October 19, 2010

Relevance Vector Machines Explained

Tristan Fletcher

www.cs.ucl.ac.uk/staff/T.Fletcher/

Introduction

This document has been written in an attempt to make Tipping's [1] Relevance Vector Machines (RVM) as simple to understand as possible for those with minimal experience of Machine Learning. It assumes knowledge of probability in the areas of Bayes' theorem and Gaussian distributions including marginal and conditional Gaussian distributions. It also assumes familiarity with matrix differentiation, the vector representation of regression and kernel (basis) functions. These latter two areas are briefly covered in the author's similar paper on Support Vector Machines which can be found through the URL on the coverpage.

The document has been split into two main sections. The first introduces the problem that needs to be solved, namely maximising the posterior probability of regression target values over some hyperparameters. It then proceeds to derive the equations required to do this. Aside from the areas mentioned above where knowledge is assumed, every mathematical step is gone through. It is therefore hoped that there is no ambiguity over any of the details in this explanation, though this does make the description a little cumbersome. The second section therefore explains from an algorithmic viewpoint the iterations that would be required to actually apply the technique.

Aside from Tipping [1], the majority of this document is based on work by MacKay [2], [3], [4], [5] and Bishop [6].

Notation

With the view of making this description as explicit (in the mathematical sense) as possible, it is worth introducing some of the notation that will be used:

- $P(A|B,C)$ is the probability of A given B and C. Note that using this notation, different representations of the parameters will not alter this probability, e.g. $P(A|B) \equiv P(A|B^{-1})$. Furthermore, for the sake of simplicity, occasionally some of the elements in the probabilistic notation will be omitted where they are not relevant, e.g. $P(A|B) \equiv P(A|B, C, D\ etc)$.

- $X \sim N(\mu, \sigma^2)$ is used to signify that X is normally distributed with mean μ and variance σ^2.

- Bold font is used to represent vectors and matrices.

1 Theory

1.1 Evidence Approximation Theory

Linear regression problems are generally based on finding the parameter vector $\mathbf{w}$ and the offset c so that we can predict y for an unknown input $\mathbf{x}$ ($\mathbf{x} \in \Re^M$):

$$y = \mathbf{w}^T \mathbf{x} + c$$

In practice we usually incorporate the offset c into $\mathbf{w}$. If there is a non-linear relationship between $\mathbf{x}$ and y then a basis function can be used:

$$y = \mathbf{w}^T \phi(\mathbf{x})$$

where $\mathbf{x} \mapsto \phi(\mathbf{x})$ is a non-linear mapping (i.e. basis function).

When attempting to calculate $\mathbf{w}$ from our our training examples, we assume that each target t_i is representative of the true model y_i, but with the addition of noise ϵ_i:

$$\begin{aligned} t_i &= y_i + \epsilon_i \\ &= \mathbf{w}^T \phi(\mathbf{x}_i) + \epsilon_i \end{aligned}$$

where ϵ_i are assumed to be independent samples from a Gaussian noise process with zero mean and variance σ^2, i.e. $\epsilon_i \sim N(0, \sigma^2) \ \forall_i$. This means that:

$$\begin{aligned} P(t_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \sim N(y_i, \sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(t_i - y_i)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(t_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2\right\} \end{aligned}$$

Looking at N training points simultaneously, so that the vector $\mathbf{t}$ represents all the individual training points t_i and the $N \times M$ design matrix $\mathbf{\Phi}$ is constructed such that the ith row represents the vector $\phi(\mathbf{x}_i)$, we have:

$$\begin{aligned} P(\mathbf{t}|\mathbf{x}_i, \mathbf{w}, \sigma^2) &= \prod_{i=1}^{N} N(\mathbf{w}^T\phi(\mathbf{x}_i), \sigma^2) \\ &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(t_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right\|^2\right\} \end{aligned}$$

When attempting to learn the relationship between $\mathbf{x}$ and y, we wish to constrain complexity and hence the growth of the weights $\mathbf{w}$ and do this by defining an explicit *prior* probability distribution on $\mathbf{w}$. Our preference for smoother and therefore less complex functions is encoded by using a zero-mean Gaussian prior over $\mathbf{w}$:

$$P(\mathbf{w}|\alpha_i) \sim N(0, \alpha_i^{-1})$$

where we have used α_i to describe the inverse variance (i.e. precision) of each w_i. If once again we look at all N points simultaneously, so that the ith element in the vector $\boldsymbol{\alpha}$ represents α_i, we have:

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} N(0, \alpha_i^{-1})$$

This means that there is an individual hyperparameter α_i associated with each weight, modifying the strength of the prior thereon.

The *posterior* probability over all the unknown parameters, given the data, is expressed as $P(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t})$. We are trying to find the $\mathbf{w}, \boldsymbol{\alpha}$ and σ^2 which maximise this posterior probability. We can decompose the posterior:

$$P(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t}) = P(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)P(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \tag{1.1}$$

Substituting β^{-1} for σ^2 to make the maths appear less cluttered, the first part of (1.1) can be expressed:

$$P(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \beta) \sim N(\mathbf{m}, \boldsymbol{\Sigma}) \tag{1.2}$$

where the mean $\mathbf{m}$ and the covariance $\boldsymbol{\Sigma}$ are given by:

$$\mathbf{m} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t} \tag{1.3}$$

$$\boldsymbol{\Sigma} = (\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} \tag{1.4}$$

and $\mathbf{A} = diag(\boldsymbol{\alpha})$.

The method for arriving at (1.2), (1.3) and (1.4), relating to conditional Gaussian distributions, lies outside the scope of this document.

In order to evaluate $\mathbf{m}$ and $\boldsymbol{\Sigma}$ we need to find the hyperparameters ($\boldsymbol{\alpha}$ and β) which maximise the second part of (1.1), which we decompose:

$$P(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \propto P(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)P(\boldsymbol{\alpha})P(\sigma^2)$$

We will assume uniform hyperpriors and hence ignore $P(\boldsymbol{\alpha})$ and $P(\sigma^2)$. Our problem is now to maximise the *evidence*:

$$P(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) \equiv P(\mathbf{t}|\boldsymbol{\alpha}, \beta) = \int P(\mathbf{t}|\mathbf{w}, \beta)P(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \tag{1.5}$$

Looking at the first component of the equation:

$$\begin{aligned}
P(\mathbf{t}|\mathbf{w},\beta) &= \prod_{i=1}^{N} N(\mathbf{y},\beta^{-1}) \\
&= \left(\frac{2\pi}{\beta}\right)^{-\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2\right\} \qquad (1.6)
\end{aligned}$$

And then at the second (where M is the dimensionality of $\mathbf{x}$):

$$\begin{aligned}
P(\mathbf{w}|\boldsymbol{\alpha}) &= \prod_{i=1}^{M} N(0,\alpha_i^{-1}) \\
&= \prod_{i=1}^{M}(2\pi\alpha_i^{-1})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\alpha_i\mathbf{w}^2\right\} \\
&= (2\pi)^{-\frac{M}{2}} \prod_{i=1}^{M}\alpha_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}\right\} \qquad (1.7)
\end{aligned}$$

Substituting (1.6) and (1.7) into (1.5):

$$\begin{aligned}
P(\mathbf{t}|\boldsymbol{\alpha},\beta) &= \int \left(\frac{2\pi}{\beta}\right)^{-\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2\right\} (2\pi)^{-\frac{M}{2}} \prod_{i=1}^{M}\alpha_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}\right\} d\mathbf{w} \\
&= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{1}{2\pi}\right)^{\frac{M}{2}} \prod_{i=1}^{M}\alpha_i^{\frac{1}{2}} \int \exp -\left\{\frac{\beta}{2}\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2 + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}\right\} d\mathbf{w} \\
&\qquad (1.8)
\end{aligned}$$

In order to simplify (1.8), we create a definition to represent the integrand:

$$E(\mathbf{w}) = \frac{\beta}{2}\left\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right\|^2 + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} \qquad (1.9)$$

This means that:

$$P(\mathbf{t}|\boldsymbol{\alpha},\beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{1}{2\pi}\right)^{\frac{M}{2}} \prod_{i=1}^{M}\alpha_i^{\frac{1}{2}} \int \exp\left\{-E(\mathbf{w})\right\} d\mathbf{w}$$

Expanding out (1.9) we get:

$$\begin{aligned}
E(\mathbf{w}) &= \frac{\beta}{2}(\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\boldsymbol{\Phi}\mathbf{w} + \mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}) + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} \\
&= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\boldsymbol{\Phi}\mathbf{w} + \beta\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} + \mathbf{w}^T\mathbf{A}\mathbf{w})
\end{aligned}$$

Substituting in (1.4) and then using $\mathbf{I} = \mathbf{\Sigma}^{-1}\mathbf{\Sigma}$:

$$\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\mathbf{\Phi}\mathbf{w} + \mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}) \\
&= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{w} + \mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w})
\end{aligned}$$

Substituting in (1.3):

$$\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t} - 2\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w} + \mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m} - \mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m}) \\
&= E(\mathbf{t}) + \frac{1}{2}(\mathbf{w} - \mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{w} - \mathbf{m})
\end{aligned}$$

where

$$E(\mathbf{t}) = \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t} - \mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m})$$

Our integrand from (1.8) now becomes:

$$\int \exp\{-E(\mathbf{w})\}\, d\mathbf{w} = \exp\{-E(\mathbf{t})\}\,(2\pi)^{\frac{M}{2}}\,|\mathbf{\Sigma}|^{\frac{1}{2}}$$

Substituting this back in, gives us:

$$P(\mathbf{t}|\boldsymbol{\alpha},\beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{1}{2\pi}\right)^{\frac{M}{2}} \prod_{i=1}^{M} \alpha_i^{\frac{1}{2}} \exp\{-E(\mathbf{t})\}\,(2\pi)^{\frac{M}{2}}\,|\mathbf{\Sigma}|^{\frac{1}{2}}$$

This is known as our *marginal likelihood* and taking logs, gives us our *log marginal likelihood*:

$$\ln P(\mathbf{t}|\boldsymbol{\alpha},\beta) = \frac{N}{2}\ln\beta - E(\mathbf{t}) - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{M}\ln\alpha_i \quad (1.10)$$

It is this equation we need to maximise with respect to $\boldsymbol{\alpha}$ and β, a process known as the *evidence approximation* procedure.

1.2 Evidence Approximation Procedure

In order to maximise our log marginal likelihood, we start by taking derivatives of (1.10) with respect to α_i and setting these to zero:

$$\begin{aligned}
\frac{d}{d\alpha_i}\ln P(\mathbf{t}|\boldsymbol{\alpha},\beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2 = 0 \\
\Rightarrow \alpha_i = \frac{1 - \alpha_i\Sigma_{ii}}{m_i^2}
\end{aligned}$$

Substituting in $\gamma_i = 1 - \alpha_i \Sigma_{ii}$ our recursive definition for the α_i which maximise (1.10) can be expressed more elegantly as:

$$\alpha_i = \frac{\gamma_i}{m_i^2}$$

We now need to differentiate (1.10) with respect to β and set these derivatives to zero:

$$\frac{d}{d\beta} \ln P(\mathbf{t}|\boldsymbol{\alpha}, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - Tr\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right]\right) = 0 \qquad (1.11)$$

In order to solve this, we first simplify the argument of the trace operator $Tr[\bullet]$:

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= \boldsymbol{\Sigma}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\beta + \mathbf{A}\right)\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= \left(\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\beta + \mathbf{A}\right)\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= \left(\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}\right)\beta^{-1}
\end{aligned}$$

Substituting this back into (1.11):

$$\begin{aligned}
&\frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - Tr\left[\frac{\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}}{\beta}\right]\right) = 0 \\
&\Rightarrow \frac{N}{\beta} - Tr\left[\frac{\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}}{\beta}\right] = \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 \\
&\Rightarrow \frac{1}{\beta}\left(N - Tr\left[\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}\right]\right) = \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 \\
&\Rightarrow \frac{1}{\beta} = \frac{\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2}{N - Tr\left[\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma}\right]} \\
&\Rightarrow \beta = \frac{N - \sum_i \gamma_i}{\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2}
\end{aligned}$$

The α_i and β which maximise our marginal likelihood are then found iteratively by setting $\boldsymbol{\alpha}$ and β to initial values, finding values for $\mathbf{m}$ and $\boldsymbol{\Sigma}$ from (1.3) and (1.4), using these to calculate new estimates for $\boldsymbol{\alpha}$ and β and repeating this process until a convergence criteria is met.

We will then be left with values for $\boldsymbol{\alpha}$ and β which maximise our marginal likelihood and which we can use to evaluate our predictive distribution over t for a new input $\mathbf{x}'$:

$$\begin{aligned}
P(t|\mathbf{x}', \boldsymbol{\alpha}, \beta) &= \int P(t|\mathbf{w}, \beta)P(\mathbf{w}|, \boldsymbol{\alpha}, \beta)d\mathbf{w} \\
&= N(\mathbf{m}^T\phi(\mathbf{x}'), \sigma^2(\mathbf{x}'))
\end{aligned}$$

This means that our estimate for t is the mean of the above distribution $\mathbf{m}^T\phi(\mathbf{x}')$.

Our confidence in our prediction is determined by the variance of this distribution $\sigma^2(\mathbf{x}')$ which is given by:

$$\sigma^2(\mathbf{x}') = \beta^{-1} + \phi(\mathbf{x}')^T\mathbf{\Sigma}\phi(\mathbf{x}')$$

1.3 Automatic Relevance Determination

Whilst carrying out the evidence approximation procedure described above, many of the α_i will tend to infinity. This has implications for the variance $\mathbf{\Sigma}$ and mean $\mathbf{m}$ of the posterior distribution over the corresponding weights in (1.2):

$$\begin{aligned} &\lim_{\alpha_i\to\infty} \mathbf{\Sigma} = \lim_{\alpha_i\to\infty} (\mathbf{A} + \beta\mathbf{\Phi}^T\mathbf{\Phi})^{-1} = 0 \\ \Rightarrow &\lim_{\alpha_i\to\infty} \mathbf{m} = \lim_{\alpha_i\to\infty} \beta\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t} = 0 \end{aligned}$$

This means that each w_i that such α_i relate to will be distributed $\alpha_i \sim N(0,0)$, i.e. will be equal to zero. The corresponding basis functions, $\phi(\mathbf{x}_i)$ should therefore be pruned from the overall design matrix $\mathbf{\Phi}$ each iteration.

The $\mathbf{x}_i$ corresponding to the remaining non-zero weights after pruning are called *relevance vectors* and are analogous to the support vectors of an SVM.

2 Application

The RVM process is an iterative one and involves repeatedly re-estimating $\boldsymbol{\alpha}$ and β until a stopping condition is met. The steps are as follows:

1. Select a suitable kernel function for the data set and relevant parameters. Use this kernel function to create the design matrix $\mathbf{\Phi}$.

2. Establish a suitable convergence criteria for $\boldsymbol{\alpha}$ and β, e.g. a threshold value for change δ_{Thresh} between one iteration's estimation of $\boldsymbol{\alpha}$ and the next $\delta = \sum_{i=1} \alpha_i^{n+1} - \alpha_i^n$ so that re-estimation will stop when $\delta < \delta_{Thresh}$.

3. Establish a threshold value α_{Thresh} which it is assumed an α_i is tending to infinity upon reaching it.

4. Choose starting values for $\boldsymbol{\alpha}$ and β.

5. Calculate $\mathbf{m} = \beta\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}$ and $\mathbf{\Sigma} = (\mathbf{A} + \beta\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$.

6. Update $\alpha_i = \frac{\gamma_i}{m_i^2}$ and $\beta = \frac{N-\sum_i \gamma_i}{||\mathbf{t}-\mathbf{\Phi}\mathbf{m}||^2}$.

7. Prune the α_i and corresponding basis functions where $\alpha_i > \alpha_{Thresh}$.

8. Repeat (5) to (7) until the convergence criteria is met.

Our hyperparameter values $\boldsymbol{\alpha}$ and β which result from the above procedure are those that maximise our marginal likelihood and hence are those used when making a new estimate of a target value t for a new input $\mathbf{x}'$:

$$t = \mathbf{m}^T\phi(\mathbf{x}') \tag{2.1}$$

The variance relating to our confidence in this estimate is given by:

$$\sigma^2(\mathbf{x}') = \beta^{-1} + \phi(\mathbf{x}')^T\mathbf{\Sigma}\phi(\mathbf{x}') \tag{2.2}$$

References

[1] M. E. Tipping, *J. Mach. Learn. Res.* **1**, 211 (2001).

[2] D. J. C. Mackay, *Neural Computation* **4**, 415 (1992).

[3] D. J. C. Mackay, *Neural Computation* **4**, 448 (1992).

[4] D. J. C. Mackay, *Bayesian methods for backprop networks*, chap. 6, pp. 211–254. Springer (1994).

[5] D. J. C. Mackay, C. Laboratory, *Neural Computation* **11**, 1035 (1999).

[6] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer (2006).