Patt. Rec. and Mach. Learning
Ch. 7: Sparse Kernel Machines

Marcin Marszalek & Pierre Mahé

January 11, 2008

Part 2: Relevance Vector Machines

Motivations

In the spirit of SVMs, define sparse kernel models:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{N} w_i k(\mathbf{x}, \mathbf{x}_i)$$

where $\{\mathbf{x}_i, t_i\}_{i=1:N}$ is the training set and many w_i's are equal to 0,

...but cope with the limitations of the SVM algorithm:

- no probabilistic interpretation
- difficulty of choosing the regularization parameter C (soft-margin)
- restricted to positive (semi) definite kernels
- no natural extension to the multiclass case
- (models are not so sparse)

Definition

The Relevance Vector Machine (RVM) is an instance of the Bayesian linear and logistic regression models where:

1. the basis functions are centered on the training points $\mathbf{x}_i$, that is $\phi(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_N]$
   - gives the "SVM-like" formulation: $y(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{N} w_i k(\mathbf{x}, \mathbf{x}_i)$
2. the following prior over the weights is used:

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

with $\mathbf{A} = \text{diag}(\alpha_1^{-1}, \ldots, \alpha_N^{-1})$.

- a vector $\boldsymbol{\alpha}$ of alpha parameters (one per training point) instead of a single α parameter in the "standard case": $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$

Outline

1. RVM for regression
   - "basics" of Bayesian linear regression
   - RVM solution
   - Intuition on sparsity
2. RVM for classification
   - "basics" of Bayesian logistic regression
   - RVM solution
3. Illustrations/remarks/conclusion

From least-squares and SVM to Bayesian models 1/2

We consider linear models of regression: $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x})$

Regularized/penalized least squares and SVM solutions:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}),$$

where:

- $E_D(\mathbf{w}) = \sum_{i=1}^{N} L(t_i, y(\mathbf{x}_i, \mathbf{w}))$ is an empirical error
- $E_W(\mathbf{w})$ is a regularization term (typically $||\mathbf{w}||^2$)

From least-squares and SVM to Bayesian models 2/2

Bayesian approach:

- work in a probabilistic framework:
  - $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$
- introduce a prior over $\mathbf{w}$:
  - typically: favor small values by $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$
- from the training data $(\mathbf{X}, \mathbf{t})$, compute the posterior of $\mathbf{w}$:
  - $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \times p(\mathbf{w}|\alpha)$
- for a new data $\mathbf{x}$, predict t according to the predictive distribution:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) d\mathbf{w}$$

Note:

- in regularized least squares, $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) = \delta(\mathbf{w} = \mathbf{w}_{MAP})$
- being "fully" or "truly" Bayesian: don't pick a value of $\mathbf{w}$, average over all possible values

Bayesian regression in practice 1/3

How to compute the predictive distribution?

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) d\mathbf{w}$$

1. Setting the regression model $p(t|\mathbf{x}, \mathbf{w}, \beta)$ and the prior $p(\mathbf{w}|\alpha)$ to Gaussians gives conjugate likelihood/prior
   - the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta)$ is Gaussian and given in closed form (see eqs 3.49 and 2.116)
2. As a result, the predictive distribution is the convolution of two Gaussians
   - it is also a Gaussian and is given in closed form (see eqs 3.57 and 2.115)

$\Rightarrow$ we can make probabilistic predictions and quantify their uncertainty (the variance of the predictive distribution depends on $\mathbf{x}$)

Bayesian regression in practice 2/3

However, the process still depends on fixed parameters α and β:

- α from the prior $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$
- β from the regression model $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$

$\Rightarrow$ the ratio α/β plays the role of a regularization parameter in the posterior (see eq 3.55).
$\Rightarrow$choosing α and β amounts to choosing λ (or C) in regularised least squares and SVMs (typically done by cross-validation)

Alternative "truly, truly" Bayesian approach $\Rightarrow$ average them out:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \int p(t|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta)p(\alpha, \beta|\mathbf{X}, \mathbf{t})d\mathbf{w}d\alpha d\beta$$

$\Rightarrow$ however, becomes untractable (need to compute $p(\mathbf{t}|\mathbf{X})$)

Bayesian regression in practice 3/3

Need to approximate the distribution:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) p(\alpha, \beta|\mathbf{X}, \mathbf{t}) d\mathbf{w} d\alpha d\beta$$

Evidence approximation method: choose single values $(\hat{\alpha}, \hat{\beta})$

1. assume $p(\alpha, \beta|\mathbf{X}, \mathbf{t})$ is sharply peaked around $\hat{\alpha}$ and $\hat{\beta}$
   - then $p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) \sim p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \hat{\alpha}, \hat{\beta})$
   - $\Rightarrow$ standard "predictive distribution", with closed-form solution
   - NB: similar to the ("not fully" Bayesian) MAP approach for estimation of $\mathbf{w}$
2. to get $(\hat{\alpha}, \hat{\beta})$, assume flat/uninformative priors $p(\alpha)$ and $p(\beta)$
   - maximizing $p(\alpha, \beta|\mathbf{X}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, \alpha, \beta) p(\alpha) p(\beta)$ is then equivalent to maximizing $p(\mathbf{t}|\mathbf{X}, \alpha, \beta)$
   - $p(\mathbf{t}|\mathbf{X}, \alpha, \beta)$ is called the marginal likelihood:

$$p(\mathbf{t}|\mathbf{X}, \alpha, \beta) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha) d\mathbf{w}$$

In the end...

- With the evidence approximation method, what we need is to maximize the marginal likelihood:

$$p(\mathbf{t}|\mathbf{X}, \alpha, \beta) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha) d\mathbf{w}$$

  $\Rightarrow$ this gives us $\hat{\alpha}$ and $\hat{\beta}$ from which we can define the approximate distribution $p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) \sim p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \hat{\alpha}, \hat{\beta})$
- Remarks:
  - flat/uninformative priors are actually justified in the sense that they define a scale-invariant model
  - maximizing the marginal likelihood allows to automatically select the appropriate complexity on the basis of the training data only
- For the "standard case", this is detailed in Sections 3.5.1/2

Coming back to the RVM model

Recall that he RVM is an instance of the previous model where:

1. the basis functions are centered on the training points $\mathbf{x}_i$, that is $\phi(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_N]$
   - gives the "SVM-like" form to $y(\mathbf{x}, \mathbf{w})$ as $\sum_{i=1}^{N} w_i k(\mathbf{x}, \mathbf{x}_i)$
2. the following prior over the weights is used:

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

with $\mathbf{A} = \text{diag}(\alpha_1^{-1}, \ldots, \alpha_N^{-1})$.

- a vector $\boldsymbol{\alpha}$ of alpha parameters (one per training point) instead of a single α parameter in the "standard case": $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$

What is changed (1/2)?

Because the prior is still Gaussian, apparently not much:

- we still have conjugate likelihood/priors, and the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \boldsymbol{\alpha}, \beta)$ is available in closed form (see eqs 7.82/7.83)
- as a result the predictive distribution:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}, \beta) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \boldsymbol{\alpha}, \beta) d\mathbf{w}$$

  is still a convolution of Gaussian and is available in closed form too (see eq 7.90)

Conclusion: we simply need to maximize the marginal likelihood $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)$ in order to get $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$.

$\Rightarrow$ Although it now depends on $N + 1$ (instead of 2) variables, the solution can be derived easily from the "standard case" (see eqs. 7.87/88).

What is changed (2/2)?

In the end: we obtain the same expression for the approximate predictive distribution $p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \hat{\boldsymbol{\alpha}}, \hat{\beta})$

- only modification: a matrix $\alpha\mathbf{I}$ is replaced by $\text{diag}(\alpha_1, \ldots, \alpha_N)$.

Striking point: the optimization drives many components of $\hat{\boldsymbol{\alpha}}$ to very large values

- as a result, the corresponding entries of $\mathbf{w}$ have a posterior distribution centered on 0, with a variance of 0
- thus, they play no role in the model and can be removed, which leads to a sparse model
- (note: this is an exemple of automatic relevance determination)

$\Rightarrow$ What is going on?

What is going on (1/2)?

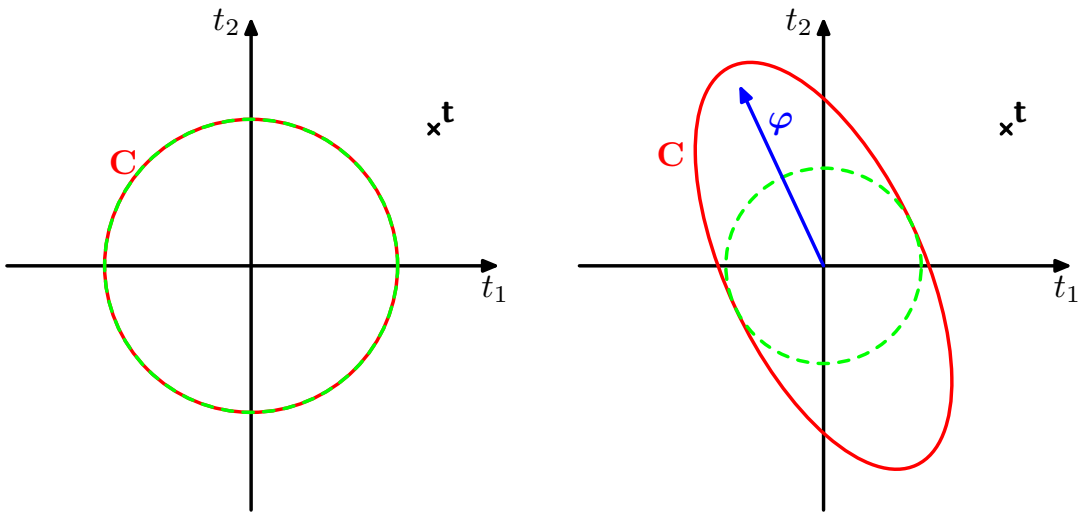
More precisely, the expression of the marginal likelihood is:

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}), \text{ with } \mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T.$$

For a sample $(\mathbf{X}, \mathbf{t})$ of size 2, with a single basis $\phi(\mathbf{x})$, we have:

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\varphi}\boldsymbol{\varphi}^T, \text{ with } \boldsymbol{\varphi} = [\phi(\mathbf{x}_1)\ \phi(\mathbf{x}_2)]^T.$$

Evidence approximation $\Rightarrow$ find α maximizing probability at $\mathbf{t} = [t_1\ t_2]^T$:



$\alpha = \infty$ $\qquad$ $\alpha = \alpha_0$

NB: $|\mathbf{C}|$ is kept constant ; red curve = unit Mahalanobis distance $\mathbf{t}^T\mathbf{C}\mathbf{t}$

What is going on (2/2)?

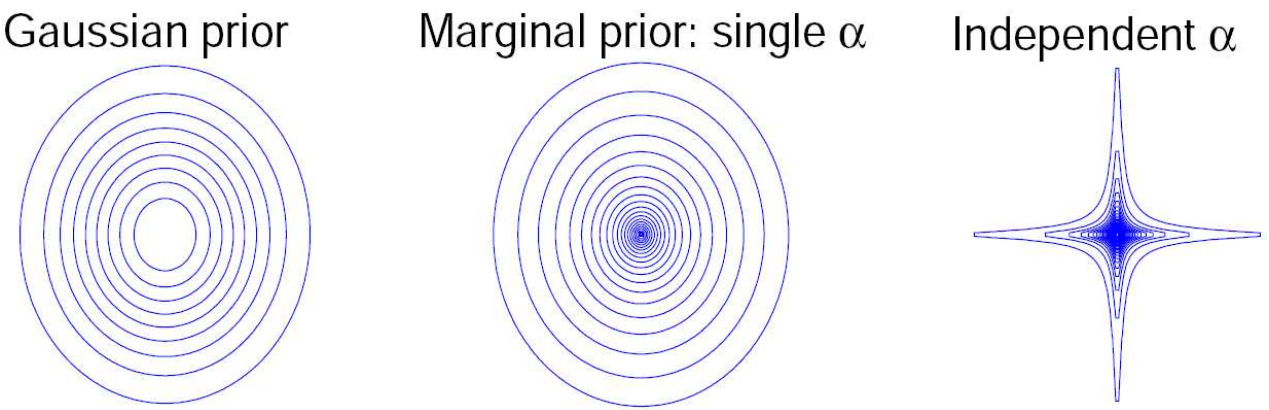
The couple $\big(p(\mathbf{w}|\boldsymbol{\alpha}), p(\boldsymbol{\alpha})\big)$ defines a hierarchical prior over $\mathbf{w}$

The "true" prior over $\mathbf{w}$ is actually given by marginalizing $\boldsymbol{\alpha}$:

$$p(\mathbf{w}) = \int p(\mathbf{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}$$

$\Rightarrow$ for RVM, this "marginal prior" decomposes as a product of Student distributions.

Illustration (for $\mathbf{w} = [w_1\ w_2]$):



Outline

1. RVM for regression
   - "basics" of Bayesian linear regression
   - RVM solution
   - Intuition on sparsity
     - (skipped: detailed analysis of sparsity)
2. RVM for classification
   - "basics" of Bayesian logistic regression
   - RVM solution
3. Illustrations/remarks/conclusion

Logistic regression

Binary classification: target variable $t \in \{0, 1\}$

Logisitic regression:

$$p(C_1|\mathbf{x}, \mathbf{w}) = \sigma(y(\mathbf{x}, \mathbf{w})),$$

with $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$ and $\sigma(x) = 1/\big((1 + \exp(-x)\big)$

Solution (regularized):

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

with:

- $E_D(\mathbf{w}) = -p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = -$log-likelihood
- $E_W(\mathbf{w})$ is a regularization term (e.g., $||w||^2$)

(Note: no closed form solution, but unique minimum – sec 4.3.3)

Bayesian logistic regression

Similar process:

- introduce a prior over $\mathbf{w}$ (e.g., $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$
- compute the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha)$
- compute the predictive distribution by averaging out $\mathbf{w}$

However: the model $p(t|\mathbf{x}, \mathbf{w})$ is not Gaussian anymore, so we don't have closed form solution for the posterior and the predictive distribution

$\Rightarrow$ As a result, the procedure is more complex

- detailed in Section 4.5
- in particular, the Laplace approximation can be used to approximate the posterior by a Gaussian
- the problem therefore boils down to averaging the logistic model w.r.t. such a Gaussian distribution

RVM for binary classification

$\Rightarrow$ An instance of Bayesian logistic regression, with the prior:

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

with $\mathbf{A} = \text{diag}(\alpha_1^{-1}, \ldots, \alpha_N^{-1})$.

After some derivations (...) we get:

- the same expression – as in the regression case – for the marginal likelihood $p(\mathbf{t}|\boldsymbol{\alpha}, \beta)$
- as a result, the same "sparsity promoting" procedure for getting the $\boldsymbol{\alpha}$

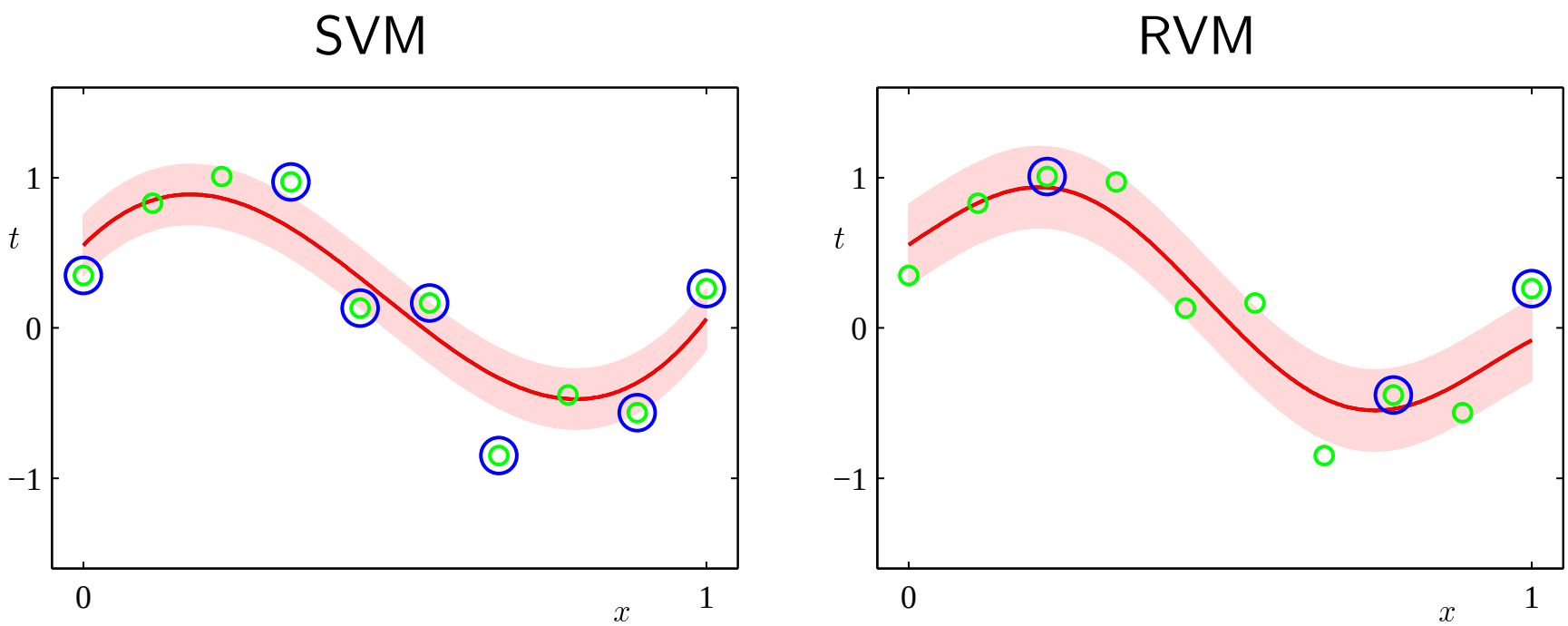
(Note: similarly to the standard logistic regression, the procedure can "readily" be extended to the multiclass case using the softmax instead of the logistic function).

Outline

1. RVM for regression
   - "basics" of Bayesian linear regression
   - RVM solution
   - Intuition on sparsity
     - (skipped: detailed analysis of sparsity)
2. RVM for classification
   - "basics" of Bayesian logistic regression
   - RVM solution
3. Illustrations/remarks/conclusion

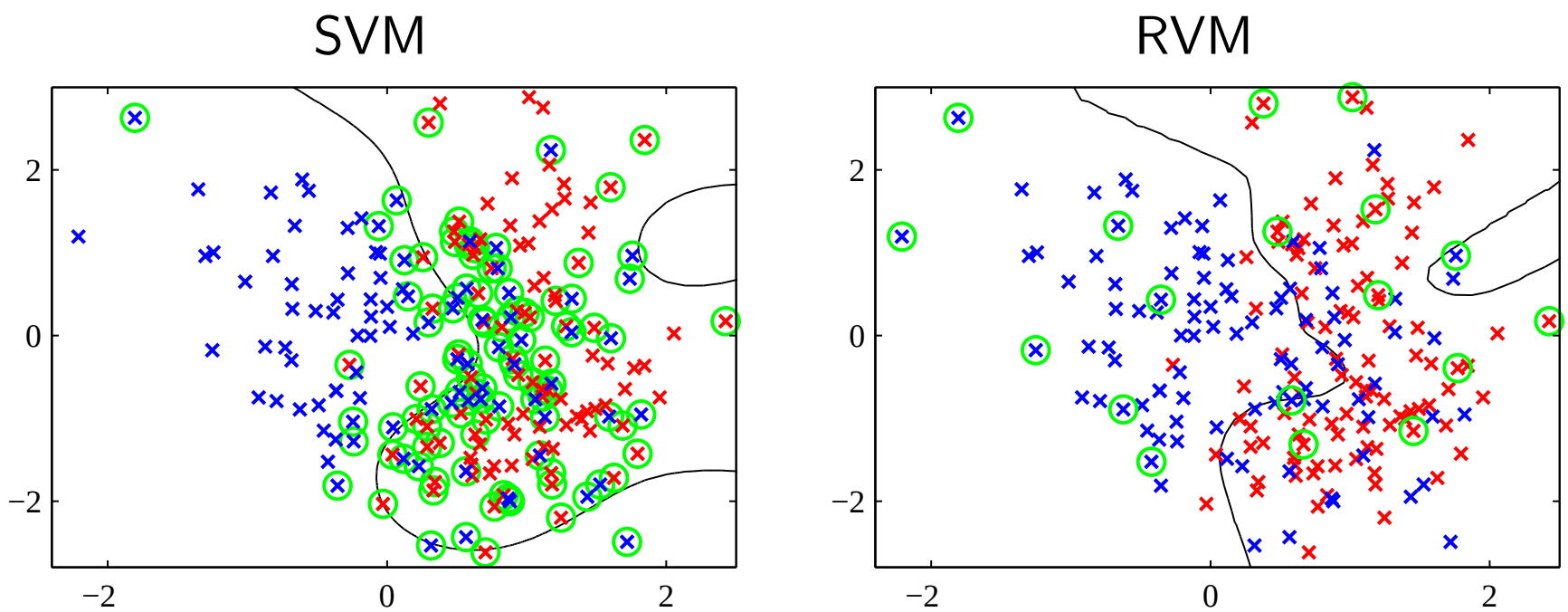
Comparaison RVM and SVM 1/2

Regression:



Comparaison RVM and SVM 2/2

Classification:



Conclusion

For comparable performance RVMs seem to give sparser models than SVMs and gives a measure of confidence in the prediction

Moreover, the mechanism is very general:

- can be applied to regression, binary and multiclass classification
- can be applied with any type of basis functions (not necessarily data-centered PSD kernels)

However: comes at the price of a more complex process...

- optimization of a non-convex function
- cubic versus quadratic complexity w.r.t. N

... which is compensated by the fact that:

- models are faster at test time (because sparser)
- the model complexity is automatically selected

Pointers

JMLR paper:
Sparse Bayesian Learing and the Relevance Vector Machine, Michael Tipping, 2001.

Tutorial:
Bayesian Inference: An Introduction to Principles and Practice in Machine Learning, Michael Tipping, 2004.