PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 7: SPARSE KERNEL MACHINE

Outlines

- Support Vector Machines
- SVM & Logistic Regression
- SVM for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

Support Vector Machines

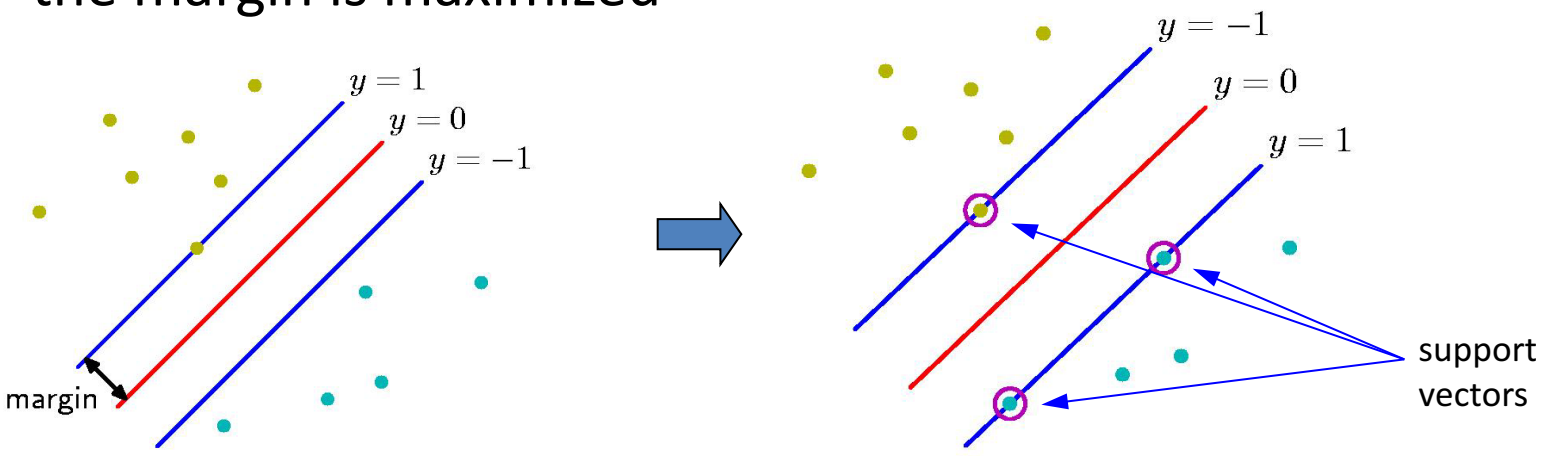
Problem settings

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

- Two-class classification using linear models
- Assume that training data set is linearly separable

Support vector machine approaches

- The decision boundary is chosen to be the one for which the margin is maximized



Maximum Margin Solution

For all data points, $t_n y(\mathbf{x}_n) > 0$

The distance of a point to the decision surface

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}$$

The maximum margin solution

$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \right] \right\}$$

$$\Rightarrow \arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1, \; n = 1, \ldots, N$$

Dual Representation

Introducing Lagrange multipliers,

→ **Find Appendix E for more details**

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) - 1 \right\}$$

Min. points satisfy the derivatives of *L* w.r.t. **w** and *b* equal 0

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n), \quad 0 = \sum_{n=1}^{N} a_n t_n$$

Dual representation

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

$$\text{subject to } a_n \geq 0, \ n = 1, \ldots, N$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

Classifying New Data

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \quad \Rightarrow \quad y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$$

Optimization subjects to

$$a_n \geq 0$$
$$t_n y(\mathbf{x}_n) - 1 \geq 0$$
$$a_n \{t_n y(\mathbf{x}_n) - 1\} = 0$$

Karush-Kuhn-Tucker (KKT) Conditions → Appendix E

$a_n = 0$ or $t_n y(\mathbf{x}_n) = 1$: support vectors

Found $\mathbf{a}$ by solving a quadratic programming problem

$$b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

→ $O(N^3)$

Example of Separable Data Classification

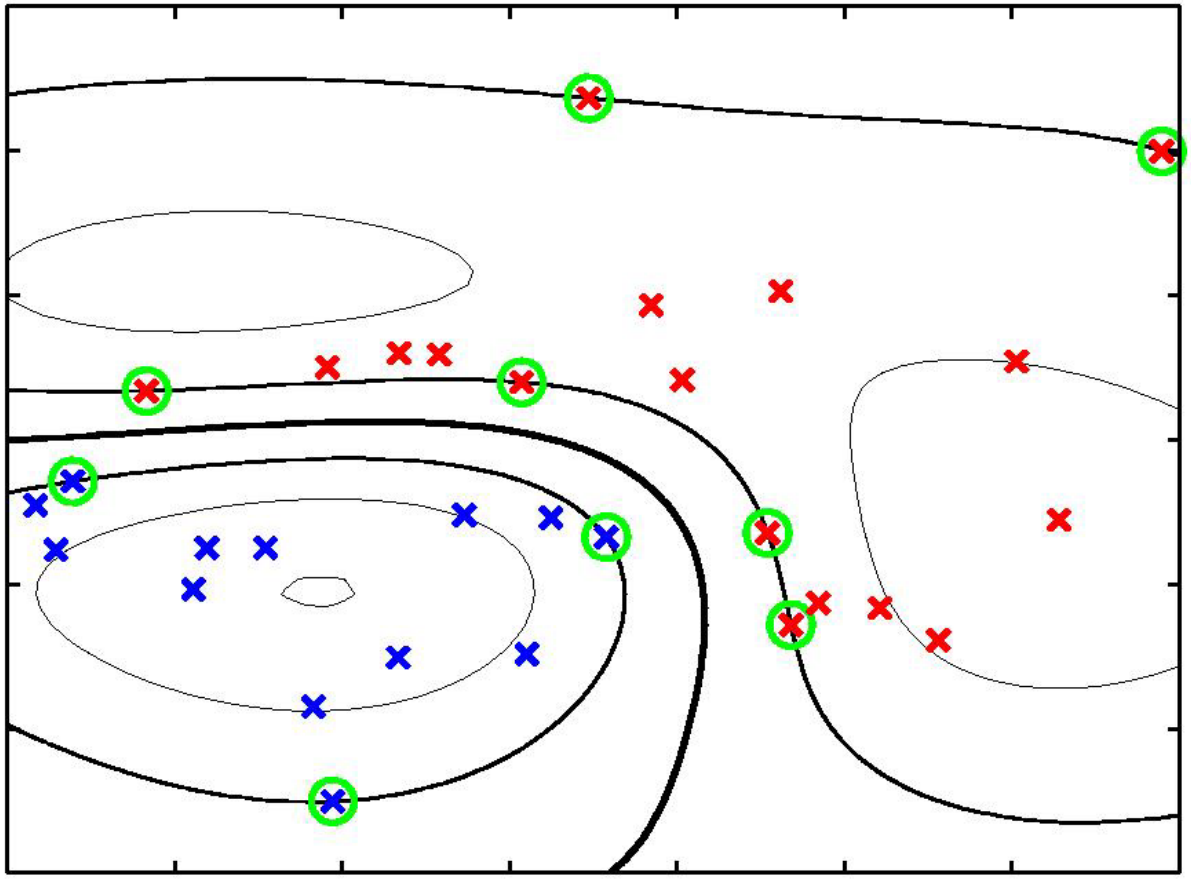
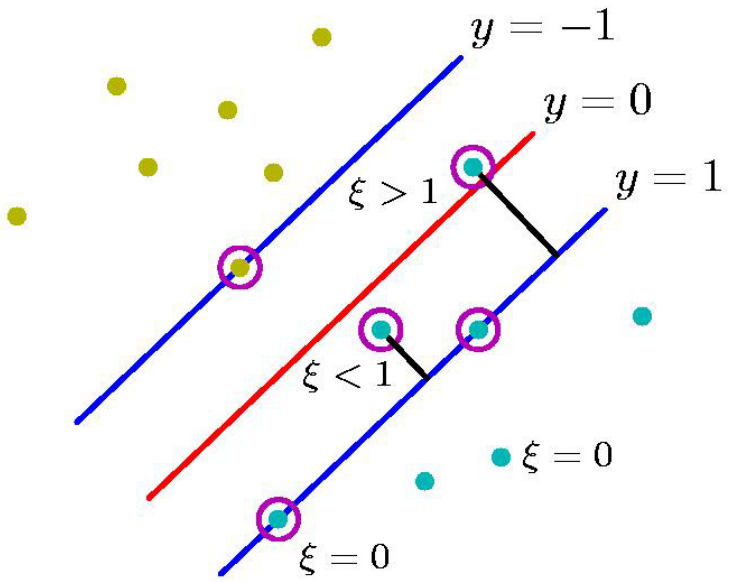


Figure 7.2

Overlapping Class Distributions

Allow some misclassified examples → soft margin

Introduce slack variables $\xi_n \geq 0, \; n = 1, \ldots, N$



$$t_n y(\mathbf{x}_n) = t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 \quad \Rightarrow \quad t_n y(\mathbf{x}_n) \geq 1 - \xi_n$$

Soft Margin Solution

Minimize $C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\mathbf{w}\|^2$

$C > 0$: trade-off between minimizing training errors and controlling model complexity

$$L(\mathbf{w},b,\mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

KKT conditions:

$$a_n \geq 0$$

$$t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0$$

$$a_n\left(t_n y(\mathbf{x}_n) - 1 + \xi_n\right) = 0$$

$$\mu_n \geq 0$$

$$\xi_n \geq 0$$

$$\mu_n\xi_n = 0$$

$a_n = 0$

or

$t_n y(\mathbf{x}_n) = 1 - \xi_n$

: support vectors

Dual Representation

Dual representation

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

$$\text{subject to } 0 \le a_n \le C,\ n = 1,\ldots,N$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

Classifying new data and obtaining *b* (➔ hard margin classifiers)

Alternative Formulation

v-SVM (Schölkopf *et al*., 2000)

$$\widetilde{L}(\mathbf{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

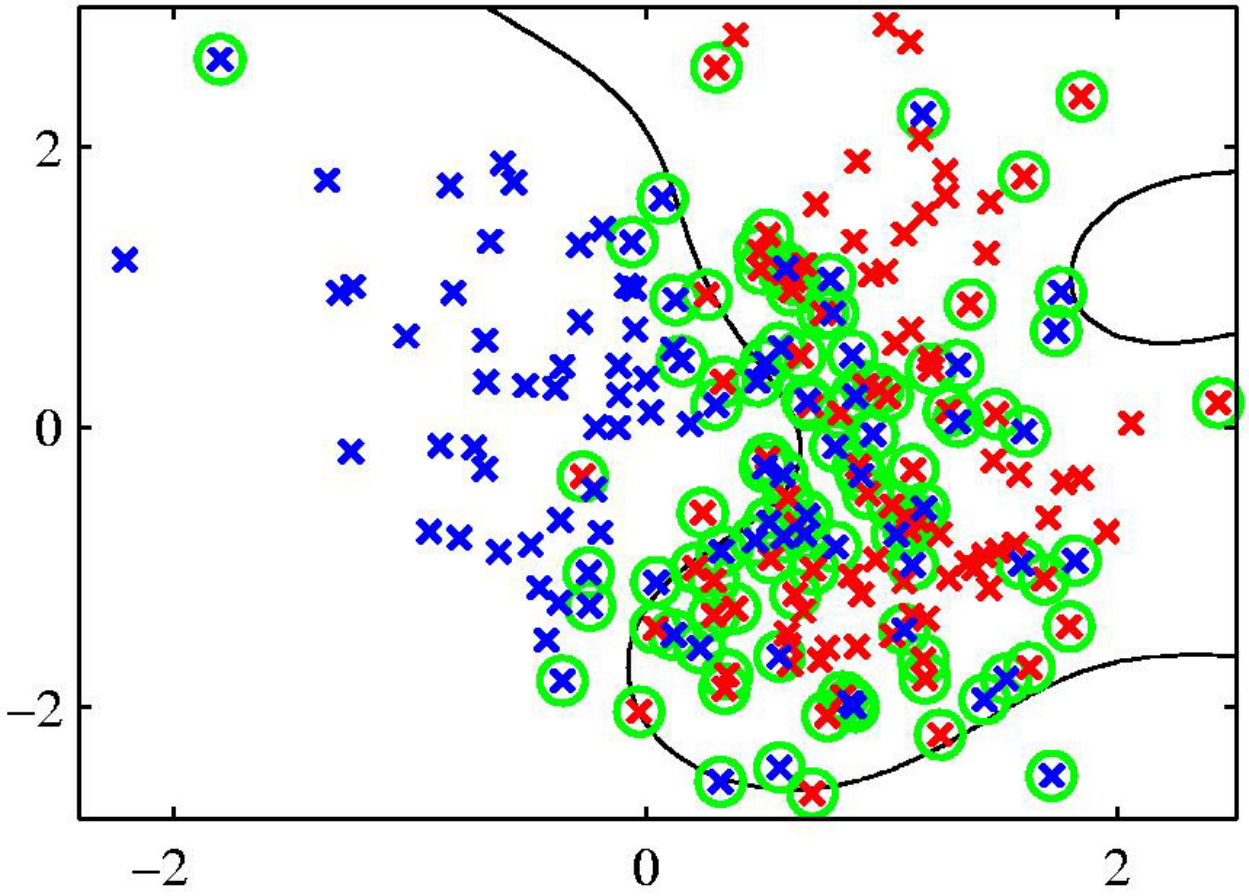
subject to $0 \le a_n \le 1/N$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$$\sum_{n=1}^{N} a_n \ge \nu$$

- Upper bound on the fraction of margin errors
- Lower bound on the fraction of support vectors

Example of Nonseparable Data Classification (*v*-SVM)



Solutions of the QP Problem

- Chunking (Vapnik, 1982)
  - Idea: the value of Lagrangian is unchanged if we remove the rows and columns of the kernel matrix corresponding to Lagrange multipliers that have value zero
  - Protected conjugate gradients (Burges, 1998)
- Decomposition methods (Osuna *et al*., 1996)
- Sequential minimal optimization (Platt, 1999)

Outlines

- Support Vector Machines
- SVM & Logistic Regression
- SVM for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

Relation to Logistic Regression

For data points on the correct side, $\xi = 0$

For the remaining points, $\xi = 1 - y_n t_n$

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\mathbf{w}\|^2 \Rightarrow \sum_{n=1}^{N} E_{SV}(y_n, t_n) + \lambda\|\mathbf{w}\|^2$$

where $\lambda = (2C)^{-1}$

$E_{SV}(y_n, t_n) = [1 - y_n t_n]_+$: hinge error function

where $[\cdot]_+$ denotes the positive part

Relation to Logistic Regression (Cont'd)

From maximum likelihood logistic regression

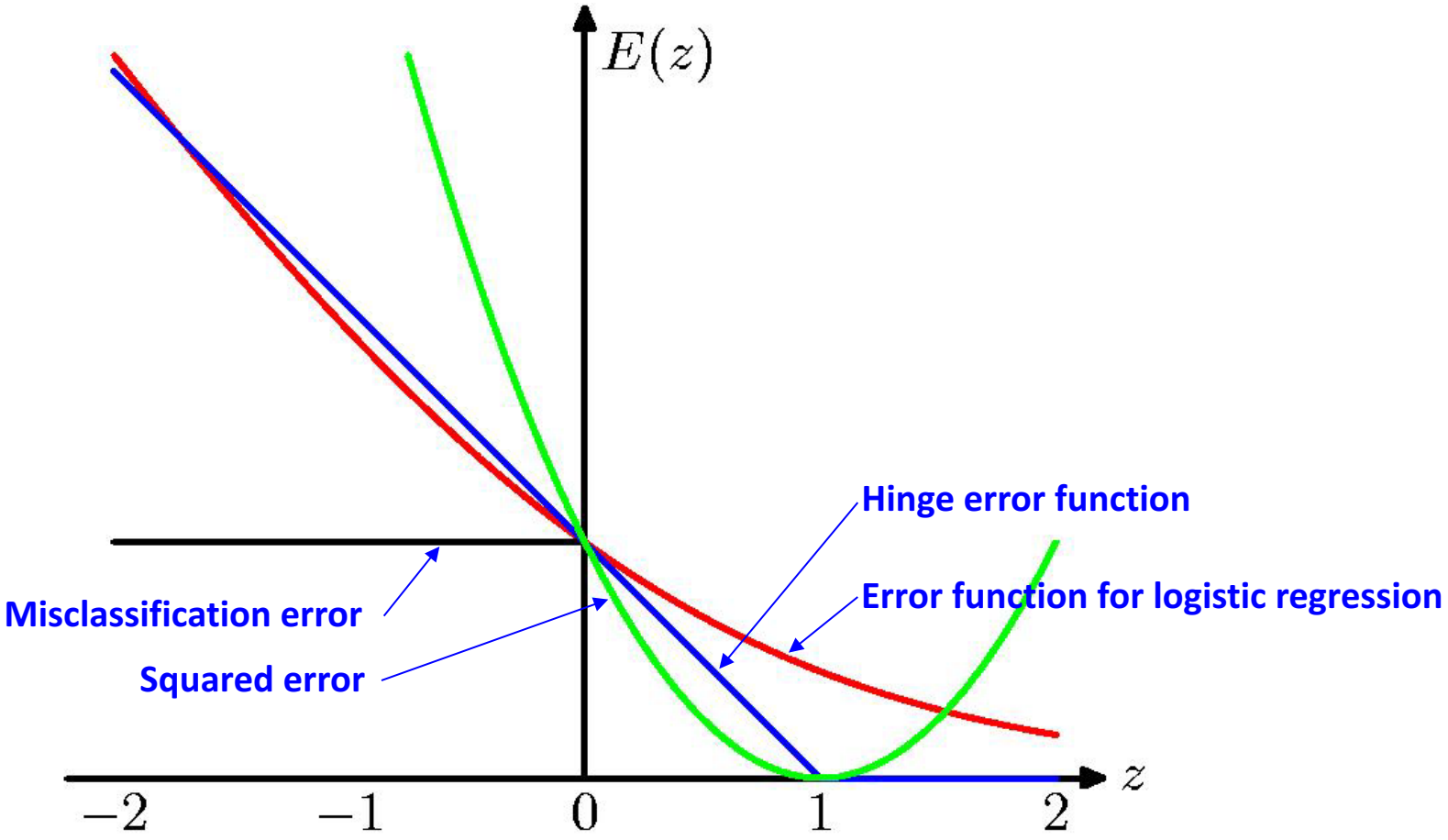
$$p(t=1 \mid y) = \sigma(y)$$
$$p(t=-1 \mid y) = 1 - \sigma(y) = \sigma(-y)$$
$$\Rightarrow p(t \mid y) = \sigma(yt)$$

Error function with a quadratic regularizer

$$\sum_{n=1}^{N} E_{LR}(y_n t_n) + \lambda \|\mathbf{w}\|^2$$

where $E_{LR}(yt) = \ln\left(1 + \exp\left(-yt\right)\right)$

Comparison of Error Functions



Multiclass SVMs

- *One-versus-the-rest*: *K* separate SVMs
  - Can lead inconsistent results (Figure 4.2)
  - Imbalanced training sets
    - Positive class: +1, negative class: -1/(*K*-1) (Lee *et al*., 2001)
- An objective function for training all SVMs simultaneously (Weston and Watkins, 1999)
- *One-versus-one*: *K*(*K*-1)/2 SVMs
- Based on error-correcting output codes (Allwein *et al*., 2000)
  - Generalization of the voting scheme of the *one-versus-one*

Outlines

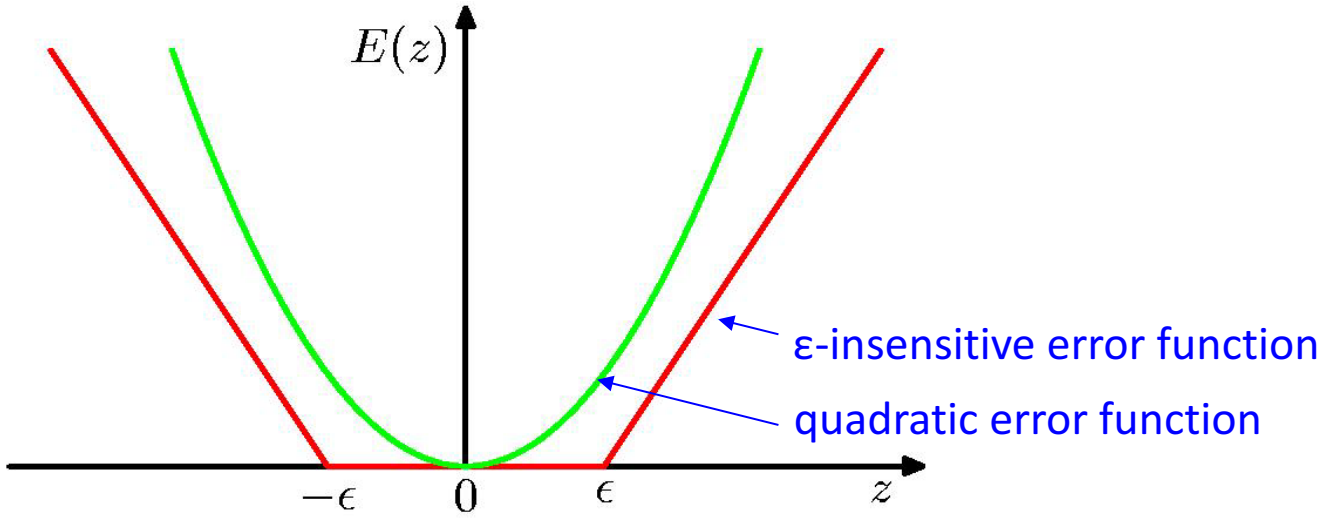
- Support Vector Machines
- SVM & Logistic Regression
- SVM for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

SVMs for Regression

Simple linear regression: minimize $\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$

ε-insensitive error function

$$E_\varepsilon\big(y(\mathbf{x})-t\big)=\begin{cases}0, & \text{if } |y(\mathbf{x})-t|<\varepsilon\\ |y(\mathbf{x})-t|-\varepsilon, & \text{otherwise}\end{cases}$$



SVMs for Regression (Cont'd)

Minimize

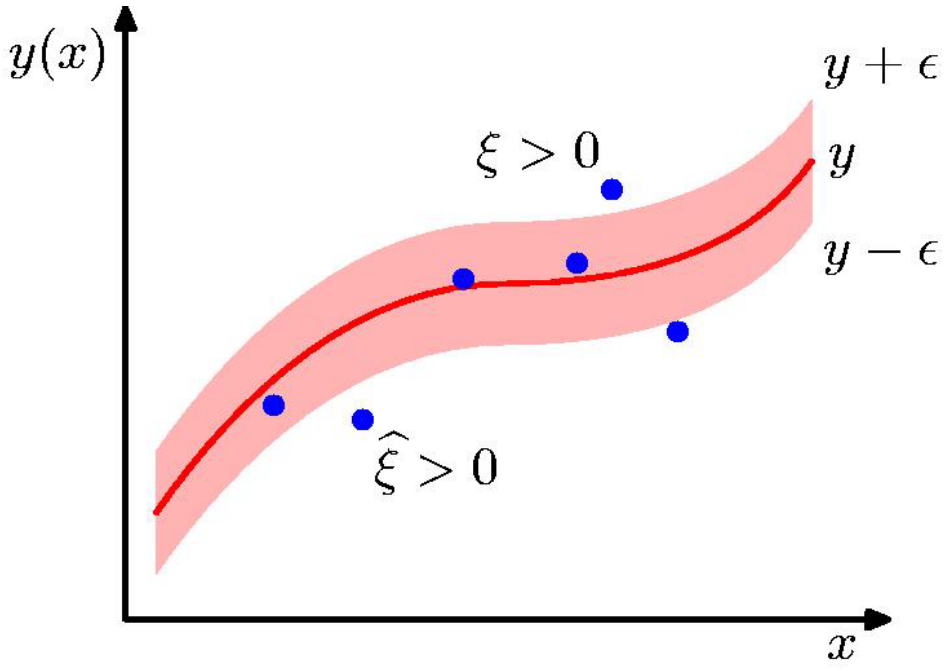
$$C\sum_{n=1}^{N} E_{\varepsilon}\left(y(\mathbf{x}_n)-t_n\right)+\frac{1}{2}\|\mathbf{w}\|^2$$

$$C\sum_{n=1}^{N} \left(\xi_n+\hat{\xi}_n\right)+\frac{1}{2}\|\mathbf{w}\|^2$$

where $t_n \le y(\mathbf{x}_n)+\varepsilon+\xi_n$

$$t_n \ge y(\mathbf{x}_n)-\varepsilon-\hat{\xi}_n$$

$$\xi_n \ge 0,\ \hat{\xi}_n \ge 0$$



Dual Problem

$$L = C\sum_{n=1}^{N}\left(\xi_n + \hat{\xi}_n\right) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}\left(\mu_n\xi_n + \hat{\mu}_n\hat{\xi}_n\right)$$

$$-\sum_{n=1}^{N} a_n\left(\varepsilon + \xi_n + y_n - t_n\right) - \sum_{n=1}^{N}\hat{a}_n\left(\varepsilon + \hat{\xi}_n - y_n + t_n\right)$$

$$\widetilde{L}(\mathbf{a},\hat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m)$$

$$-\varepsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n$$

subject to $0 \le a_n, \hat{a}_n \le C$

$$\sum_{n=1}^{N}\left(a_n - \hat{a}_n\right) = 0$$

Predictions

$$\mathbf{w} = \sum_{n=1}^{N} (a_n - \hat{a}_n)\phi(\mathbf{x}_n)$$ **(from derivatives of the Lagrangian)**

$$y(\mathbf{x}) = \sum_{n=1}^{N} (a_n - \hat{a}_n)k(\mathbf{x}, \mathbf{x}_n) + b$$

KKT conditions:

$$a_n(\varepsilon + \xi_n + y_n - t_n) = 0$$
$$\hat{a}_n(\varepsilon + \hat{\xi}_n - y_n + t_n) = 0$$
$$(C - a_n)\xi_n = 0$$
$$(C - \hat{a}_n)\hat{\xi}_n = 0$$

$$b = t_n - \varepsilon - \mathbf{w}^T\phi(\mathbf{x}_n) = t_n - \varepsilon - \sum_{m=1}^{N} (a_m - \hat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m)$$

Alternative Formulation

v-SVM (Schölkopf *et al*., 2000)

$$\tilde{L}(\mathbf{a},\hat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\hat{a}_n)(a_m-\hat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m)$$

$$+\sum_{n=1}^{N}(a_n-\hat{a}_n)t_n$$

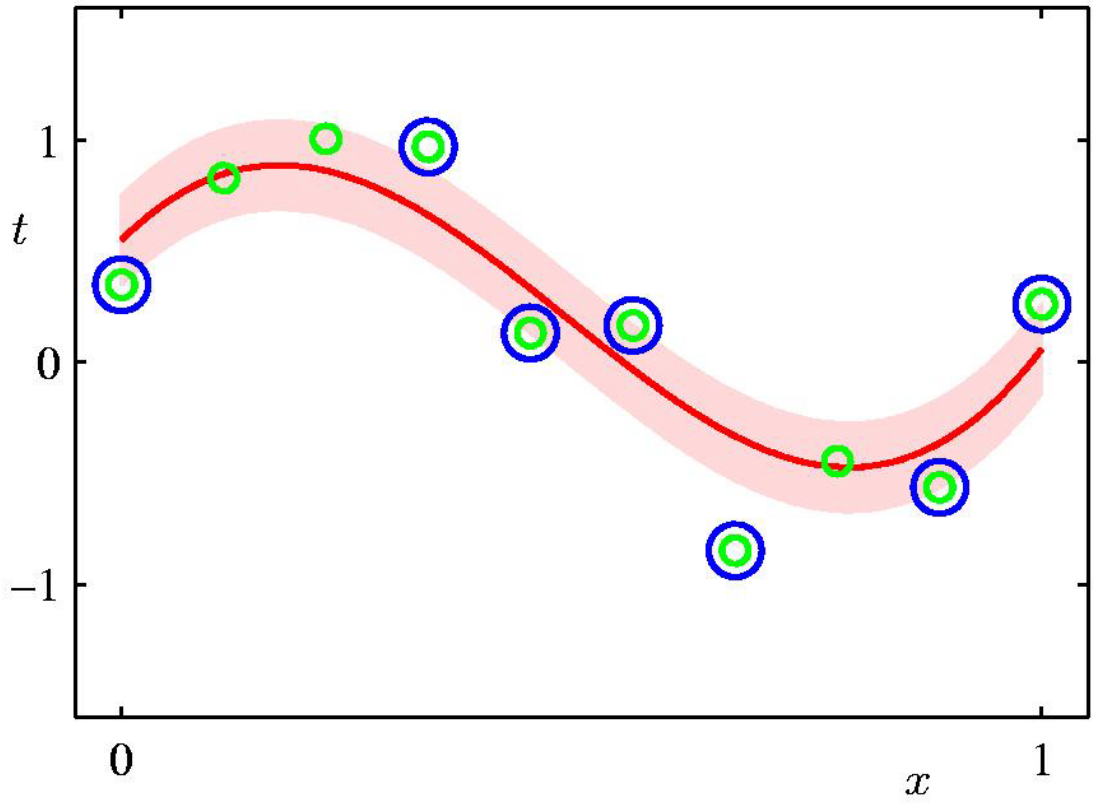
$$\text{subject to } 0 \le a_n, \hat{a}_n \le C/N$$

$$\sum_{n=1}^{N}(a_n-\hat{a}_n)=0$$

$$\sum_{n=1}^{N}(a_n+\hat{a}_n)\le \nu C$$

fraction of points lying outside the tube

Example of *v*-SVM Regression



Outlines

- Support Vector Machines
- SVM & Logistic Regression
- SVM for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

Relevance Vector Machines

SVM

- Outputs are decisions rather than posterior probabilities
- The extension to $K>2$ classes is problematic
- There is a complexity parameter C
- Kernel functions are centered on training data points and required to be positive definite

RVM

- Bayesian sparse kernel technique
- Much sparser models
- Faster performance on test data

Outlines

- Support Vector Machines
- SVM & Logistic Regression
- SVM for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

RVM for Regression

RVM is a linear form in Chapter 3 with a modified prior

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

$$\text{where } y(\mathbf{x}) = \sum_{i=1}^{M} w_i \phi_i(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) \Rightarrow y(\mathbf{x}) = \sum_{i=1}^{N} w_n k(\mathbf{x}, \mathbf{x}_n) + b$$

$$\beta = \sigma^{-2}$$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(t_n \mid \mathbf{x}_n, \mathbf{w}, \beta^{-1})$$

$$p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{n=1}^{N} N(w_i \mid 0, \alpha_i^{-1})$$

RVM for Regression (Cont'd)

$$p(\mathbf{w} \mid \mathbf{t}, \mathbf{X}, \boldsymbol{\alpha}, \beta) = N(\mathbf{w} \mid \mathbf{m}, \Sigma)$$

From the result (3.49) for linear regression models

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

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

where $\boldsymbol{\Phi}$: $N \times M$ matrix with elements $\Phi_{ni} = \phi_i(\mathbf{x}_n)$

$$\mathbf{A} = diag(\alpha_i)$$

α and β are determined using *evidence approximation* (type-2 maximum likelihood) (Section 3.5)

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

$$\ln p(\mathbf{t} \mid \mathbf{X}, \boldsymbol{\alpha}, \beta) = \ln N(\mathbf{t} \mid \mathbf{0}, \mathbf{C})$$

$$= -\frac{1}{2}\left\{N \ln(2\pi) + \ln|\mathbf{C}| + \mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right\}$$

⇒ Maximize

where $\mathbf{t} = (t_1, \ldots, t_N)^T$, $\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T$

RVM for Regression (Cont'd)

Two approaches

By derivatives of marginal likelihood

$$\alpha_i^{new} = \frac{\gamma_i}{m_i^2}, \quad \left(\beta^{new}\right)^{-1} = \frac{\left\|\mathbf{t} - \mathbf{\Phi m}\right\|^2}{N - \sum_i \gamma_i}$$

where $\gamma_i = 1 - \alpha_i \Sigma_{ii}$

EM algorithm → Section 9.3.4 Predictive distribution

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

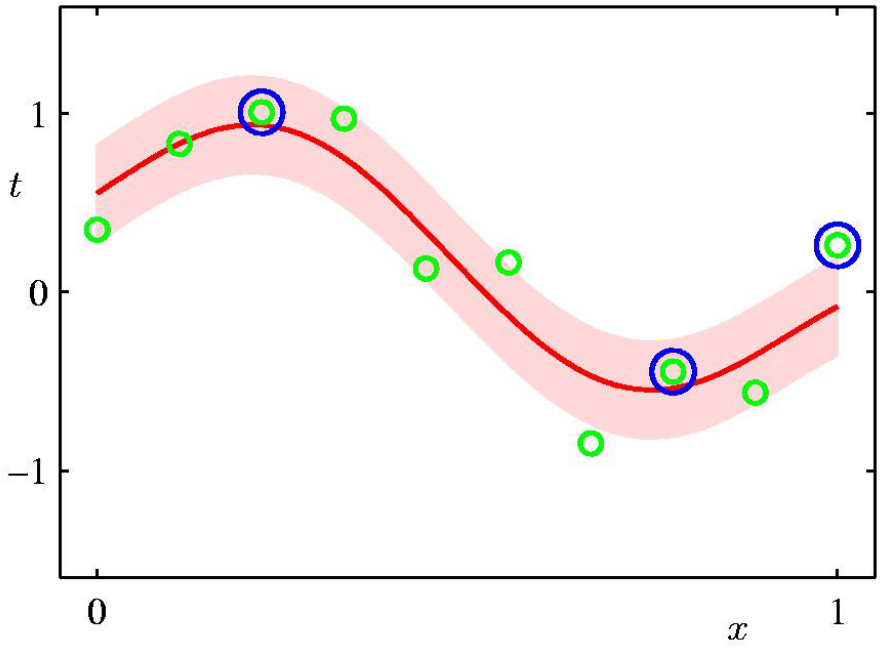
$$= N\left(t \mid \mathbf{m}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2(\mathbf{x})\right)$$

→ Section 3.3.2

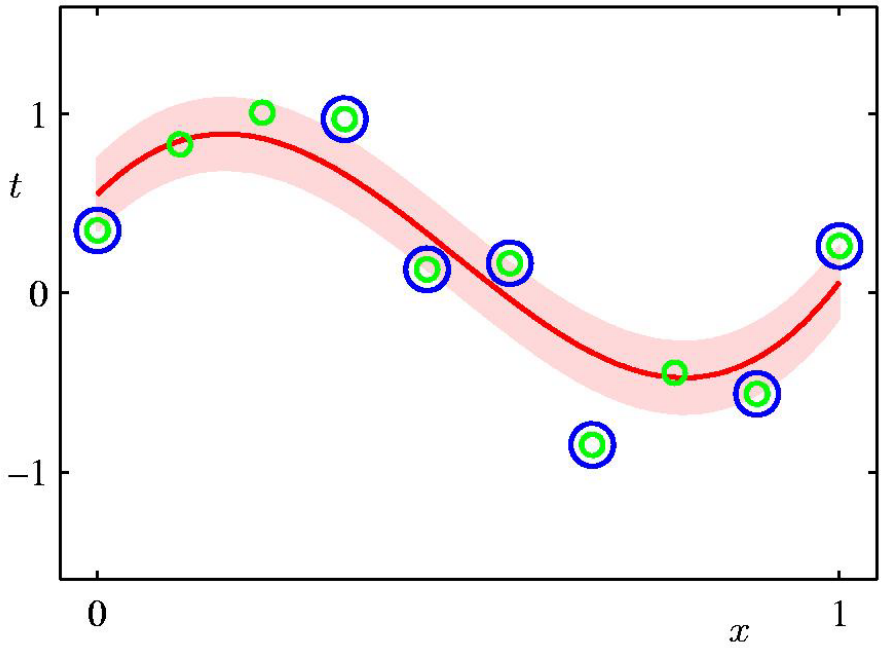
where $\sigma^2(\mathbf{x}) = \left(\beta^*\right)^{-1} + \boldsymbol{\phi}(\mathbf{x})^T \Sigma \boldsymbol{\phi}(\mathbf{x})$

Example of RVM Regression

RVM regression



v-SVM regression



More compact than SVM

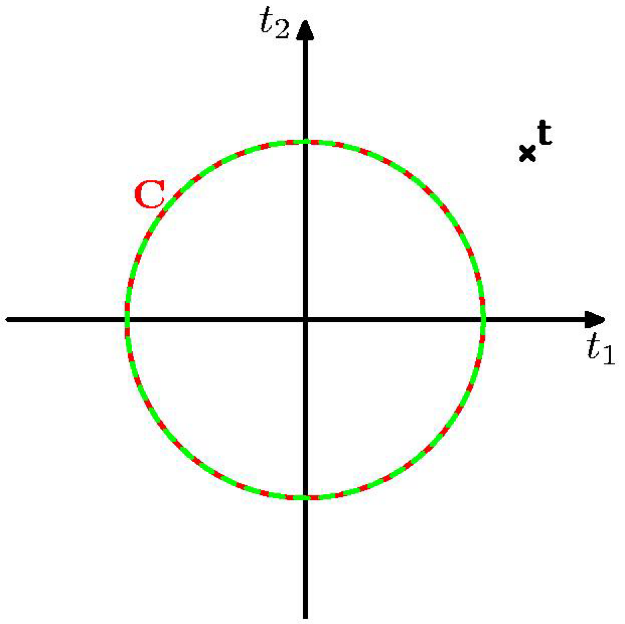
Parameters are determined automatically

Require more training time than SVM

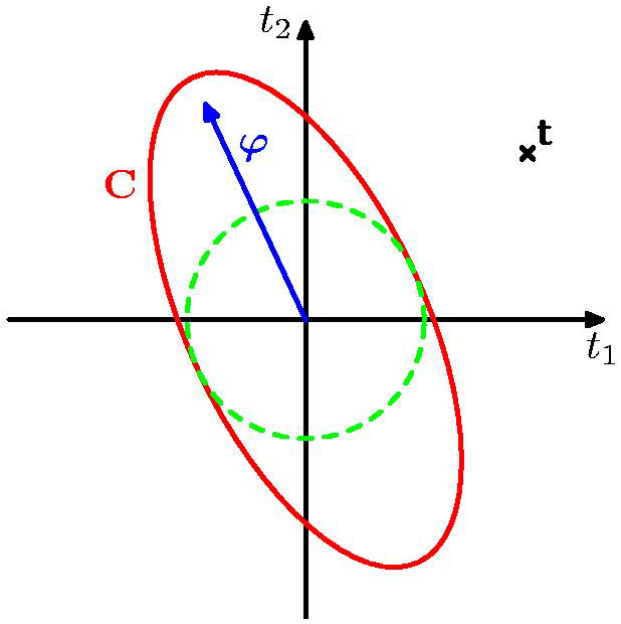
Mechanism for Sparsity

$$p(\mathbf{t} \mid \alpha, \beta) = N(\mathbf{t} \mid \mathbf{0}, \mathbf{C})$$

$$\text{where } \mathbf{t} = (t_1, t_2)^T, \ \mathbf{C} = \beta^{-1}\mathbf{I} + \alpha^{-1}\varphi\varphi^T$$



only isotropic noise, α = ∞



a finite value of α

Sparse Solution

Pull out the contribution from α_i in

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T$$

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \sum_{j \neq i} \alpha_j^{-1}\varphi_j\varphi_j^T + \alpha_i^{-1}\varphi_i\varphi_i^T \qquad \text{where } \varphi_i : i\text{th column of } \mathbf{\Phi}$$

$$= \mathbf{C}_{-i} + \alpha_i^{-1}\varphi_i\varphi_i^T$$

$$\left|\mathbf{C}\right| = \left|\mathbf{C}_{-i}\right|\left|1 + \alpha_i^{-1}\varphi_i^T\mathbf{C}_{-i}^{-1}\varphi_i\right| \qquad \rightarrow \text{Using (C.7), (C.15) in Appendix C}$$

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\varphi_i\varphi_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \varphi_i^T\mathbf{C}_{-i}^{-1}\varphi_i}$$

Sparse Solution (Cont'd)

For log marginal likelihood function *L*,

$$L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-1}) + \lambda(\alpha_i)$$

$$\lambda(\alpha_i) = \frac{1}{2}\left[\ln \alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right]$$

where $s_i = \varphi_i^T \mathbf{C}_{-i}^{-1} \varphi_i$

$q_i = \varphi_i^T \mathbf{C}_{-i}^{-1} \mathbf{t}$

- Sparsity: measures the extent to which φ_i overlaps with the other basis vectors
- Quality of φ_i : represents a measure of the alignment of the basis vector with the error between **t** and $\mathbf{y}_{-i}$

Stationary points of the marginal likelihood w.r.t.α_i

$$\Rightarrow \frac{d\lambda(\alpha_i)}{d\alpha_i} = \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

Sequential Sparse Bayesian Learning Algorithm

1. Initialize β
2. Initialize using φ_1, with $\alpha_1 = s_1^2 / \left(q_1^2 - s_1\right)$, with the remaining $\alpha_{j(j \neq i)} = \infty$
3. Evaluate Σ and $\mathbf{m}$ for all basis functions
4. Select a candidate φ_i
5. If $q_i^2 > s_i$, $\alpha_i < \infty$ (φ_i is already in the model), update $\alpha_i = s_i^2 / \left(q_i^2 - s_i\right)$
6. If $q_i^2 > s_i$, $\alpha_i = \infty$, add φ_i to the model, and evaluate $\alpha_i = s_i^2 / \left(q_i^2 - s_i\right)$
7. If $q_i^2 \leq s_i$, $\alpha_i < \infty$, remove φ_i from the model, and set $\alpha_i = \infty$
8. Update β
9. Go to 3 until converged

Outlines

- Support Vector Machines
- SVM & Logistic Regression
- SVM for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

RVM for Classification

Probabilistic linear classification model (Chapter 4)

$y(\mathbf{x}, \mathbf{w}) = \sigma\left(\mathbf{w}^T \phi(\mathbf{x})\right)$ with ARD prior $p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{n=1}^{N} N(w_i \mid 0, \alpha_i^{-1})$

- Initialize $\boldsymbol{\alpha}$
- Build a Gaussian approximation to the posterior distribution
- Obtain an approximation to the marginal likelihood
- Maximize the marginal likelihood (re-estimate $\boldsymbol{\alpha}$) until converged

RVM for Classification (Cont'd)

The posterior distribution is obtained by maximizing

$$\ln p(\mathbf{w}\,|\,\mathbf{t},\boldsymbol{\alpha}) = \ln\{p(\mathbf{t}\,|\,\mathbf{w})\,p(\mathbf{w}\,|\,\boldsymbol{\alpha})\} - \ln p(\mathbf{t}\,|\,\boldsymbol{\alpha})$$

$$= \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} + \text{const}$$

where $\mathbf{A} = diag(\alpha_i)$

➡ Iterative reweighted least squares (IRLS) from Section 4.3.3

$$\nabla \ln p(\mathbf{w}\,|\,\mathbf{t},\boldsymbol{\alpha}) = \boldsymbol{\Phi}^T(\mathbf{t}-\mathbf{y}) - \mathbf{A}\mathbf{w}$$

$$\nabla\nabla \ln p(\mathbf{w}\,|\,\mathbf{t},\boldsymbol{\alpha}) = -(\boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \mathbf{A})$$

where $\mathbf{B}: N \times N$ diagonal matrix, $b_n = y_n(1-y_n)$,

$\boldsymbol{\Phi}$: design matrix, $\Phi_{ni} = \phi_i(\mathbf{x}_n)$

➡ Resulting Gaussian approximation to the posterior distribution

$$\mathbf{w}^* = \mathbf{A}^{-1}\boldsymbol{\Phi}^T(\mathbf{t}-\mathbf{y}),\ \Sigma = (\boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \mathbf{A})^{-1}$$

RVM for Classification (Cont'd)

Marginal likelihood using Laplace approximation (Section 4.4)

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

$$= p(\mathbf{t}\,|\,\mathbf{w}^*)p(\mathbf{w}^*\,|\,\boldsymbol{\alpha})(2\pi)^{M/2}\left|\Sigma\right|^{1/2}$$

Set the derivative of the marginal likelihood equal to zero, and rearranging then gives

$$\alpha_i^{new} = \frac{\gamma_i}{\left(w_i^*\right)^2} \quad \text{where} \quad \gamma_i = 1 - \alpha_i \Sigma_{ii}$$

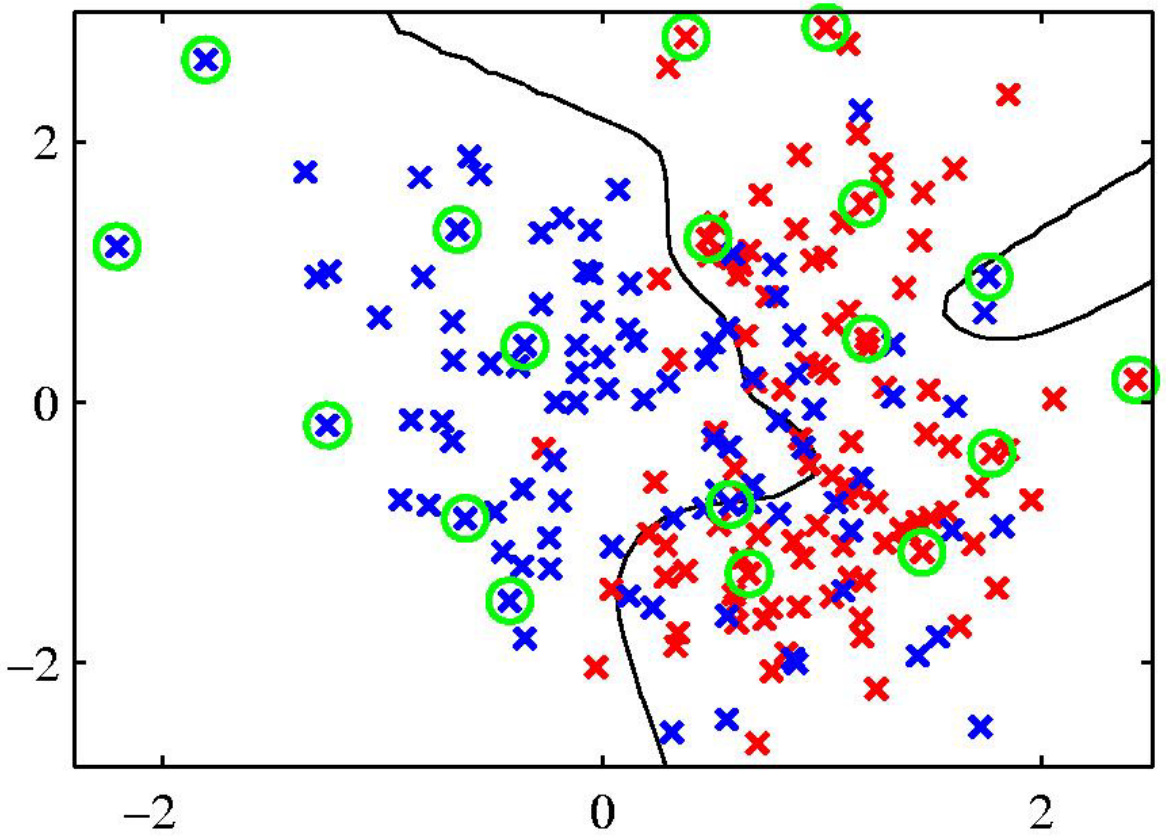
If we define $\hat{\mathbf{t}} = \boldsymbol{\Phi}\mathbf{w}^* + \mathbf{B}^{-1}(\mathbf{t} - \mathbf{y})$

$$\ln p(\mathbf{t}\,|\,\boldsymbol{\alpha},\beta) = -\frac{1}{2}\left\{N\ln(2\pi) + \ln\left|\mathbf{C}\right| + (\hat{\mathbf{t}})^T\mathbf{C}^{-1}\hat{\mathbf{t}}\right\}$$

Same in the regression case

where $\mathbf{C} = \mathbf{B} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T$

Example of RVM Classification



HW6

SVM： 7.1 7.3 7.4 7.6 7.7

RVM： 7.10 7.12 7.13 7.15