu_i^2} \quad \text{and} \quad (\sigma^2)^{\text{new}} = \frac{\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{N - \sum_{i=0}^{N}\gamma_i}, \tag{8}$$

where μ_i is the i-th component of $\boldsymbol{\mu}$ in Eq. 6 and $\gamma_i = 1 - \alpha_i\Sigma_{ii}$ with Σ_{ii} being the i-th diagonal element of $\boldsymbol{\Sigma}$ in Eq. 6. During the optimization, many of the hyperparameters α_i tend to infinity and the corresponding weights w_i become zero; the vectors $\mathbf{x}_i$ corresponding to the non-zero weights are considered as **relevance vectors**.

By considering $\mathbf{w}$ probabilistic and using the notion of conditional independence [1], the predictive distribution of y_t given a test vector $\mathbf{x}_t$ is

$$p(y_t|\mathbf{y}, \mathbf{x}_t, \mathcal{X}) = \int_{\sigma^2}\int_{\boldsymbol{\alpha}}\int_{\mathbf{w}} p(y_t|\mathbf{x}_t, \mathbf{w}, \boldsymbol{\alpha}, \sigma^2)p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{y}, \mathcal{X})d\mathbf{w}d\boldsymbol{\alpha}d\sigma^2 \tag{9}$$

[1]To use Eq. 2.116 of [1], we consider $\mathbf{x}$ and $\mathbf{y}$ in Eq. 2.116 as our $\mathbf{w}$ and $\mathbf{y}$, respectively. Also, $\boldsymbol{\Lambda}$ and $\mathbf{L}$ in Eq. 2.113 and Eq. 2.114 are our $\mathbf{A}$ and $\sigma^{-2}\mathbf{I}$, respectively. Moreover, $\boldsymbol{\mu}$ and $\mathbf{b}$ in Eq. 2.113 and Eq. 2.114 are zero vectors in our case.

where

$$p(y_t|\mathbf{x}_t, \mathbf{w}, \boldsymbol{\alpha}, \sigma^2) = p(y_t|\mathbf{x}_t, \mathbf{w}, \sigma^2) \tag{10}$$

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{y}, \mathcal{X}) = p(\mathbf{w}|\mathbf{y}, \mathcal{X}, \boldsymbol{\alpha}, \sigma^2)p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y}, \mathcal{X}). \tag{11}$$

Instead of computing the posterior $p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y})$ in Eq. 11, Tipping [2] used a delta function at the most probable values of $\boldsymbol{\alpha}$ and σ^2 as an approximation. Therefore, using Eq. 11 and assuming uniform priors for $\boldsymbol{\alpha}$ and σ^2, Eq. 9 reduces to

$$\begin{aligned} p(y_t|\mathbf{y}, \mathbf{x}_t, \mathcal{X}) &= \int_{\sigma^2}\int_{\boldsymbol{\alpha}}\int_{\mathbf{w}} p(y_t|\mathbf{x}_t, \mathbf{w}, \boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}})p(\mathbf{w}|\mathbf{y}, \mathcal{X}, \boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}})p(\boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}}|\mathbf{y})d\mathbf{w}d\boldsymbol{\alpha}d\sigma^2 \\ &= \int_{\mathbf{w}} p(y_t|\mathbf{x}_t, \mathbf{w}, \boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}})p(\mathbf{w}|\mathbf{y}, \mathcal{X}, \boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}})d\mathbf{w} \\ &= \int_{\mathbf{w}} \mathcal{N}(y_t|\boldsymbol{\phi}(\mathbf{x}_t)^\mathsf{T}\mathbf{w}, \sigma^2_{\text{MP}})\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\text{MP}}, \boldsymbol{\Sigma}_{\text{MP}})d\mathbf{w} \end{aligned} \tag{12}$$

where

$$\begin{aligned} (\boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}}) &= \arg\max_{\boldsymbol{\alpha},\sigma^2} p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y}, \mathcal{X}) \\ &= \arg\max_{\boldsymbol{\alpha},\sigma^2} p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2, \mathcal{X})p(\boldsymbol{\alpha})p(\sigma^2) \\ &= \arg\max_{\boldsymbol{\alpha},\sigma^2} \int p(\mathbf{y}|\mathbf{w}, \sigma^2, \mathcal{X})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\ &= \arg\max_{\boldsymbol{\alpha},\sigma^2} \int \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})d\mathbf{w} \\ &= \arg\max_{\boldsymbol{\alpha},\sigma^2} \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T}). \end{aligned} \tag{13}$$

and

$$\boldsymbol{\mu}_{\text{MP}} = \sigma^{-2}_{\text{MP}}\boldsymbol{\Sigma}_{\text{MP}}\boldsymbol{\Phi}^\mathsf{T}\mathbf{y} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\text{MP}} = \left(\sigma^{-2}_{\text{MP}}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \mathbf{A}\right)^{-1}. \tag{14}$$

Because both terms in the integrand of Eq. 12 are Gaussians, the predictive distribution is also a Gaussian:[2]

$$p(y_t|\mathbf{y}, \mathcal{X}, \mathbf{x}_t, \boldsymbol{\alpha}_{\text{MP}}, \sigma^2_{\text{MP}}) = \mathcal{N}(y_t|g(\mathbf{x}_t), \sigma^2_t) \tag{15}$$

[2] Again, we make use of the marginal and conditional Guassian formulae in Eqs. 2.113–2.115 of [1] to derive Eq. 16. Specifically, in Eqs. 2.113–2.115 of [1], we substitute $\mathbf{x}$ by $\mathbf{w}$, $\mathbf{y}$ by y_t, $\boldsymbol{\mu}$ by $\boldsymbol{\mu}_{\text{MP}}$, $\boldsymbol{\Lambda}^{-1}$ by $\boldsymbol{\Sigma}_{\text{MP}}$, $\mathbf{L}^{-1}$ by σ^2_{MP}, $\mathbf{b}$ by 0, and $\mathbf{A}$ by $\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}$.

with

$$g(\mathbf{x}_t) = \boldsymbol{\mu}_{\text{MP}}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x}_t) \quad \text{and} \quad \sigma_t^2 = \sigma_{\text{MP}}^2 + \boldsymbol{\phi}(\mathbf{x}_t)^{\mathsf{T}} \boldsymbol{\Sigma}_{\text{MP}} \boldsymbol{\phi}(\mathbf{x}_t). \tag{16}$$

2.2. RVM Classification

When RVM is applied to classification, the target conditional distribution $p(y|\mathbf{x})$ is assumed to follow a Bernoulli distribution. Assume that we have a set of training i-vectors $\mathcal{X} = \{\mathcal{X}_s, \mathcal{X}_b\}$ and $y_i = 1$ when $\mathbf{x}_i \in \mathcal{X}_s$ and $y_i = 0$ when $\mathbf{x}_i \in \mathcal{X}_b$, the likelihood of the training data set can be written as [2]:

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{N} \sigma\left(f(\mathbf{x}_i; \mathbf{w})\right)^{y_i} \left\{1 - \sigma\left(f(\mathbf{x}_i; \mathbf{w})\right)\right\}^{1-y_i}, \quad y_i \in \{0, 1\}. \tag{17}$$

where

$$N = |\mathcal{X}_s| + |\mathcal{X}_b|;\ \mathbf{y} = [y_1, \ldots, y_N]^{\mathsf{T}};\ \ \mathbf{w} = [w_0, \ldots, w_N]^{\mathsf{T}} \tag{18}$$

and $\sigma\{\cdot\}$ is the logistic sigmoid link function $\sigma(z) = \frac{1}{1+e^{-z}}$. Similar to RVM regression, RVM classification also introduces a zero-mean Gaussian prior distribution over $\mathbf{w}$ as defined in Eq. 4.

Using Eq. 17 and Eq. 4, we can obtain the posterior distribution of $\mathbf{w}$:

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})}{\int p(\mathbf{y}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w}} = \frac{g(\mathbf{w})}{p(\mathbf{y}|\boldsymbol{\alpha})}, \tag{19}$$

where we have defined $g(\mathbf{w}) \equiv p(\mathbf{y}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$. Taking logarithm of $g(\mathbf{w})$, we have

$$\begin{aligned}
\log g(\mathbf{w}) &= \log p(\mathbf{y}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha}) \\
&= \sum_{i=1}^{N} \left\{y_i \log\left[\sigma(f(\mathbf{x}_i; \mathbf{w}))\right] + (1 - y_i) \log\left[1 - \sigma(f(\mathbf{x}_i; \mathbf{w}))\right]\right\} \\
&\qquad - \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w} + \text{const} \\
&= \sum_{i=1}^{N} \left\{y_i \log\left[\sigma\left(\boldsymbol{\phi}(\mathbf{x}_i)^{\mathsf{T}}\mathbf{w}\right)\right] + (1 - y_i) \log\left[1 - \sigma\left(\boldsymbol{\phi}(\mathbf{x}_i)^{\mathsf{T}}\mathbf{w}\right)\right]\right\} \\
&\qquad - \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w} + \text{const},
\end{aligned} \tag{20}$$

where we have used $f(\mathbf{x}_i; \mathbf{w}) = \boldsymbol{\phi}(\mathbf{x}_i)^{\mathsf{T}}\mathbf{w}$. Note that because $p(\mathbf{y}|\mathbf{w})$ in

Eq. 17 is not a Gaussian, we cannot analytically perform the integration in Eq. 19 to obtain a closed-form solution for $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$. One possible solution is to use the Laplace's method [1] to approximate $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$ by a Gaussian distribution. The idea is to find a Gaussian approximation $q(\mathbf{w})$ with mean $\mathbf{w}_0$ equals to a mode of $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$. This can be achieved by approximate $\log g(\mathbf{w})$ by a Taylor expansion around $\mathbf{w}_0$:

$$\log g(\mathbf{w}) \approx \log g(\mathbf{w}_0) - \frac{1}{2} (\mathbf{w} - \mathbf{w}_0)^\mathsf{T} \mathbf{H} (\mathbf{w} - \mathbf{w}_0), \tag{21}$$

where $\mathbf{H}$ is a Hessian matrix

$$\begin{aligned} \mathbf{H} &= -\nabla\nabla_\mathbf{w} \log g(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}_0} \qquad (22) \\ &= \frac{\partial}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} \log g(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}_0} \\ &= \sum_{i=1}^{N} \sigma\left(\phi(\mathbf{x}_i)^\mathsf{T}\mathbf{w}_0\right)\left[1 - \sigma\left(\phi(\mathbf{x}_i)^\mathsf{T}\mathbf{w}_0\right)\right]\phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_i) + \mathbf{A} \\ &= \boldsymbol{\Phi}^\mathsf{T}\mathbf{B}\boldsymbol{\Phi} + \mathbf{A} \qquad (23) \end{aligned}$$

where $\mathbf{B}$ is an $(N+1) \times (N+1)$ diagonal matrix with diagonal elements

$$b_{ii} = \sigma\left(\phi(\mathbf{x}_i)^\mathsf{T}\mathbf{w}_0\right)\left[1 - \sigma\left(\phi(\mathbf{x}_i)^\mathsf{T}\mathbf{w}_0\right)\right], \tag{24}$$

and $\boldsymbol{\Phi}$ and $\boldsymbol{\phi}(\mathbf{x}_t)$ are defined in Eq. 3.

The value of $\mathbf{w}_0$ can be obtained by using iterative reweighted least squares (IRLS) as follows:

$$\mathbf{w}_0^{\text{new}} = \mathbf{w}_0^{\text{old}} - (\mathbf{H}^{\text{old}})^{-1} \nabla_\mathbf{w} \log g(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}_0^{\text{old}}} \tag{25}$$

where

$$\nabla_\mathbf{w} \log g(\mathbf{w}) = \boldsymbol{\Phi}^\mathsf{T}\left(\mathbf{y} - \left[\sigma(\phi(\mathbf{x}_1)^\mathsf{T}\mathbf{w}), \ldots, \sigma(\phi(\mathbf{x}_N)^\mathsf{T}\mathbf{w})\right]^\mathsf{T}\right) - \mathbf{A}\mathbf{w}.$$

At convergency, the gradient is zero and therefore we have

$$\mathbf{w_0} \to \mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T}\left(\mathbf{y} - \left[\sigma(\phi(\mathbf{x}_1)^\mathsf{T}\mathbf{w}_0), \ldots, \sigma(\phi(\mathbf{x}_N)^\mathsf{T}\mathbf{w}_0)\right]^\mathsf{T}\right)$$

Taking exponential of Eq. 21 and noting that $q(\mathbf{w}) \propto g(\mathbf{w})$, we have

$$\begin{aligned} q(\mathbf{w}) &= \frac{|\mathbf{H}|^{1/2}}{(2\pi)^{(N+1)/2}} \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\mathsf{T}}\mathbf{H}(\mathbf{w}-\mathbf{w}_0)\right\} \\ &= \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{H}^{-1}), \end{aligned} \tag{26}$$

which is a Gaussian distribution with mean $\mathbf{w}_0$ and covariance matrix $\mathbf{H}^{-1}$.

We then use $q(\mathbf{w})$ to approximate the posterior $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$ around the mode $\mathbf{w}_0$. Comparing the covariance matrix $\mathbf{H}^{-1}$ in Eq. 22 with that in Eq. 2.117 of [1] reveals that $\mathbf{B}$ is the precision matrix of $p(\mathbf{y}|\mathbf{w})$ (see Eq. 2.114 of [1]). As a result, using Eq. 2.116 of [1], we obtain the posterior mean of the weights in $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$ as

$$\mathbf{w}_{\text{MP}} = \mathbf{H}^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{B}\mathbf{y}. \tag{27}$$

Given Eqs. 8, 22, 24, 25, and 27, we may proceed the estimation of $\boldsymbol{\alpha}$ as follows. First, we initialize $\boldsymbol{\alpha}$ to obtain $\mathbf{A}$. Then, we initialize $\mathbf{w}$ and use Eq. 25 to estimate $\mathbf{w}_0$. We then plug this $\mathbf{w_0}$ into Eq. 22 and Eq. 24 to obtain $\mathbf{H}$ and $\mathbf{B}$, respectively, followed by estimating $\mathbf{w}_{\text{MP}}$ using Eq. 27. A new estimation of $\boldsymbol{\alpha}$ is then obtained by maximizing the likelihood $p(\mathbf{y}|\boldsymbol{\alpha})$, i.e., using Eq. 8 without σ^2. Then, the cycle is repeated.

Appendix: Estimating Hyperparameters

In Section 3.5.1 and Section 3.5.2 of Bishop's book, the hyperparameters α is estimated by using the fact that the eigenvalues of $(\alpha\mathbf{I} + \mathbf{L})$ are $(\alpha + \lambda_i)$, where λ_i's are the eigenvalues of symmetric matrix $\mathbf{L}$. However, this property does not hold if the diagonal elements of the first matrix are not equal, which is the case in RVM where $\mathbf{A} = \text{diag}\{\alpha_0, \ldots, \alpha_N\}$, i.e., the eigenvalues of $(\mathbf{A} + \mathbf{L})$ are not equal to $\alpha_i + \lambda_i, i = 0, \ldots, N$.

Instead of completing the square over $\mathbf{w}$, the optimal value of $\boldsymbol{\alpha}$ and σ^2 can be obtained by maximizing the following marginal likelihood with respect to $\boldsymbol{\alpha}$ and σ^2:

$$\begin{aligned} p(\mathbf{y}|\mathcal{X}, \boldsymbol{\alpha}, \sigma^2) &= \int p(\mathbf{y}|\mathcal{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\ &= \int \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})d\mathbf{w} \\ &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}}), \end{aligned} \tag{28}$$

where $\mathbf{A} = \text{dig}\{\alpha_0, \ldots, \alpha_N\}$. Taking logarithm of Eq. 28 and ignoring terms independent of $\boldsymbol{\alpha}$ and σ^2, the log-likelihood of $\mathbf{y}$ becomes

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = -\frac{1}{2}\log|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}}| - \frac{1}{2}\mathbf{y}^{\mathsf{T}}(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}})^{-1}\mathbf{y}. \quad (29)$$

The first term can be expressed as

$$\begin{aligned} &-\frac{1}{2}\log|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}}| \\ &\qquad = \frac{1}{2}\left(\log|\mathbf{A}| - \log|\sigma^2\mathbf{I}| - \log|\mathbf{A} + \sigma^{-2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}|\right) \\ &\qquad = \frac{1}{2}\left(\sum_{i=0}^{N}\log\alpha_i - N\log\sigma^2 + \log|\boldsymbol{\Sigma}|\right) \end{aligned} \quad (30)$$

where $\boldsymbol{\Sigma} = \left(\mathbf{A} + \sigma^{-2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}$ and we have used the determinant identity

$$|\mathbf{A}||\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}}| = |\sigma^2\mathbf{I}||\mathbf{A} + \sigma^{-2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}|.$$

Using the Woodbury inversion identity, the second term in Eq. 29 can be expressed as

$$\begin{aligned} &-\frac{1}{2}\mathbf{y}^{\mathsf{T}}(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathsf{T}})^{-1}\mathbf{y} \\ &\qquad = -\frac{1}{2}\mathbf{y}^{\mathsf{T}}\left[\sigma^{-2}\mathbf{I} - \sigma^{-2}\boldsymbol{\Phi}(\mathbf{A} + \sigma^{-2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\sigma^{-2}\right]\mathbf{y} \\ &\qquad = -\frac{\sigma^{-2}}{2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y} - \mathbf{y}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathsf{T}}\sigma^{-2}\mathbf{y}\right] \\ &\qquad = -\frac{1}{2}\sigma^{-2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y} - \mathbf{y}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\mu}\right] \\ &\qquad = -\frac{1}{2}\left[\sigma^{-2}\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2 + \sigma^{-2}\mathbf{y}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\mu} - \sigma^{-2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\mu}\right] \\ &\qquad = -\frac{1}{2}\left[\sigma^{-2}\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2 + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \sigma^{-2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\mu}\right] \\ &\qquad = -\frac{1}{2}\left[\sigma^{-2}\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2 + \boldsymbol{\mu}^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu}\right] \end{aligned} \quad (31)$$

where $\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{y}$.

Combining Eq. 30 and Eq. 31, the log-likelihood is

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\alpha},\sigma^2) = \frac{1}{2}\left(\sum_{i=0}^{N}\log\alpha_i - N\log\sigma^2 + \log|\boldsymbol{\Sigma}| - \sigma^{-2}\|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\mu}\|^2 - \boldsymbol{\mu}^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu}\right) \tag{32}$$

Let $p_i = \log\alpha_i$ so that $\alpha_i = e^{p_i}$ and that $\frac{\partial\alpha_i}{\partial p_i} = e^{p_i} = \alpha_i$. Setting

$$\frac{\partial\mathcal{L}(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)}{\partial p_i} = 0,$$

we have

$$\alpha_i(\mu_i^2 + \Sigma_{ii}) = 1, \tag{33}$$

where μ_i is the i-th component of $\boldsymbol{\mu}$ in Eq. 16 and Σ_{ii} is the i-th diagonal element of $\boldsymbol{\Sigma}$ in Eq. 16. Note that we have used the derivatives $\frac{\partial}{\partial p_i}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu} = \alpha_i\mu_i^2$ and

$$\begin{aligned}\frac{\partial\log|\boldsymbol{\Sigma}|}{\partial p_i} &= -\frac{\partial\log|\boldsymbol{\Sigma}^{-1}|}{\partial p_i} = -\mathrm{tr}\left\{\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial p_i}\right\} = -\mathrm{tr}\left\{\boldsymbol{\Sigma}\frac{\partial\mathbf{A}}{\partial p_i}\right\}\\ &= -\Sigma_{ii}\frac{\partial\alpha_i}{\partial p_i} = -\Sigma_{ii}\alpha_i.\end{aligned}$$

Define $\gamma_i = 1-\alpha_i\Sigma_{ii}$ and substitute $\Sigma_{ii} = (1-\gamma_i)/\alpha_i$ into Eq. 33, we obtain the update equation for α_i as follows:

$$\alpha_i^{\mathrm{new}} = \frac{\gamma_i}{\mu_i^2} \tag{34}$$

Let $q = \log\sigma^{-2}$ so that $\frac{\partial\sigma^{-2}}{\partial q} = \sigma^{-2}$. Then, we compute the derivative

$$\begin{aligned}\frac{\partial\mathcal{L}(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)}{\partial q} &= \frac{1}{2}\left[N - \sigma^{-2}\|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\mu}\|^2 - \mathrm{tr}\left\{\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial q}\right\}\right]\\ &= \frac{1}{2}\left[N - \sigma^{-2}\|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\mu}\|^2 - \mathrm{tr}\left\{\sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right\}\right]\\ &= \frac{1}{2}\left[N - \sigma^{-2}\|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\mu}\|^2 - \sum\nolimits_i\gamma_i\right]\end{aligned} \tag{35}$$

Setting $\frac{\partial\mathcal{L}(\mathbf{y}|\boldsymbol{\alpha},\sigma^2)}{\partial q} = 0$, we obtain

$$(\sigma^2)^{\mathrm{new}} = \frac{\|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{N - \sum_{i=0}^{N}\gamma_i}. \tag{36}$$

References

[1] Bishop, C., 2006. Pattern recognition and machine learning. springer, New York.

[2] Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. The journal of machine learning research 1, 211–244.