PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 7: SPARSE KERNEL MACHINES

Learning Objectives

1、 What are support vector machines?

2、 What are maximum (soft) margin classifiers?

3、 What the relation between SVMs and logistic regression?

4、 How to use SVMs for regression?

5、 What are relevance vector machines?

6、 How to use RVMs for regression?

7、 How to use RVMs for classification?

8、 What is the mechanism for RVMs to have sparse solutions?

Outlines

- Support Vector Machines
- SVM and 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
  - ✓ Kernel functions are centered on training data points and required to be positive definite
- RVM
  - ✓ Bayesian regression and classification frameworks
  - ✓ Bayesian sparse kernel technique
  - ✓ Much sparser models
  - ✓ Faster performance on test data

Outlines

- Support Vector Machines
- SVMs and Logistic Regression
- SVMs for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

RVM for Regression I

- RVM is a linear form with a modified prior

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

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

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

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

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

Each data sample has a weight

RVM for Regression II

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

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 $\Phi_{ni} = \phi_i(\mathbf{x}_n)$

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

<u>α and β are determined using *evidence approximation*</u>

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

Prior Predictive Distribution

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

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

Two steps

① From derivatives of the marginal likelihood, we have

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

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

Σ_{ii} : ith diagonal element of Σ

② Predictive distribution

Posterior Predictive Distribution

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

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

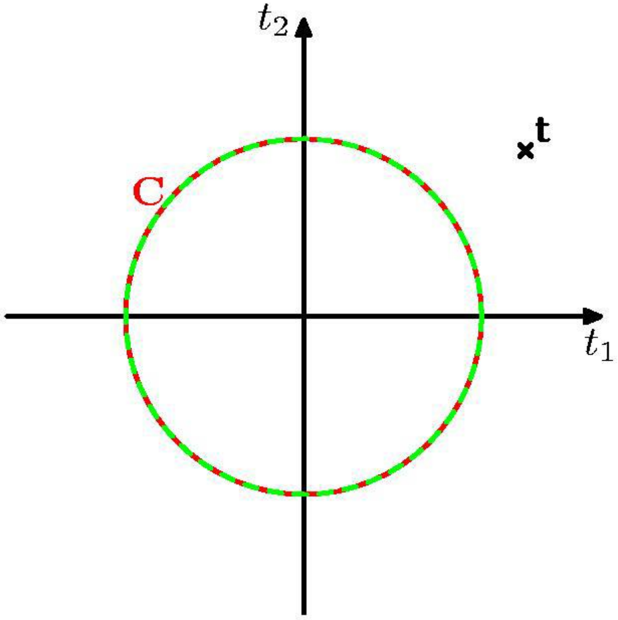
where $\sigma^2(\mathbf{x}) = \left(\beta^*\right)^{-1} + \phi(\mathbf{x})^T \Sigma \phi(\mathbf{x})$

Mechanism for Sparsity

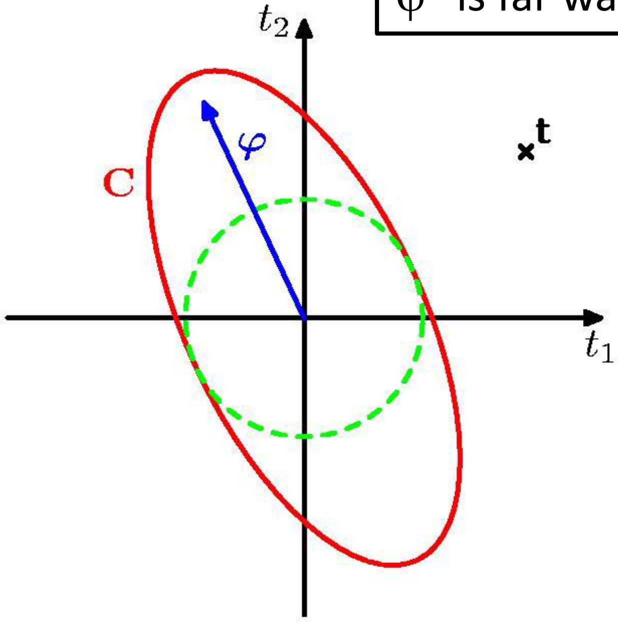
$$p(\mathbf{t} \mid \alpha, \beta) = N(\mathbf{t} \mid \mathbf{0}, \mathbf{C})$$

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

φ is far way from x^t



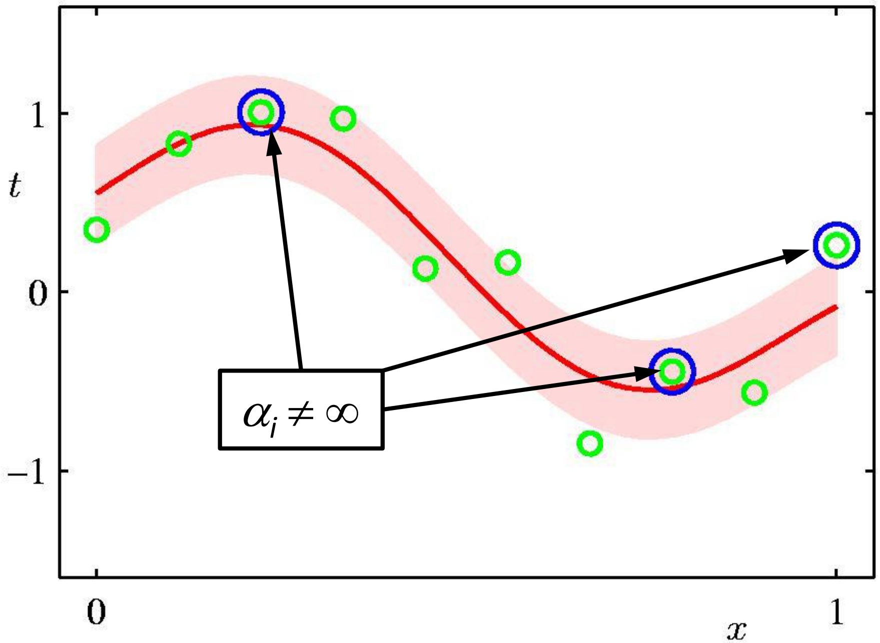
only isotropic noise, α = ∞



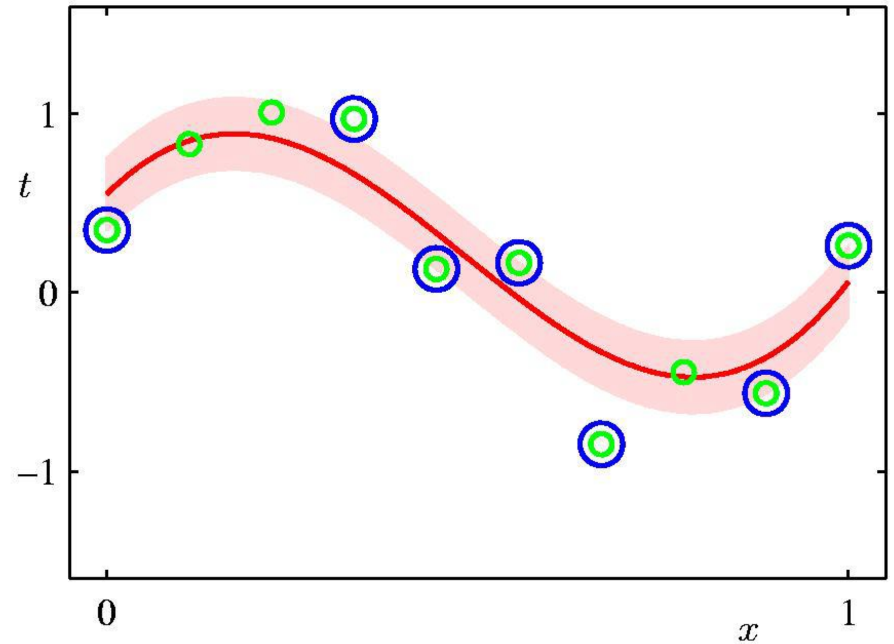
a finite value of α

Examples of RVM Regression

RVM regression



ν-SVM regression



More compact than SVM (3 relevance vectors v.s. 7 support vectors)

Parameters are determined automatically

Require more training time than SVM

Sparse Solution I

Pull out the contribution from α_i in

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

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \sum_{j \neq i} \alpha_j^{-1}\varphi_j\varphi_j^T + \alpha_i^{-1}\varphi_i\varphi_i^T$$

where $\varphi_i : i$th column of $\mathbf{\Phi}$

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

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

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 II

- Then log marginal likelihood function *L* becomes,

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

$L(\boldsymbol{\alpha}_{-i})$: omitting α_i

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

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

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

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

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
- SVMs and Logistic Regression
- SVMs for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification

RVM for Classification

- Probabilistic linear classification model with Gaussian prior

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

- 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)

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

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

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

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

⇨ Resulting Gaussian approximation to the posterior distribution

$$\mathbf{w}^* = \mathbf{A}^{-1}\mathbf{\Phi}^T(\mathbf{t}-\mathbf{y}),\ \Sigma = \left(\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A}\right)^{-1} \quad \Leftarrow \quad \nabla \ln p(\mathbf{w}|\mathbf{t},\alpha) = 0$$

RVM for Classification (Cont'd)

- Marginal likelihood using Laplace approximation

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

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

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

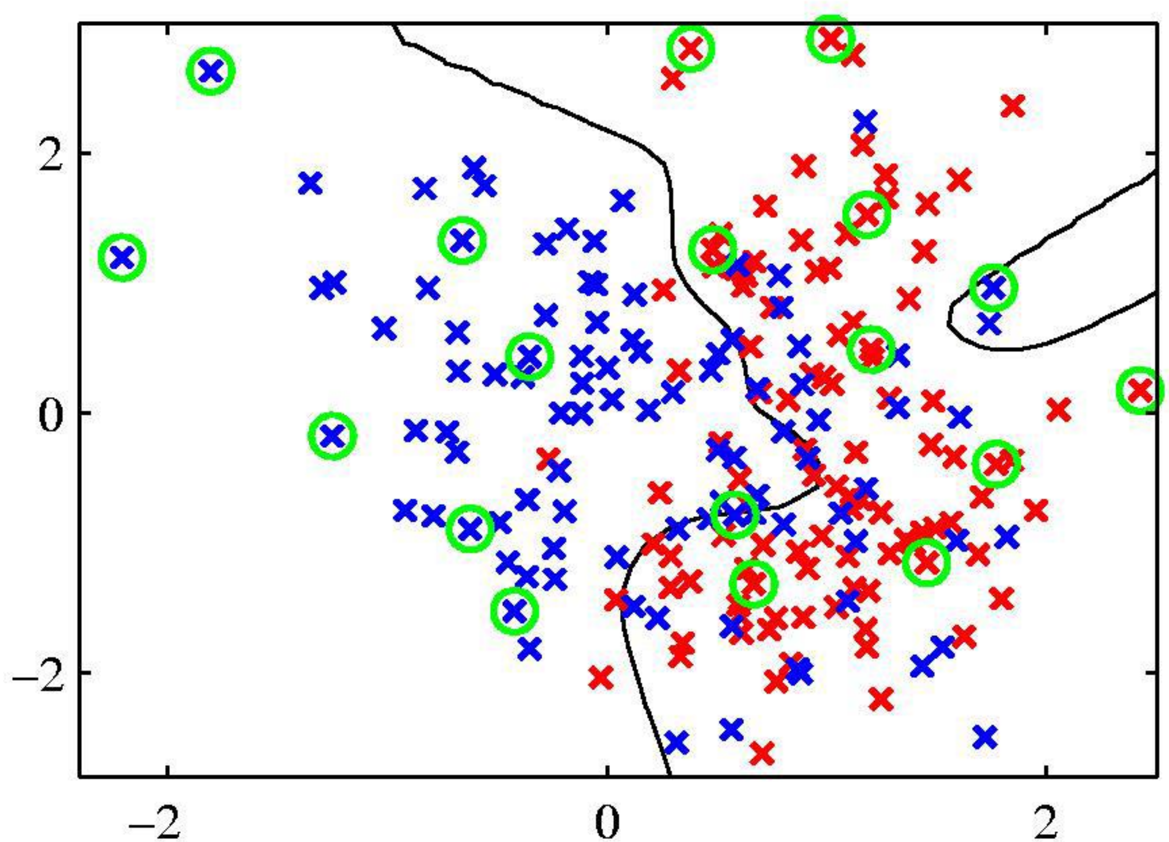
$$\alpha_i^{new} = \frac{\gamma_i}{\left(w_i^*\right)^2} \quad \text{where} \quad \gamma_i = 1 - \alpha_i \Sigma_{ii}$$

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

$$\ln p(\mathbf{t}\,|\,\boldsymbol{\alpha}, \beta) = -\frac{1}{2}\left\{N\ln(2\pi) + \ln\left|\mathbf{C}\right| + (\hat{\mathbf{t}})^T\mathbf{C}^{-1}\hat{\mathbf{t}}\right\} \Rightarrow \boxed{\textbf{Same in the regression case}}$$

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

Example of RVM Classification



Summary

- Support Vector Machines
- SVMs and Logistic Regression
- SVMs for Regression
- Relevance Vector Machines
- RVMs for Regression
- RVMs for Classification