

Lecture 18.

SVM (II): Non-separable Cases

Outline

Non-separable pattern classification

Additional term in cost function

Primal problem formulation

Dual problem formulation

KKT condition

Implication of Minimizing $\|w\|$

Let D denote the diameter of the smallest hyper-ball that encloses all the input training vectors $\{x_1, x_2, \dots, x_N\}$.

The set of optimal hyper-planes described by the equation

$$W_o^T x + b_o = 0$$

has a VC-dimension h bounded from above as

$$h \leq \min \{ \lceil D^2/\rho^2 \rceil, m_o \} + 1$$

where m_o is the dimension of the input vectors, and $\rho = 2/\|w_o\|$ is the margin of the separation of the hyper-planes.

VC-dimension determines the *complexity of the classifier structure, and usually the smaller the better.*

Non-separable Cases

Recall that in linearly separable case, each training sample pair (x_i, d_i) represents a linear inequality constraint

$$d_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, N \quad (*)$$

If the training samples are not linearly separable, the constraint can be modified to yield a *soft constraint*:

$$d_i(w^T x_i + b) \geq 1 - \zeta_i, \quad i = 1, 2, \dots, N \quad (**)$$

$\{\zeta_i; 1 \leq i \leq N\}$ are known as *slack variables*.

Note that originally, (*) is a normalized version of

$d_i g(x_i)/|w| \geq \rho$. With the slack variable ζ_i , that eq. becomes $d_i g(x_i)/|w| \geq \rho(1 - \zeta_i)$. Hence with the slack variable, we allow some samples x_i fall within the gap. Moreover, if $\zeta_i > 1$, then the corresponding (x_i, d_i) is mis-classified because the sample will fall on the *wrong side* of the hyper-plane H.

Non-Separable Case

Since $\zeta_i > 1$ implies misclassification, the cost function must include a term to minimize the number of samples that are misclassified:

$$\Phi(W, \zeta) = W^T W / 2 + \lambda \sum_{i=1}^N I(\zeta_i - 1)$$

where λ is a Lagrange multiplier. But this formulation is non-convex and a solution is difficult to find using existing nonlinear optimization algorithms.

Hence, we may instead use an approximated cost function

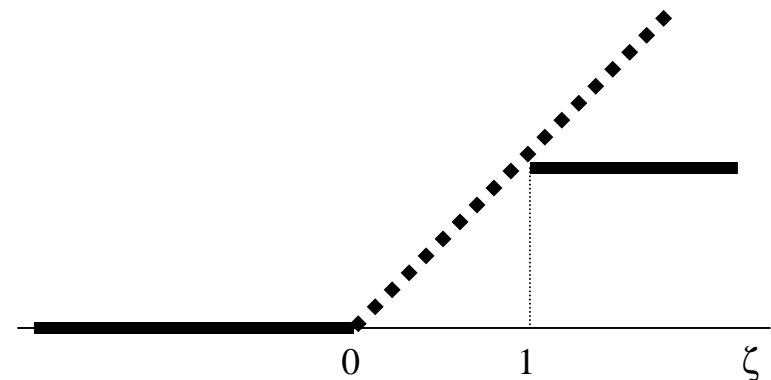
$$\Phi(W, \zeta) = \frac{1}{2} W^T W + C \sum_{i=1}^N \zeta_i$$

With this approximated cost function, the goal is to maximize ρ (minimize $\|W\|$) while minimize $\zeta_i (\geq 0)$.

ζ_i : not counted if x_i outside gap and on the correct side.

$0 < \zeta_i < 1$: x_i inside gap, but on the correct side.

$\zeta_i > 1$: x_i on the wrong side (inside or outside gap).



Primal Problem Formulation

Primal Optimization Problem Given $\{(x_i, d_i); 1 \leq i \leq N\}$.

Find w, b such that

$$\Phi(w, \zeta) = \frac{1}{2} w^T w + C \sum_{i=1}^N \zeta_i$$

is minimized subject to the constraints

- (i) $\zeta_i \geq 0$, and
- (ii) $d_i(w^T x_i + b) \geq 1 - \zeta_i$ for $i = 1, 2, \dots, N$.

Using α_i and μ_i as Lagrange multipliers, the unconstrained cost function becomes

$$\Phi(w, \zeta) = \frac{1}{2} w^T w + C \sum_{i=1}^N \zeta_i - \sum_{i=1}^N \alpha_i (d_i (w^T x_i + b) - 1 + \zeta_i) - \sum_{i=1}^N \mu_i \zeta_i$$

Dual Problem Formulation

Note that

$$\frac{\partial \Phi(w, \zeta)}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^N \alpha_i d_i x_i$$

$$\frac{\partial \Phi(w, \zeta)}{\partial \zeta_i} = 0 \Rightarrow C - \alpha_i - \mu_i = 0$$

Dual Optimization Problem Given $\{(x_i, \zeta_i); 1 \leq i \leq N\}$. Find Lagrange multipliers $\{\alpha_i; 1 \leq i \leq N\}$ such that

$$Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j x_i^T x_j$$

is maximized subject to the constraints

- (i) $0 \leq \alpha_i \leq C$ (a user-specified positive number) and
- (ii) $\sum_{i=1}^N \alpha_i d_i = 0$

Solution to the Dual Problem

By the Karush-Kuhn-Tucker condition:

for $i = 1, 2, \dots, N$,

$$(i) \quad \alpha_i [d_i (w^T x_i + b) - 1 + \zeta_i] = 0 \quad (*)$$

$$(ii) \quad \mu_i \zeta_i = 0$$

At optimal point $\alpha_i + \mu_i = C$. Thus, one may deduce that

if $0 < \alpha_i < C$, then $\zeta_i = 0$ and $d_i(w^T x_i + b) = 1$

if $\alpha_i = C$, then $\zeta_i \geq 0$ and $d_i(w^T x_i + b) = 1 - \zeta_i \leq 1$

if $\alpha_i = 0$, then $d_i(w^T x_i + b) \geq 1$: x_i is not a support vector

Finally, the optimal solutions are:

$$w_o = \sum_{i=1}^N \alpha_i d_i x_i$$

$$b_o = \left(\sum_{i \in I_o} (1 - d_i w_o^T x_i) \right) / \left(\sum_{i \in I_o} d_i \right)$$

where $I_o = \{i; 0 < \alpha_i < C\}$