

Support Vector Classifiers and Machines

Sections 12.1–12.3

CN700 *Elements of Statistical Learning*, April 22, 2008

satyavarta/sat@cns.bu.edu
Auditory Neuroscience Lab,
Dept Cognitive and Neural Systems
Boston University

Classification Task

Outline

- ① SEPARABLE CASE
- ② NONSEPARABLE CASE
- ③ Computation of Optimal Hyperplane
- ④ Support Vector Machines
- ⑤ SVM for Regression

Geometry of Hyperplanes

signed distance $\beta^{*T}(x - x_0)$

Classification Task

Hyperplane $\{x : f(x) = x^T \beta + \beta_0 = 0\}$ $\|\beta\| = 1$

Rule $G(x) = \text{sign}[x^T \beta + \beta_0]$

Misclassification $y_i f(x_i) < 0$

“Optimal Solution”

Hyperplane $\{x : f(x) = x^T \beta + \beta_0 = 0\}$ $\|\beta\| = 1$

Rule $G(x) = \text{sign}[x^T \beta + \beta_0]$

Misclassification: $y_i f(x_i) < 0$

“Optimal Solution”: Support Vectors

Hyperplane $\{x : f(x) = x^T \beta + \beta_0 = 0\}$ $\|\beta\| = 1$

Rule $G(x) = \text{sign}[x^T \beta + \beta_0]$

Misclassification: $y_i f(x_i) < 0$

Optimal Solution, Formally

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C, i = 1, 2, \dots, N$

Optimal Solution, simplified

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C, i = 1, 2, \dots, N$

Optimal Solution, simplified

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C, i = 1, 2, \dots, N$

$$\max_{\beta, \beta_0} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C\|\beta\|, i = 1, 2, \dots, N$

Optimal Solution, simplified

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C, i = 1, 2, \dots, N$

$$\max_{\beta, \beta_0} C$$

subject to $y_i(x_i^T \beta + \beta_0) \geq C \|\beta\|, i = 1, 2, \dots, N$
 Arbitrarily scaled β works, pick $C = \frac{1}{\|\beta\|}$

$$\min_{\beta, \beta_0} \|\beta\|$$

subject to $y_i(x_i^T \beta + \beta_0) \geq 1, i = 1, 2, \dots, N$

Outline

- ① SEPARABLE CASE
- ② NONSEPARABLE CASE
- ③ Computation of Optimal Hyperplane
- ④ Support Vector Machines
- ⑤ SVM for Regression

Optimal Solution for Nonseparable Case

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq C \\ \forall i$$

$$\max_{\beta, \beta_0, \|\beta\|=1} C$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq C(1 - \xi_i) \\ \forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const}$$

Optimal Solution, simplified

$$\min_{\beta, \beta_0} \|\beta\|$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq 1 \\ \forall i$$

$$\min_{\beta, \beta_0} \|\beta\|$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i \\ \forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const}$$

Outline

- ① SEPARABLE CASE
- ② NONSEPARABLE CASE
- ③ Computation of Optimal Hyperplane**
- ④ Support Vector Machines
- ⑤ SVM for Regression

$$\begin{aligned} & \min_{\beta, \beta_0} \|\beta\| \\ & \text{with } y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i \\ \forall i \quad & \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const} \end{aligned}$$

$$\min_{\beta, \beta_0} \|\beta\|$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i$$

$$\forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const}$$

$$\min_{\beta, \beta_0} \|\beta\| + \gamma \sum_{i=1}^N \xi_i$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i$$

$$\forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const}$$

$$\min_{\beta, \beta_0} \|\beta\|$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i$$

$$\forall i \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const}$$

$$\min_{\beta, \beta_0} \|\beta\| + \gamma \sum_{i=1}^N \xi_i$$

$$\text{with } y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i$$

$$\forall i, \quad \xi_i \geq 0, \quad \sum_{i=1}^N \xi_i \leq \text{const}$$

$$L_P = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

minimize w.r.t. β , β_0 , and ξ_i

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

minimize w.r.t. β , β_0 , and ξ_i

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

minimize w.r.t. β , β_0 , and ξ_i

$$\beta = \sum_{i=1}^N \alpha_i y_i x_i \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \alpha_i = \gamma - \mu_i, \forall i$$

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

minimize w.r.t. β , β_0 , and ξ_i

$$\beta = \sum_{i=1}^N \alpha_i y_i x_i \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \alpha_i = \gamma - \mu_i, \forall i$$

Lagrange Dual

$$L_D = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$

maximize under $0 \leq \alpha_i \leq \gamma$ and $\sum \alpha_i y_i = 0$ and KKT conditions:

$$\alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] = 0$$

$$\mu_i \xi_i = 0$$

$$y_i(x_i^T \beta + \beta_0) - (1 - \xi_i) \geq 0$$

Support Vectors

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

maximize w.r.t. β , β_0 , and ξ_i

$$\boxed{\beta = \sum_{i=1}^N \alpha_i y_i x_i} \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \alpha_i = \gamma - \mu_i, \forall i$$

Lagrange Dual

$$L_D = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$

maximize under $0 \leq \alpha_i \leq \gamma$ and $\sum \alpha_i y_i = 0$ and KKT conditions:

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] = 0$$

$$\mu_i \xi_i = 0$$

$$y_i (x_i^T \beta + \beta_0) - (1 - \xi_i) \geq 0$$

Support Vectors

Convex Quadratic Programming problem

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

maximize w.r.t. β , β_0 , and ξ_i

$$\beta = \sum_{i=1}^N \alpha_i y_i x_i \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \alpha_i = \gamma - \mu_i, \forall i$$

Lagrange Dual

$$L_D = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$

maximize under $0 \leq \alpha_i \leq \gamma$ and $\sum \alpha_i y_i = 0$ and KKT conditions:

$$\alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] = 0$$

$$\mu_i \xi_i = 0$$

$$y_i(x_i^T \beta + \beta_0) - (1 - \xi_i) \geq 0$$

Example

$\gamma = 10,000$ slack is low
62% points are in the support

Example

$\gamma = 0.01$ slack is high
85% points are in the support

Outline

- ① SEPARABLE CASE
- ② NONSEPARABLE CASE
- ③ Computation of Optimal Hyperplane
- ④ Support Vector Machines**
- ⑤ SVM for Regression

Support Vector Classifier

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

maximize w.r.t. β , β_0 , and ξ_i

$$\beta = \sum_{i=1}^N \alpha_i y_i x_i \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \alpha_i = \gamma - \mu_i, \forall i$$

Lagrange Dual

$$L_D = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$

maximize under $0 \leq \alpha_i \leq \gamma$ and $\sum \alpha_i y_i = 0$ and KKT conditions:

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] = 0$$

$$\mu_i \xi_i = 0$$

$$y_i (x_i^T \beta + \beta_0) - (1 - \xi_i) \geq 0$$

$$\hat{G}(x) = \text{sign}[\hat{f}(x)] = \text{sign}[x^T \hat{\beta} + \hat{\beta}_0]$$

Support Vector Machine

Lagrange Primal

$$L_p = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (\mathbf{h}(x_i)^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

maximize w.r.t. β , β_0 , and ξ_i

$$\beta = \sum_{i=1}^N \alpha_i y_i \mathbf{h}(x_i) \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \alpha_i = \gamma - \mu_i, \forall i$$

Lagrange Dual

$$L_D = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \mathbf{h}(x_i)^T \mathbf{h}(x_j)$$

maximize under $0 \leq \alpha_i \leq \gamma$ and $\sum \alpha_i y_i = 0$ and KKT conditions:

$$\alpha_i [y_i (\mathbf{h}(x_i)^T \beta + \beta_0) - (1 - \xi_i)] = 0$$

$$\mu_i \xi_i = 0$$

$$y_i (\mathbf{h}(x_i)^T \beta + \beta_0) - (1 - \xi_i) \geq 0$$

$$\hat{G}(x) = \text{sign}[\hat{f}(\mathbf{h}(x))] = \text{sign}[\mathbf{h}(x)^T \hat{\beta} + \hat{\beta}_0]$$

Kernel Trick

x always appears as $\langle h(x), h(x') \rangle$

Rewrite as $K(x, x') = \langle h(x), h(x') \rangle$

Any symmetric positive semidefinite function K works.

Kernel Trick

x always appears as $\langle h(x), h(x') \rangle$

Rewrite as $K(x, x') = \langle h(x), h(x') \rangle$

Any symmetric positive semidefinite function K works.

Polynomial	$K(x, x') = (1 + \langle x, x' \rangle)^d$
Radial basis	$K(x, x') = \exp(-\ x - x'\ ^2/c)$
Neural network	$K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$

Example: Degree 4 polynomials in feature space

Example: Radial kernel in feature space

Claims

- “Kernel trick unique to SVM”
- “Can finesse curse of dimensionality”

SVM as Penalization Method

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum (\xi_i)$$

where $(y_i f(x_i)) \geq 1 - \xi_i$

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum [1 - y_i f(x_i)]_+$$

Familiar problem: $\min \text{penalty} + \text{loss}$

SVM Loss Function

Loss function $L(Y, f(X))$

$$[1 - Yf(X)]_+$$

$$(Y - f(X))^2$$

$$\log(1 + \exp(-Yf(X)))$$

Minimizing function

$$f(X) = +1 \text{ if } \Pr(Y = +1|X) \geq \frac{1}{2}, -1, \text{ o/w}$$

$$f(X) = \Pr(Y = +1|X) - \Pr(Y = -1|X)$$

$$f(X) = \log \text{ odds}$$

Kernel Trick not unique to SVM

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum [1 - y_i f(x_i)]_+$$

Theory of reproducing kernel Hilbert spaces guarantees a **finite dimensional solution** of the form

$$f(x) = \beta_0 + \sum \alpha_j K(x, x_j)$$

- Common to smoothing splines, additive and interaction spline models.
- Can use any loss function (except binomial log-likelihood loss)

Curse of Dimensionality

Class I Standard normal X_1, X_2, X_3, X_4

Class II Standard normal conditioned on $9 \leq X_j^2 \leq 16$

Curse of Dimensionality

Class I Standard normal X_1, X_2, X_3, X_4

Class II Standard normal conditioned on $9 \leq X_j^2 \leq 16$

Opportunity Kernel cannot concentrate on subspaces

Outline

- ① SEPARABLE CASE
- ② NONSEPARABLE CASE
- ③ Computation of Optimal Hyperplane
- ④ Support Vector Machines
- ⑤ SVM for Regression

SVM for Regression

Classification

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum (\xi_i)$$

where $(y_i f(x_i)) \geq 1 - \xi_i$

Regression

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum (??)$$

where $|y_i - f(x_i)| \leq C + \epsilon$

SVM for Regression

Classification

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum (\xi_i)$$

where $(y_i f(x_i)) \geq 1 - \xi_i$

Regression

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum \alpha_i (|y_i - f(x_i)| - \epsilon)_+$$

where $C + \epsilon \geq |y_i - f(x_i)|$

SVM for Regression

Classification

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum (\xi_i)$$

where $(y_i f(x_i)) \geq 1 - \xi_i$

Regression

$$\min_{\beta_0, \beta} \|\beta\|^2 + \gamma \sum \alpha_i (|y_i - f(x_i)| - \epsilon)_+$$

where $C + \epsilon \geq |y_i - f(x_i)|$

Kernel Trick in Regression

$$f(x) = \sum_m \beta_m h_m(x) + \beta_0$$

$$\min H(\beta, \beta_0) = \sum_i V(y_i - f(x_i)) + \sum \beta_m^2$$

has solution of the form

$$\beta_m = \sum \alpha_i y_i h_m(x_i)$$

$$\hat{f}(x) = \sum \hat{\beta}_m h_m(x) + \hat{\beta}_0$$

Amenable to kernel trick due to inner product in $\hat{f}(x)$:

$$\hat{f}(x) = \sum \hat{\alpha}_i K(x, x_i)$$

Notes

- Multiclass extension by 1 vs rest, or multinomial loss

Notes

- Multiclass extension by 1 vs rest, or multinomial loss
- VC Dimension applicable