Reinforced multiclass Support Vector Machines

Yufeng Liu
University of North Carolina at Chapel Hill
http://www.unc.edu/~yfliu

Joint work with Ming Yuan (Georgia Tech)
Outline

• Background on Classification
• Large-Margin Classifiers: SVM
• The Regularization Framework
• New Family of Multicategory SVMs
• Fisher Consistency
• Learning Theory
• Numerical Examples and Discussions
General Framework

- Supervised learning: Given training data $\{(x_i, y_i)_{i=1}^{n}\}$, i.i.d. $\sim$ unknown $P(x, y)$.
  - input $x_i \in \mathbb{R}^d$ as predictor;
  - outcome $y_i$ as class.

- Build a prediction model, or classifier:
  - enable us to do prediction.

- Good classifier: accurately predicts the class $y$ for given $x$. 
Classification Methods

- Traditional statistical methods
  Linear/Quadratic Discriminant Analysis, Nearest Neighbor, Logistic Regression, etc.

- Machine learning
  Margins →
  - Boosting (Freund & Schapire, 1997)
  - $\psi$-Learning (Shen et al., 2003, Liu & Shen, 2006); Robust SVM (Wu & Liu, 2007)
Binary Large-Margin Classifiers

- $y \in \{-1, 1\}$;
  Estimate $f(x)$ with classification rule $\text{sign}[f(x)] : \mathbb{R}^d \rightarrow \{-1, 1\}$,
  $\hat{y} = +1$ if $f(x) > 0$ and $\hat{y} = -1$ if $f(x) < 0$.

- $y_i f(x_i)$: functional margin.

- Correction classification if $y_i f(x_i) > 0$.

- The $0 - 1$ loss: $I(y_i f(x_i) > 0)$. \

Optimization of SVM

- SVM solves

\[ \min_{b, w} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} V_{SVM}(y_i f(x_i)). \]

- \( V_{SVM}(u) = [1 - u]_+ \) (Hinge Loss).

- Nonlinear learning can be achieved by basis expansion or kernel learning.
Regularization

- SVM fits into regularization framework.
- Regularization is common in nonparametric statistics.
- Regularization framework:
  \[
  \min_{f} \quad J(f) + C \sum_{i=1}^{n} V(y_{i}f(x_{i})).
  \]
  - Regularization term \(J(f)\): the roughness penalty of \(f\):
    Typical choices \(\frac{1}{2}\|w\|^2\) (\(L_2\) penalty) or
    \(\sum_{j=1}^{d} |w_{j}|\) (\(L_1\) penalty).
  - Loss \(V\): data fit measure.
Large-Margin Loss Functions

- The loss $V(u)$ is typically non-increasing in $u$.
- $u = y_i f(x_i)$.
- SVM: $[1 - u]_+$
  The minimizer $f^*(x) = \text{sign}(p(x) - 1/2)$;
  $p(x) = P(Y = 1|X = x)$ (Lin, 2002).
- PLR: $\log(1 + \exp\{-u\})$
  $f^*(x) = \ln(p(x)/(1 - p(x)))$.
- AdaBoost: $\exp\{-u\}$ (Friedman et al., 2000)
  $f^*(x) = 0.5 \ln(p(x)/(1 - p(x)))$. 
Different Losses

![Different Losses](image)

- 0–1
- SVM
- AdaBoost(Exp)
- Logistic

Loss $V(u)$

- $u$
From Binary to Multicategory

- Require a novel treatment.
- Label: $\{-1, +1\} \rightarrow \{1, 2, \ldots, k\}$.
- $k$-class
  - Construct decision function vector $\mathbf{f} = (f_1, \ldots, f_k)$. ($k = 2$ only one $f$)
  - Classifier: $\arg\max_{j=1,\ldots,k} f_j(x)$. ($k = 2 : \text{sign}(f)$)
- Sum-to-zero constraint $\sum_{j=1}^k f_j(x) = 0$. 
Find $\mathbf{f} = (f_1, f_2, f_3)$ and use $\text{argmax}_j f_j(\mathbf{x})$ to do classification.
Multicategory Framework

- Multiple comparison (Liu & Shen, 2006):
  \[ g(x, y) = \{ f_y(x) - f_j(x), \forall j \neq y \}. \]
  - Compare class \( y \) with rest \( k - 1 \) classes.
  - \( g(x, y) = f_y(x) - f_{3-y}(x) \) when \( k = 2 \).

- \( f \) yields correct classification for \((x, y)\) if \( g(x, y) > 0_{k-1} \).

- 0–1 loss: \( I(\min(g(X, Y)) \leq 0) \).

- Solve
  \[
  \min_f \sum_{j=1}^{k} \frac{\lambda}{2} J(f_j) + \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i)
  \]
Consistency

- Bayes rule

\[ T^*(x) = \arg \max_j P(y = j|x) \]

- Fisher Consistency (Lin, 2002)

\[ \arg \max_j f^* = T^*(x), \text{ where } f^* = \arg \min_f EV(f(x), y) \]

- Classification calibrated (Bartlett et al., 2003)
- Infinite sample consistency (Zhang, 2004)
Existing Formulations

- Goal: force $f_y(x)$ to be big.


$$V(f(x), y) = \sum_{j \neq y} [1 - (f_y - f_j)]_+$$

- Crammer and Singer (2001), Liu and Shen (2006)

$$V(f(x), y) = [1 - (f_y - \max_{j \neq y} f_j)]_+$$

- These formulations are not always Fisher consistent.
One versus Rest

\[ V(f(x), y) = [1 - f_y]^+ + \sum_{j \neq y}[1 + f_j]^+ \]

- Solve \( k \) binary problems
- SVMs commonly used
- Not always Fisher consistent.
Multicategory SVM – Lee et al. (2004)

- \( f = (f_1, \ldots, f_k) \)

\[
V(f(x), y) = \sum_{j \neq y} [1 + f_j(x)]_+ \quad \text{subject to} \quad \sum_j f_j = 0
\]

- Population solution
  - \( f_j^*(x) = k - 1 \) if \( j = \arg\max_c P(y = c|x) \), \(-1\) otherwise
  - Example: \( k = 3 \)

\[
f^* = \begin{cases} (2, -1, -1) & \text{for class 1} \\ (-1, 2, -1) & \text{for class 2} \\ (-1, -1, 2) & \text{for class 3} \end{cases}
\]

- Fisher consistent.
A General Family of MSVM

Reinforced multicategory hinge loss:

\[ V(f(x), y) = \gamma [a - f_y(x)]_+ + (1 - \gamma) \sum_{j \neq y} [1 + f_j(x)]_+ \]

subject to \( \sum_{j=1}^{K} f_j = 0 \)

- Second term encourages \( f_j(x) \to -1 \) for \( j \neq y \)
- First term encourages \( f_y(x) \to a \)
- Without constraint and \( a = 1, \gamma = 1/2 \) – one-versus-rest
- \( \gamma = 0 \) – MSVM of Lee et al. (2004)
• $u = g(f(x), y)$

• $0 - 1$ loss: $I(\min_j u_j \leq 0)$.

• Reinforced loss:
  \[
  \gamma[(k - 1) - \sum_{l=1}^{k-1} u_l/k] + (1 - \gamma) \sum_{j=1}^{k-1} \left[1 + \sum_{l=1}^{k-1} \frac{u_l}{k} - u_j\right]_+.
  \]

• $k = 3$: reinforced losses with $\gamma = 1, 0, 0.5$ correspond to 2, 4, 6 jointing planes.
Figure 1: $0 - 1$ loss, $\gamma = 1, 0, 0.5$. 
• If $\gamma = 0$, the minimizer of $E[V(f(X), Y)]$ is

\[ f_j^*(x) = \begin{cases} k - 1 & \text{if } j = T^*(x); \\ -1 & \text{otherwise} \end{cases} \]

• If $\gamma = 1$, the minimizer of $E[V(f(X), Y)]$ is

\[ f_j^*(x) = \begin{cases} -(k - 1)a & \text{if } j = \arg \min_c P(y = c|x); \\ a & \text{otherwise} \end{cases} \]
**Theorem** Assume $k > 2$. The reinforced hinge loss function under the sum-to-zero constraint is Fisher consistent if and only if $\gamma \leq 1/2$.

- If $\gamma \leq 1/2$, the minimizer of $E[V(f(X), Y)]$ is
  
  \[
  f_j^*(x) = \begin{cases} 
  k - 1 & \text{if } j = T^*(x); \\
  -1 & \text{otherwise}
  \end{cases}
  \]

- If $\gamma > 1/2$, there exists
  \[
  P(y = 1|x) > P(y = 2|x) \geq \ldots \geq P(y = k|x)
  \]
  such that the minimizer of $E[V(f(X), Y)]$ satisfies
  
  \[
  f_1^*(x) = f_2^*(x).
  \]
• \( \hat{f} \): Solution of reinforced SVM.

• \( f_V = \arg \min_{f \in \mathcal{F}} E[V(f(X), Y)] \).

• Study convergence rate of \( \hat{f} \) in terms of its excess risk:
  \[
e_V(\hat{f}, f_V) = E[V(\hat{f}(X), Y) - V(f_V(X), Y)].\]

• \( \mathcal{F}^V(\ell) = \{ V(f(x), y) - V(f_V(x), y) : f \in \mathcal{F}(\ell) \} \),
  \( \mathcal{F}(\ell) = \{ f \in \mathcal{F} : J(f) \leq \ell \} \),
  \( J(f) = \frac{1}{2} \sum_{j=1}^{k} J(f_j) \), and
  \( J_V = \max(J(f_V), 1) \).
Assumptions

• **(Boundedness)** Assume \( x \in S \subset [-B, B]^d \); \( B < \infty \), and there exists a constant \( T > 0 \) such that \( V(f(x), y) \leq T \) for \( \forall f \in \mathcal{F} \).

• **(Variance)** There exist constants \( 0 \leq \beta \leq 1 \) and \( c_1 > 0 \) such that for any sufficiently small \( \delta > 0 \),

\[
\sup_{\{f \in \mathcal{F}: e_{V}(f, f_{V}) \leq \delta\}} \text{Var}(V(f(x), y) - V(f_{V}(x), y)) \leq c_1 \delta^\beta.
\]

• **(Metric entropy)** For some constants \( c_i > 0; i = 2, \cdots, 4 \), there exists some \( \epsilon_n > 0 \),

\[
\sup_{\ell \geq 1} \phi(\epsilon_n, \ell) \leq c_4 n^{1/2},
\]

where \( \phi(\epsilon, \ell) = \int_{c_3 L}^{c_1 \epsilon^2} H_B^{1/2}(w, \mathcal{F}^V(\ell))dw / L \), and \( L = L(\epsilon, \lambda, \ell) = \min(\epsilon^2 + \lambda(\ell - 1)J_V, 1) \).
THEOREM Under the previous assumptions, for the estimator $\hat{f}$, there exists a constant $c_5 > 0$ such that

$$P(e_V(\hat{f}, f_V) \geq 2\delta_n^2) \leq 3.5 \exp(-c_5 n (\lambda J_V)^{2-\beta})$$

provided that $\lambda^{-1} \geq 2\delta_n^{-2} J_V$, where $\delta_n^2 = \min(\epsilon_n^2, 1)$. 
Example

- $S = [0, 1]$.
- True conditional probability: $q_1 \geq 2 \max(q_2, q_3)$.

$$(P_1(x), P_2(x), P_3(x)) = \begin{cases} 
(q_1, q_2, q_3) & x \leq 1/3 \\
(q_2, q_1, q_3) & 1/3 < x \leq 2/3 \\
(q_3, q_2, q_1) & x > 2/3
\end{cases}$$

- Bayes classifier yields sets $[0,1/3), [1/3,2/3), \text{ and } [2/3,1]$ for the corresponding three classes.

- $\mathcal{F}_1 = \{f : f_j(x) = \sum_{i=1}^{n} v_{ij} K(x_i, x) + b_j, \sum_{j=1}^{3} f_j(x) = 0, J(f) \leq M_1, \max_j |f_j(x)| \leq M_2, x \in [0, 1]\}$ with Gaussian kernel $K(s, t) = \exp(||s - t||^2/\sigma^2)$.

- From the previous theorem

$$e_V(\hat{f}, f_V) = O_P\left(\frac{\log^2 n}{n}\right)$$
Optimization

\[ f_j(x) = b_j + \sum_{i=1}^{n} K(x, x_i) v_{ij} \]

\[
\min_{\Theta} \frac{\lambda}{2} \sum_{j=1}^{k} v_j^T K v_j + \frac{1}{n} \sum_{i=1}^{n} \left( \gamma [a - b_i - K_i^T v_{y_i}] + 
+ (1 - \gamma) \sum_{j \neq y_i} [1 + b_i + K_i^T v_j]_+ \right),
\]

\[
\text{s.t.} \quad e \sum_{j=1}^{k} b_j + K \sum_{j=1}^{k} v_j = 0.
\]

where \( \Theta = \{ v_j, b_j \}^{k}_{j=1} \), \( K = \{ K(x_i, x_i') \} \), \( e = (1, 1, \cdots, 1)^T \).
Margin Formulation

\[
\begin{align*}
\min_{\Theta, \xi} & \quad \frac{n \lambda}{2} \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{K} \mathbf{v}_j + \sum_{i=1}^{n} \left( \gamma \xi_{iy_i} + (1 - \gamma) \sum_{j \neq y_i} \xi_{ij} \right), \\
\text{s.t.} & \quad \xi_{ij} \geq 0; \ i = 1, \ldots, n, \ j = 1, \ldots, k, \\
& \quad \xi_{iy_i} + (b_{y_i} + \mathbf{K}_{iy_i}^T \mathbf{v}_{y_i} - a) \geq 0; \ i = 1, \ldots, n, \\
& \quad \xi_{ij} - (b_j + \mathbf{K}_{ij}^T \mathbf{v}_j + 1) \geq 0; \ i = 1, \ldots, n, \ j \neq y_i, \\
& \quad \left( \sum_{j=1}^{k} b_j \right) e + \mathbf{K} \left( \sum_{j=1}^{k} \mathbf{v}_j \right) = 0.
\end{align*}
\]
Dual Problem

\[ v_j = \frac{1}{n\lambda} \left[ (\alpha_j \cdot (e - L_j) - \alpha_j \cdot L_j) - (\bar{\alpha} - \bar{\alpha}) \right]. \]

and

\[ \min_{\alpha} \frac{1}{2} \sum_{j=1}^{k} \langle [(\alpha_j \cdot (e - L_j) - \alpha_j \cdot L_j) - (\bar{\alpha} - \bar{\alpha})], K[(\alpha_j \cdot (e - L_j) - \alpha_j \cdot L_j) - (\bar{\alpha} - \bar{\alpha})] \rangle \]

\[ -n\lambda \sum_{i=1}^{n} a\alpha_{iy_i} - n\lambda \sum_{i=1}^{n} \sum_{j \neq y_i} \alpha_{ij} \]

s.t. \[ 0 \leq \alpha_{ij} \leq A_{ij}; i = 1, \ldots, n, j = 1, \ldots, k, \]

\[ 1^T[(\alpha_j \cdot (e - L_j) - \alpha_j \cdot L_j) - (\bar{\alpha} - \bar{\alpha})] = 0 \]
Solve for \(b\)

\[
\min_{b, \eta} \sum_{j=1}^{k} \left( \gamma \eta_{iy_i} + (1 - \gamma) \sum_{j \neq y_i} \eta_{ij} \right)
\]

subject to

\[
\sum_{j=1}^{k} b_j = 0
\]

\[
\eta_{ij} \geq 0; i = 1, \ldots, n, j = 1, \ldots, k,
\]

\[
\eta_{iy_i} + (b_{y_i} + K_i^T v_{.y_i} - a) \geq 0; i = 1, \ldots, n,
\]

\[
\eta_{ij} - (b_j + K_i^T v_{.j} + 1) \geq 0; i = 1, \ldots, n, j \neq y_i.
\]

To sum up

- Obtain \(v\) by solving QP
- Obtain \(b\)'s by solving LP
Numerical Examples

- Examine effect of $\gamma$ on classification accuracy.
- Choice $\gamma = 0, 0.1, \ldots, 1$.
- Use training (model building), tuning (tuning parameter selection) and test sets (evaluation).
- Performance measure: test error.
• $P(Y = 1) = P(Y = 2) = P(Y = 3) = 1/3$,
  $P(X|Y = 1) \sim N(\mu = (0, 2)^T, 1.5^2 I_2)$,
  $P(X|Y = 2) \sim N(\mu = (-\sqrt{3}, -1)^T, 1.5^2 I_2)$,
  $P(X|Y = 3) \sim N(\mu = (\sqrt{3}, -1)^T, 1.5^2 I_2)$.

• Kernel: $K(u, v) = \langle u, v \rangle$, $K(u, v) = (1 + \langle u, v \rangle)^2$.

• The Bayes error is 0.2039.
Test Errors & SDs of Linear Example

Figure 2: Plots of the average estimated test errors & standard errors of reinforced MSVMs based on 100 replications with $\gamma = 0, 0.1, \ldots, 1$. 
Figure 3: Plots of the typical classification boundaries of the reinforced MSVMs with $\gamma = 0, 0.5, 1$ using linear kernel ($n = 50$).
Summary of Linear Example

- Linear kernel: $\gamma = 0$ gives the worst performance.
- Polynomial kernel: $\gamma \in [0, 0.5]$ works better than $\gamma > 0.5$.
- Middle range values of $\gamma$ such as 0.5 give the most accurate and stable performance.
- Fisher consistency is only a pointwise consistency result.
Nonlinear Example

- $P(Y = 1) = P(Y = 2) = P(Y = 3) = 1/3,$
  $P(X|Y = 1) \sim N(\mu = (2, 0)^T, 1.5^2 I_2),$
  $P(X|Y = 2) \sim 0.5N(\mu = (0, 2)^T, 1.5^2 I_2) + 0.5N(\mu = (0, -2)^T, 1.5^2 I_2),$
  $P(X|Y = 3) \sim N(\mu = (-2, 0)^T, 1.5^2 I_2).$

- Kernel: $K(u, v) = (1 + \langle u, v \rangle)^2, K(s, t) = \exp(||s - t||^2/2\sigma^2).$

- The Bayes error is 0.2883.
Figure 4: Plots of the average estimated test errors & standard errors of reinforced MSVMs based on 100 replications with $\gamma = 0, 0.1, \ldots, 1$. 
Figure 5: Plots of the typical classification boundaries of the reinforced MSVMs with $\gamma = 0, 0.5, 1$ using Gaussian kernel ($n = 100$)
Summary of Nonlinear Example

- $\gamma$’s close to 1 give worse accuracy than smaller $\gamma$’s.
- $\gamma = 0.5$ gives the best or close to the best performance.
- Gaussian kernel works slightly better than the polynomial kernel.
Summary and Future Work

- A general class of MSVM that includes Lee et al. (2004)’s MSVM and closely mimics the one-versus-rest SVM
- A continuum of MSVMs in the family is Fisher consistent
- Theoretical convergence rates are available and computation can be done fairly efficiently
- Numerical results indicate $\gamma = 1/2$ performs best
- Application on other large margin losses.
- Automatic variable selection via use of different penalty functions.