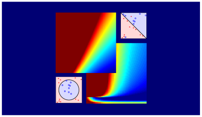
Machine Learning Techniques
(機器學習技巧)



Lecture 4: Soft-Margin SVM

Hsuan-Tien Lin (林軒田)
htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)

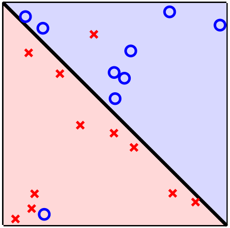
Agenda

Lecture 4: Soft-Margin SVM

- Soft-Margin SVM: Primal
- Soft-Margin SVM: Dual
- Soft-Margin SVM: Solution
- Soft-Margin SVM: Selection

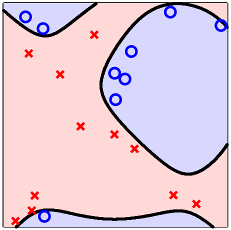
Cons of Hard-Margin SVM

recall: SVM can still overfit :-(



$\mathbf{\Phi}_1$

- part of reasons: $\mathbf{\Phi}$
- other part: **separable**



$\mathbf{\Phi}_4$

if always insisting on **separable** ($\Longrightarrow$ **shatter**),
have power to **overfit to noise**

Give Up on Some Examples

want: **give up** on some noisy examples

pocket

$$\min_{b,\mathbf{w}} \quad \sum_{n=1}^{N} [\![y_n \neq \mathrm{sign}(\mathbf{w}^T\mathbf{z}_n + b)]\!]$$

hard-margin SVM

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for all } n$$

combination:

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} [\![y_n \neq \mathrm{sign}(\mathbf{w}^T\mathbf{z}_n + b)]\!]$$

$$\begin{aligned} \text{s.t.} \quad & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for } \textbf{correct } n \\ & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq -\infty \text{ for } \textbf{incorrect } n \end{aligned}$$

C: trade-off of large margin & noise tolerance

Soft-Margin SVM (1/2)

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} [\![y_n \neq \text{sign}(\mathbf{w}^T\mathbf{z}_n + b)]\!]$$

$$\text{s.t.} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \infty \cdot [\![y_n \neq \text{sign}(\mathbf{w}^T\mathbf{z}_n + b)]\!]$$

- $[\![\cdot]\!]$: non-linear—**not QP anymore :-(** (dual? kernel?)
- cannot distinguish **small error** (slightly away from fat boundary) or **large error** (a...w...a...y... from fat boundary)

- record '**margin violation**' by ξ_n—**linear constraints**
- penalize with **margin violation** instead of **error count** —**quadratic objective**

soft-margin SVM: $$\min_{b,\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} \xi_n$$

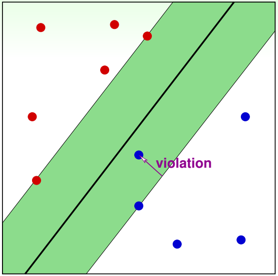
$$\text{s.t.} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \text{ and } \xi_n \geq 0 \text{ for all } n$$

Soft-Margin SVM (2/2)

- record '**margin violation**' by ξ_n
- penalize with **margin violation**

$$\min_{b,\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} \xi_n$$

$$\text{s.t.} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \text{ and } \xi_n \geq 0 \text{ for all } n$$



- parameter C: trade-off of large margin & margin violation
  - large C: want less margin violation
  - small C: want large margin
- QP of $\tilde{d} + 1 + N$ variables, $2N$ constraints

next: remove dependence on $\tilde{d}$ by
soft-margin SVM primal $\Rightarrow$ **dual**?

Fun Time

Lagrange Dual

primal: $$\min_{b,\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} \xi_n$$

s.t. $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n$ and $\xi_n \geq 0$ for all n

Lagrange function with Lagrange multipliers α_n and β_n

$$\begin{aligned}\mathcal{L}(b, \mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad = \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} \xi_n \\ & + \sum_{n=1}^{N} \alpha_n \cdot \left(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)\right) + \sum_{n=1}^{N} \beta_n \cdot (-\xi_n)\end{aligned}$$

want: Lagrange dual

$$\max_{\alpha_n \geq 0,\ \beta_n \geq 0} \left(\min_{b,\mathbf{w},\boldsymbol{\xi}} \quad \mathcal{L}(b, \mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right)$$

Simplify ξ_n and β_n

$$\max_{\alpha_n \geq 0,\ \beta_n \geq 0} \left(\min_{b,\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{n=1}^{N} \xi_n + \sum_{n=1}^{N} \alpha_n \cdot \left(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)\right) + \sum_{n=1}^{N} \beta_n \cdot (-\xi_n) \right)$$

- $\frac{\partial \mathcal{L}}{\partial \xi_n} = 0 = C - \alpha_n - \beta_n$
- no loss of optimality if solving with implicit constraint $\beta_n = C - \alpha_n$ and explicit constraint $0 \leq \alpha_n \leq C$: β_n removed

ξ can also be removed :-), like how we removed b

$$\max_{0 \leq \alpha_n \leq C,\ \beta_n = C - \alpha_n} \left(\min_{b,\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n\left(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)\right) \cancel{+ \sum_{n=1}^{N} (C - \alpha_n - \beta_n) \cdot \xi_n} \right)$$

Other Simplifications

$$\max_{0 \le \alpha_n \le C,\ \beta_n = C - \alpha_n} \left(\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)\right) \right)$$

familiar? :-)

- inner problem **same as hard-margin SVM**
- $\frac{\partial \mathcal{L}}{\partial b} = 0$: no loss of optimality if solving with constraint $\sum_{n=1}^{N} \alpha_n y_n = 0$
- $\frac{\partial \mathcal{L}}{\partial w_i} = 0$: no loss of optimality if solving with constraint $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n$

standard dual can be derived
using the same steps as Lecture 18

Standard Soft-Margin SVM Dual

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} y_n\alpha_n = 0;$$

$$0 \leq \alpha_n \leq C, \text{for } n = 1, 2, \ldots, N;$$

$$\text{implicitly} \quad \mathbf{w} = \sum_{n=1}^{N}\alpha_n y_n \mathbf{z}_n;$$

$$\beta_n = C - \alpha_n, \text{for } n = 1, 2, \ldots, N$$

—only difference to hard-margin: upper bound on α_n

another (convex) **QP**,
with N variables & $2N+1$ constraints

Fun Time

Kernel Soft-Margin SVM

Kernel Soft-Margin SVM Algorithm

1. $q_{n,m} = y_n y_m K(\mathbf{x}_n, \mathbf{x}_m)$; $\mathbf{c} = -\mathbf{1}_N$; $(\mathrm{P}, \mathbf{r})$ for equ./lower-bound/**upper-bound** constraints
2. $\boldsymbol{\alpha} \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{c}, \mathrm{P}, \mathbf{r})$
3. $b \leftarrow ?$
4. return SVs and their α_n as well as b such that for new $\mathbf{x}$,

$$g_{\mathrm{SVM}}(\mathbf{x}) = \mathrm{sign}\left(\sum_{\text{SV indices } n} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b \right)$$

- almost the same as hard-margin
- more flexible than hard-margin
  —**primal/dual always solvable**

remaining question: step 3?

Solving for b

hard-margin SVM

complementary slackness:

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{x}_n + b)) = 0$$

- SV ($\alpha_m > 0$)
  $\Rightarrow b = y_m - \mathbf{w}^T\mathbf{x}_m$

soft-margin SVM

complementary slackness:

$$\alpha_n(1 - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) = 0$$

$$(C - \alpha_n)\xi_n = 0$$

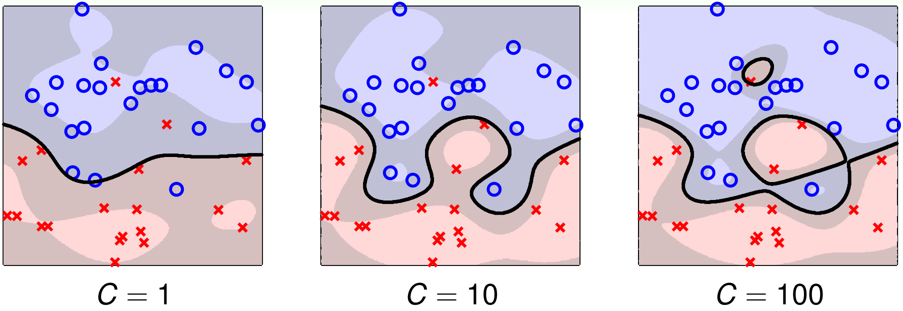
- SV ($\alpha_m > 0$)
  $\Rightarrow b = y_m - y_m\xi_m - \mathbf{w}^T\mathbf{x}_m$
- unbounded ($\alpha_m < C$)
  $\Rightarrow \xi_m = 0$

solve unique b with unbounded SV $(\mathbf{x}_m, y_m)$:

$$b = y_m - \sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}_m)$$

—range of b otherwise

Soft-Margin Gaussian SVM in Action



- large $C \Longrightarrow$ less noise tolerance $\Longrightarrow$ 'overfit'?
- **warning: SVM can still overfit :-(**

soft-margin Gaussian SVM:
need careful selection of (γ, C)

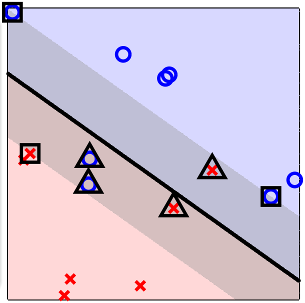
Physical Meaning of α_n

complementary slackness:

$$\alpha_n(1-\xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) = 0$$

$$(C - \alpha_n)\xi_n = 0$$

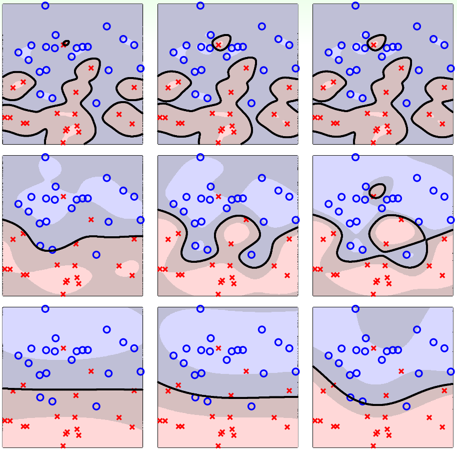
- non SV ($0 = \alpha_n$): $\xi_n = 0$, 'away from'/on fat boundary
- □ unbounded SV ($0 < \alpha_n < C$): $\xi_n = 0$, on fat boundary, locates b
- △ bounded SV ($\alpha_n = C$): $\xi_n =$ violation amount, 'violate'/on fat boundary



α_n can be used for **data analysis**

Fun Time

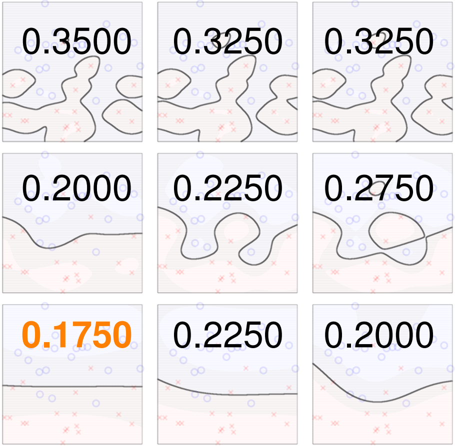
Practical Need: Model Selection



- complicated even for (C, γ) **of Gaussian SVM**
- more combinations if including other kernels or parameters

how to select? **validation :-)**

Selection by Cross Validation



- $E_{cv}(C, \gamma)$: 'non-smooth' function of (C, γ) —**difficult to optimize**
- proper models can be chosen by ***V*-fold cross validation** on **a few grid values of (C, γ)**

E_{cv}: very popular criteria for soft-margin SVM

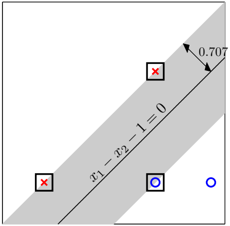
Leave-One-Out CV Error for SVM

recall: $E_{\text{loocv}} = E_{\text{cv}}$ with N folds

claim: $E_{\text{loocv}} \le \frac{\#\text{SV}}{N}$

- for $(\mathbf{x}_N, y_N)$: if optimal $\alpha_N = 0$ (non-SV)
  $\Longrightarrow (\alpha_1, \alpha_2, \ldots, \alpha_{N-1})$ still optimal when leaving out $(\mathbf{x}_N, y_N)$
  key: **what if there's better** α_n?
- SVM: $g^- = g$ when leaving out non-SV

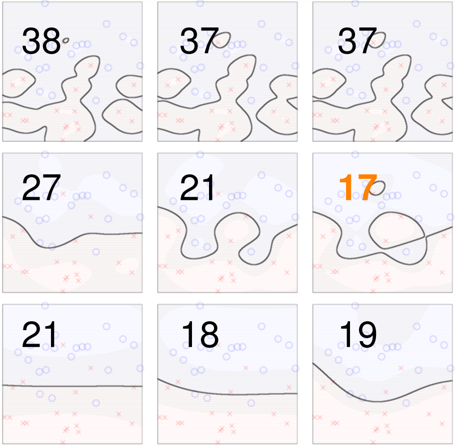
$$\begin{aligned} e_{\text{non-SV}} &= \text{err}(g^-, \text{non-SV}) \\ &= \text{err}(g, \text{non-SV}) = 0 \\ e_{\text{SV}} &\le 1 \end{aligned}$$



motivation from hard-margin SVM: only **SVs needed**

scaled $\#$SV bounds leave-one-out CV error

Selection by # SV



- nSV(C, γ): 'non-smooth' function of (C, γ) —**difficult to optimize**
- **just an upper bound!**
- dangerous models can be ruled out by **nSV** on **a few grid values of (C, γ)**

nSV: often used as a **safety check** if computing E_{cv} not allowed

Fun Time

Summary

Lecture 4: Soft-Margin SVM

- Soft-Margin SVM: Primal
  add margin violations ξ_n
- Soft-Margin SVM: Dual
  adds upper bound to α_n
- Soft-Margin SVM: Solution
  formulated by bounded/unbounded SVs
- Soft-Margin SVM: Selection
  cross-validation, or approximately nSV