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)
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 :-(

- part of reasons: Φ
- other part: separable

if always insisting on separable (⇒ shatter), have power to overfit to noise
Give Up on Some Examples

want: *give up* on some noisy examples

**Pocket**

$$\min_{b,w} \sum_{n=1}^{N} \left[ y_n \neq \text{sign}(w^T z_n + b) \right]$$

**Hard-Margin SVM**

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

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

**Combination**

$$\min_{b,w} \frac{1}{2} w^T w + C \cdot \sum_{n=1}^{N} \left[ y_n \neq \text{sign}(w^T z_n + b) \right]$$

s.t. $$y_n(w^T x_n + b) \geq 1 \text{ for correct } n$$

$$y_n(w^T x_n + b) \geq -\infty \text{ for incorrect } n$$

*C*: trade-off of large margin & noise tolerance
Soft-Margin SVM (1/2)

\[
\begin{align*}
\min_{b,w} & \quad \frac{1}{2} w^T w + C \cdot \sum_{n=1}^{N} \left[ y_n \neq \text{sign}(w^T z_n + b) \right] \\
\text{s.t.} & \quad y_n(w^T x_n + b) \geq 1 - \infty \cdot \left[ y_n \neq \text{sign}(w^T z_n + b) \right]
\end{align*}
\]

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

- record ‘\textbf{margin violation}’ by \(\xi_n\)—\textbf{linear constraints}
- penalize with \textbf{margin violation} instead of \textbf{error count}
  —\textbf{quadratic objective}

soft-margin SVM: \[
\begin{align*}
\min_{b,w,\xi} & \quad \frac{1}{2} w^T w + C \cdot \sum_{n=1}^{N} \xi_n \\
\text{s.t.} & \quad y_n(w^T x_n + b) \geq 1 - \xi_n \text{ and } \xi_n \geq 0 \text{ for all } n
\end{align*}
\]
Soft-Margin SVM (2/2)

- record ‘margin violation’ by $\xi_n$
- penalize with margin violation

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

s.t. $y_n(w^T x_n + b) \geq 1 - \xi_n$ and $\xi_n \geq 0$ 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
Soft-Margin SVM

**Soft-Margin SVM: Dual**

---

**Lagrange Dual**

primal:
\[
\min_{b,w,\xi} \frac{1}{2} w^T w + C \cdot \sum_{n=1}^{N} \xi_n \\
\text{s.t.} \quad y_n(w^T x_n + b) \geq 1 - \xi_n \text{ and } \xi_n \geq 0 \text{ for all } n
\]

Lagrange function with Lagrange multipliers \(\alpha_n\) and \(\beta_n\)
\[
\mathcal{L}(b, w, \xi, \alpha, \beta) = \frac{1}{2} w^T w + C \cdot \sum_{n=1}^{N} \xi_n \\
+ \sum_{n=1}^{N} \alpha_n \cdot (1 - \xi_n - y_n(w^T x_n + b)) + \sum_{n=1}^{N} \beta_n \cdot (-\xi_n)
\]

want: Lagrange dual

\[
\max_{\alpha_n \geq 0, \beta_n \geq 0} \left( \min_{b,w,\xi} \mathcal{L}(b, w, \xi, \alpha, \beta) \right)
\]
Simplify $\xi_n$ and $\beta_n$

$$\max_{\alpha_n \geq 0, \beta_n \geq 0} \left( \min_{b, w, \xi} \frac{1}{2} w^T w + C \cdot \sum_{n=1}^{N} \xi_n ight)$$

$$+ \sum_{n=1}^{N} \alpha_n \cdot (1 - \xi_n - y_n(w^T x_n + b)) + \sum_{n=1}^{N} \beta_n \cdot (-\xi_n)$$

• $\frac{\partial L}{\partial \xi_n} = 0 = C - \alpha_n - \beta_n$

• no loss of optimality if solving with implicit constraint $\beta_n = C - \alpha_n$ and explicit constraint $0 \leq \alpha_n \leq C$: $\beta_n$ removed

$\xi$ can also be removed :-), like how we removed $b$

$$\max_{0 \leq \alpha_n \leq C, \beta_n = C - \alpha_n} \left( \min_{b, w, \xi} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n(1 - y_n(w^T z_n + b)) ight)$$

$$+ \sum_{n=1}^{N} (C - \alpha_n - \beta_n) \cdot \xi_n$$
Other Simplifications

\[
\max_{0 \leq \alpha_n \leq C, \ \beta_n = C - \alpha_n} \left( \min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n (1 - y_n (w^T z_n + b)) \right)
\]

familiar? :-)

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

\[
w = \sum_{n=1}^{N} \alpha_n y_n z_n
\]

standard dual can be derived using the same steps as Lecture 18
Standard Soft-Margin SVM Dual

\[
\begin{aligned}
\min_{\alpha} & \quad \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m z_n^T z_m - \sum_{n=1}^{N} \alpha_n \\
\text{subject to} & \quad \sum_{n=1}^{N} y_n \alpha_n = 0; \\
& \quad 0 \leq \alpha_n \leq C, \text{ for } n = 1, 2, \ldots, N; \\
\text{implicitly} & \quad w = \sum_{n=1}^{N} \alpha_n y_n z_n; \\
& \quad \beta_n = C - \alpha_n, \text{ for } n = 1, 2, \ldots, N
\end{aligned}
\]

—only difference to hard-margin: upper bound on \( \alpha_n \)

another (convex) \( \text{QP} \),
with \( N \) variables & \( 2N + 1 \) constraints
Fun Time
Kernel Soft-Margin SVM

Kernel Soft-Margin SVM Algorithm

1. $q_{n,m} = y_n y_m K(x_n, x_m); \ c = -1^N; (P, r)$ for
   equ./lower-bound/upper-bound constraints

2. $\alpha \leftarrow \text{QP}(Q, c, P, r)$

3. $b \leftarrow ?$

4. return SVs and their $\alpha_n$ as well as $b$ such that for new $x$,
   \[
g_{\text{SVM}}(x) = \text{sign} \left( \sum_{\text{SV indices } n} \alpha_n y_n K(x_n, x) + b \right)
   \]

- 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:
\[ \alpha_n (1 - y_n (w^T x_n + b)) = 0 \]

- SV ($\alpha_m > 0$)
  \[ b = y_m - w^T x_m \]

**soft-margin SVM**

complementary slackness:
\[ \alpha_n (1 - \xi_n - y_n (w^T x_n + b)) = 0 \]
\[ (C - \alpha_n) \xi_n = 0 \]

- SV ($\alpha_m > 0$)
  \[ b = y_m - y_m \xi_m - w^T x_m \]
- unbounded ($\alpha_m < C$)
  \[ \xi_m = 0 \]

solve unique $b$ with unbounded SV ($x_m, y_m$):
\[ b = y_m - \sum_{n=1}^{N} \alpha_n y_n K(x_n, x_m) \]

—range of $b$ otherwise
Soft-Margin SVM: Solution

Soft-Margin Gaussian SVM in Action

- Large $C \Rightarrow$ less noise tolerance $\Rightarrow$ ‘overfit’?
- Warning: SVM can still overfit :-(

Soft-margin Gaussian SVM:
Need careful selection of $(\gamma, C)$
Physical Meaning of $\alpha_n$

complementary slackness:

$$\alpha_n (1 - \xi_n - y_n(w^T x_n + b)) = 0$$

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

- 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 ($\alpha_n = C$): $\xi_n = \text{violation amount}$, ‘violate’/on fat boundary

$\alpha_n$ can be used for data analysis
Fun Time
Practical Need: Model Selection

- complicated even for $(C, \gamma)$ 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 $(C, \gamma)$ — difficult to optimize
- proper models can be chosen by $V$-fold cross validation on a few grid values of $(C, \gamma)$

$E_{cv}$: 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: $E_{\text{loocv}} \leq \frac{\#SV}{N}$

- for $(x_N, y_N)$: if optimal $\alpha_N = 0$ (non-SV)
  $\implies (\alpha_1, \alpha_2, \ldots, \alpha_{N-1})$ still optimal when leaving out $(x_N, y_N)$

key: what if there’s better $\alpha_n$?

- SVM: $g^- = g$ when leaving out non-SV

$$e_{\text{non-SV}} = \text{err}(g^-, \text{non-SV})$$
$$= \text{err}(g, \text{non-SV}) = 0$$
$$e_{\text{SV}} \leq 1$$

motivation from hard-margin SVM: only SVs needed

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

- $n_{SV}(C, \gamma)$: ‘non-smooth’ function of $(C, \gamma)$ —difficult to optimize
- just an upper bound!
- dangerous models can be ruled out by $n_{SV}$ on a few grid values of $(C, \gamma)$

$n_{SV}$: often used as a safety check if computing $E_{cv}$ not allowed
Fun Time
Lecture 4: Soft-Margin SVM

- **Soft-Margin SVM: Primal**
  - add margin violations $\xi_n$

- **Soft-Margin SVM: Dual**
  - adds upper bound to $\alpha_n$

- **Soft-Margin SVM: Solution**
  - formulated by bounded/unbounded SVs

- **Soft-Margin SVM: Selection**
  - cross-validation, or approximately nSV