v-Support Vector Machine as Conditional Value-at-Risk Minimization

Akiko Takeda (Keio University)
Masashi Sugiyama (Tokyo Institute of Technology)
Binary Classification Problem

Find a decision function $f(x) = \hat{w}^\top x + \hat{b}$ based on given training samples $(x_1, y_1), \ldots, (x_m, y_m)$ to correctly classify new samples.

EX.) diagnosis of diabetes

$x_i \in \mathbb{R}^n$ medical examination

$y_i \in \{\pm 1\}$ tested positive/negative

$i \in M := \{1, 2, \ldots, m\}$
C-SVC (C-support vector classification)

\[
\min_{\mathbf{w}, b, z} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} z_i
\]

s.t. \( y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - z_i, \quad (i \in M) \)
\( z \geq 0 \)

Two conflicting goals

- minimizing the training error
- maximizing the margin

- the trade-off between these goals is controlled by C
- C is a rather unintuitive parameter
**ν-SVC**

Scholkopf, Smola, Williamson & Bartlett ('00)

\[ \min_{w, b, z, \rho} \frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{m} \sum_{i=1}^{m} z_i \]

s.t. \[ y_i(w^T x_i + b) \geq \rho - z_i \quad (i \in M) \]

\[ z \geq 0, \quad \rho \geq 0 \]

**C** is replaced by an intuitive parameter **ν**

- C-SVC with \( C = \left| \frac{1}{m \rho^*} \right| \) ↔ ν-SVC
- ν controls # of margin errors and SVs
  \( \left( \frac{\# \text{MEs}}{m} \leq \nu \leq \frac{\# \text{SVs}}{m} \right) \)
- admissible values of ν are limited
  \( \nu \in (\nu_{\text{min}}, \nu_{\text{max}}] \subseteq (0, 1] \)
Chang & Lin (‘01) showed that for
the opt. solutions of $\nu$-SVC and $C$-SVC are the same.

$$\nu_{\text{max}} = 2 \frac{\min(m_+, m_-)}{m}$$

Admissible values of $\nu$

Chang & Lin (‘01) showed that for $\nu \in ([\nu_{\text{min}}, \nu_{\text{max}}] \subseteq (0, 1]$, the opt. solutions of $\nu$-SVC and $C$-SVC are the same.

- $\nu = 0$: 0 optimal value
  - $(w^* = 0, b^* = 0, \rho^* = 0)$

- $\nu_{\text{min}}$ to $\nu_{\text{max}}$: $\nu$-SVC has no solution (unbounded)

- $\nu = 1$: modification up to 1 is trivial
  - The cost is modified to balance the classes.
  - The separation with the smallest negative margin is found

Perez-Cruz, Weston, Hermann & Scholkopf (‘03)
Parameter $\nu$ (diagnosis of diabetes)

Training error ... error rates for training samples
Test error ... error rates for test samples

good prediction with $\nu$-SVC

Test error

Training error

$\nu = \nu_{\text{min}}$

$\nu$-SVC

Perez-Cruz, Weston, Hermann & Scholkopf ('03)
Extended \( \nu \)-SVC (E\( \nu \)-SVC)

Perez-Cruz, Weston, Hermann & Scholkopf ('03)

\[
\begin{align*}
\min_{w,b,z,\rho} & \quad -\nu \rho + \frac{1}{m} \sum_{i=1}^{m} z_i \\
\text{s.t.} & \quad y_i (w^T x_i + b) \geq \rho - z_i, \quad (i \in M) \\
& \quad z \geq 0, \quad w^T w = 1
\end{align*}
\]

Nonconvex optimization

- The margin \( \rho^* \) is negative for \( \nu \in (0, \nu_{\min}] \).
- A non-trivial solution is obtained even for the range.
- An iterative algorithm was proposed for a (local ?) solution.
Two Open Issues

• How the negative margin influences on the generalization performance

• How nice properties the Ev-SVC algorithm has
  – in terms of local optimality, finite convergence etc.…

The goal is to give new theoretical insights into these issues.
Outline

• Overview of SVCs for binary classification
  – C-SVC, ν-SVC, Ev-SVC

• New interpretation of Ev-SVC
  – Ev-SVC as minimization of the Conditional Value-at-Risk (CVaR)
    – Minimizing the CVaR improves generalization performance of classifiers

• A new efficient optimization procedure for Ev-SVC
\[
\begin{align*}
\langle \hat{w}, x \rangle + \hat{b} &> 0 \\
\langle \hat{w}, x \rangle + \hat{b} &< 0 \\
\langle \hat{w}, x \rangle + \hat{b} &= 0
\end{align*}
\]

\[g(\hat{w}, \hat{b}; x_i, y_i) = -y_i(\hat{w}^\top x_i + \hat{b})/\|\hat{w}\|\]

\[y_i(\hat{w}^\top x_i + \hat{b}) > 0\]

- \(g < 0\) correctly classified,
- \(g > 0\) misclassified

\[
\begin{align*}
g(\hat{w}, \hat{b}; x_1, y_1) &\quad g(\hat{w}, \hat{b}; x_2, y_2) \\
1 &\quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7
\end{align*}
\]
How to find \((w, b)\) of the hyperplane?

- Consider a histogram of risk scores:
  \[ g(w, b; x_i, y_i), \quad i = 1, \ldots, m \]

- Find a solution \((w, b)\) which minimizes the mean risk score over a set of "bad" training samples.

\[ \text{CVaR: } \phi_{\beta}(w, b) \]

- High risk of misclassification
- "Bad" training samples with the worst \((1-\beta) \times 100\%\) scores

\[ 100\beta - \text{percentile} \]

\[ \alpha_{\beta}(w, b) \]

\[ \text{mean risk score (CVaR)} \]

\[ \phi_{\beta}(w, b) \]

\[ \Rightarrow \text{Minimize} \]
CVaR Minimization

100\(\beta\) -percentile
\(\alpha_\beta(w, b)\)

bad training samples
\(g(w, b; x, y)\)

Density

risk measure in finance

100\(\beta\) -percentile
\(\alpha_\beta(w, b)\)

Minimize

Rockafellar-Uryasev (1999)

\[
\min_{w, b} \phi_\beta(w, b) = \min_{w, b, \alpha} \alpha + \frac{1}{1 - \beta} \mathbb{E} \left[ [g(w, b; x, y) - \alpha]^+ \right]
\]

optimal sol. : \(\alpha^* = \alpha_\beta(w^*, b^*)\)

\([X]^+ := \max\{X, 0\}\)
New interpretation of Ev-SVC

CVaR minimization: \[
\min_{\mathbf{w}, b, z, \alpha} \quad \phi_{\beta}(\mathbf{w}, b)
\]
\[
\min_{\mathbf{w}, b} \quad \alpha + \frac{1}{(1 - \beta)m} \sum_{i=1}^{m} z_i
\]
s.t. \[
z_i + y_i (\mathbf{w}^\top \mathbf{x}_i + b) + \alpha \geq 0 \quad (i = 1, \ldots, m)
\]
\[
z \geq 0, \quad \mathbf{w}^\top \mathbf{w} = 1
\]

Theorem: Ev-SVC is equivalent to CVaR minimization (minCVaR) with \( \beta = 1 - \nu \)

\[\rho^* = -\alpha_{1-\nu}\]
\[\rho^* : \text{margin of Ev-SVC}\]
negative margin \( \Rightarrow \alpha_{1-\nu} > 0\)
Convex / Nonconvex Program

(Ev-SVC)

\[
\begin{align*}
\min_{w,b,z,\rho} & \quad -\nu \rho + \frac{1}{m} \sum_{i=1}^{m} z_i \\
\text{s.t.} & \quad z_i + y_i(w^T x_i + b) - \rho \geq 0 \quad (i = 1, \ldots, m) \\
& \quad z \geq 0, \quad w^T w = 1
\end{align*}
\]

\[w^T w \geq 1\]

\[w^T w \leq 1\]

Nonconvex Program

\[\phi_{1-\nu}(w^*, b^*) > 0\]

Convex Program

\[\phi_{1-\nu}(w^*, b^*) = 0\]

\(\nu = 0\)

(unknown a-priori)

\(\nu = \hat{\nu}\)

\[\phi_{1-\nu}(w^*, b^*) = 0\]

\(\nu = 1\)

opt. val of (Ev-SVC) \(\phi_{1-\nu}(w^*, b^*)\) is decreasing w.r.t \(\nu \in (0, 1)\)
Outline

• Overview of SVCs for binary classification
  – C-SVC, ν-SVC, Ev-SVC

• New interpretation of Ev-SVC
  – Ev-SVC as minimization of the Conditional Value-at-Risk (CVaR)
  – Minimizing the CVaR improves generalization performance of classifiers

• A new efficient optimization procedure for Ev-SVC
New generalization error bounds of $E_{\nu}$-SVC include the CVaR risk measure. Minimizing the CVaR lowers the bound, and it justifies the use of $E_{\nu}$-SVC & $\nu$-SVC.

- **Case 1**: $\nu = 0$
  - Convex Program: $\phi_{1-\nu}(w^*, b^*) < 0$

- **Case 2**: $\nu = \nu$ (approximately)
  - Nonconvex Program: $\phi_{1-\nu}(w^*, b^*) < 0$

- **Case 3**: $\nu = 1$
  - Nonconvex Program: $\phi_{1-\nu}(w^*, b^*) < 0$
**Generalization Error Bound (case 1)**

**Theorem:** (case 1)

For a feasible sol. \((w, b)\) of \((\nu\text{-SVC})\), the inequality:

\[
    \text{(generalization error with } f(x) = w^\top x + b) \\
    \leq \nu + \sqrt{\frac{2}{m}} \left( \frac{4c^2(1 + B_R^2)^2}{\phi_{1-\nu}(w, b)^2} \log_2(2m) - 1 + \log \left( \frac{2}{\delta} \right) \right)
\]

holds with probability at least \(1 - \delta\)

\(\Rightarrow\) CVaR min. gives an opt. solution which minimizes the bound.

\(\Rightarrow\) \(\nu\text{-SVC}\) is reasonable.

**Proof**

Bartlett ('98), Williamson et al ('01)  
Rockafellar & Uryasev ('02)

Generalization error analysis + results of CVaR min.
Generalization Error Bound (cases 2&3)

For a feasible sol. \((w, b)\) of \((E_{\nu}-SVC)\)

\[
\nu + \sqrt{\frac{2}{m}} \left( \frac{4c^2(1 + B^2_R)^2}{\alpha_{1-\nu}(w, b)^2} \log_2(2m) - 1 + \log \left( \frac{2}{\delta} \right) \right)
\]
holds with probability at least \(1 - \delta\)

Find an opt. sol \((w, b)\) minimizing \(\phi_{1-\nu}(w, b)\)

\[
\nu - \sqrt{\frac{2}{m}} \left( \frac{4c^2(1 + B^2_R)^2}{\phi_{1-\nu}(w, b)^2} \log_2(2m) - 1 + \log \left( \frac{2}{\delta} \right) \right)
\]
holds with probability at least \(1 - \delta\)
Outline

• Overview of SVCs for binary classification
  – C-SVC, \(\nu\)-SVC, Ev-SVC

• New interpretation of Ev-SVC

• A new efficient optimization procedure for Ev-SVC
  – Modification for a local optimum search algorithm of Ev-SVC
  – Global optimum search algorithm
  – Finite convergence
2-Step Algorithm for $\nu$-SVC

Step 1. Solve ($\nu$-SVC) via e.g., SMO, and obtain $$(\hat{w}, \hat{b}, \hat{\alpha}, \hat{z})$$.

If $\phi_{1-\nu}(\hat{w}, \hat{b}) < 0$, terminate ($\nu > \hat{\nu}$).

Step 2. Otherwise, find a global optimum of $\nu$-SVC with a global optimization algorithm, which consists of local optimum search & cutting plane method

$\Rightarrow$ global optimal algorithm with finite convergence

$E\nu$-SVC: Nonconvex Program Convex Program

$\nu = 0$ (unknown a-priori) $\nu = \hat{\nu}$ $\nu = 1$

equivalent to $\nu$-SVC
At the $k$th iteration, the following linear program is solved

\[
\min_{\mathbf{w}, b, z, \rho} \quad -\nu \rho + \frac{1}{m} \sum_{i=1}^{m} z_i \\
\text{s.t.} \quad z_i + y_i (\mathbf{w}^\top \mathbf{x}_i + b) - \rho \geq 0 \quad (i = 1, \ldots, m) \\
z \geq 0, \quad \hat{\mathbf{w}}_{k-1}^\top \mathbf{w} = 1
\]

Space of $\mathbf{w}$

\[
\hat{\mathbf{w}}_1^\top \mathbf{w} = 1, \quad \hat{\mathbf{w}}_1 = \frac{\mathbf{w}_1^*}{\|\mathbf{w}_1^*\|}
\]

Linearization

\[
\text{obj}(\hat{\mathbf{w}}_0) \geq \text{obj}(\mathbf{w}_1^*) > \text{obj}(\hat{\mathbf{w}}_1) \geq \ldots.
\]

Objective values are decreased

\[
\Rightarrow \text{finite convergence}
\]

KKT optimality condition of (LP)

\[
\Rightarrow \text{local optimality}
\]
Properties of Local Optimization Algorithm

✓ After finite steps, we have a converged solution.
✓ The solution is a local minimizer.

Linearized surface for \( (LP_{k+1}) \) is updated at

(Our Algorithm)
\[
\hat{w}_k = \frac{w_k^*}{\|w_k^*\|}
\]

(Ev-SVC Algorithm)  
Perez-Cruz et al. (‘03)
\[
\hat{w}_k = \gamma \hat{w}_{k-1} + (1 - \gamma) w_k^* \\
e.g., \gamma = \frac{9}{10}
\]
finite convergence ???
Cutting Planes for Global Optimization

After finding a local solution, narrow the feasible set of $E^v$-SVC by adding a concavity cut / facial cut

Concavity cut cuts off a corner of $D$
Facial cut removes a face of $D$

Add cuts to $E^v$-SVC until $\tilde{D}$ includes no faces of $D$

# of faces of $D$ is finite

$D$ = feasible set of $E^v$-SVC
$\tilde{D} = D +$ three additional cuts

Global sol. is a corner of $D$

→ finite convergence
The global opt. algorithm: finite convergence, but not fast

Only concavity cuts were added to $E_{\nu}$-SVC

- There are a few local solutions.
- $E_{\nu}$-SVC with small $\nu$ requires many cuts.
- The first local sol. is sufficiently good.

Liver-disorders (UCI dataset)
Conclusion

- Ev-SVC (Perez-Cruz et al., ‘03) generalizes well.
  - It is not clear why it generalizes well.
  - Algorithmic properties are not clear.

- We showed Ev-SVC is equivalent to minimizing Conditional Value-at-Risk (CVaR), a popular risk measure in finance.

- We derived generalization error bounds for Ev-SVC, and showed that Ev-SVC reduces the bounds.
  - The use of Ev-SVC and ν-SVC is justified.

- Although Ev-SVC is nonconvex, we gave an algorithm that can find a global solution within finite iterations.