Classification with Reject Option

Bartlett and Wegkamp (2008)
Wegkamp and Yuan (2010)

February 17, 2012
Outline

. Introduction
  .. Classification with reject option

. Spirit of the papers – BW2008
  .. Infinite sample consistency
  .. Bounding excess risk
  .. Convergence of excess risk

. Results
  .. Generalized hinge loss (SVM) – BW2008
  .. Further generalization – WY2010
Introduction – reject option

classification:

- observed data:
  \[ \{X, Y\} \in \mathcal{X} \times \{\pm 1\} \]

- discriminant function:
  \[ f : \mathcal{X} \rightarrow \mathbb{R}, \text{ typically classify by } \text{sign}(f) \]

reject option:
withhold for cost \( d \) when \( |f| < \delta \), and proceed as usual if \( |f| > \delta \)
for specified \( d, \delta \)

motivation:
want to avoid making hard decisions
Introduction – loss

loss as a function of $Yf(X)$

typically:

$$\ell(z) = \begin{cases} 
1 & \text{if } z \leq 0 \\
0 & \text{if } z > 0 
\end{cases}$$

reject option:

$$\ell_{d,\delta}(z) = \begin{cases} 
1 & \text{if } z < -\delta \\
0 & \text{if } z \geq \delta \\
d & \text{if } |z| \leq \delta 
\end{cases}$$

for some $d \in [0, 1/2]$, $\delta \in (0, 1)$
Introduction – loss

loss as a function of $Yf(X)$

reject option:

\[ \ell_{d,\delta}(z) = \begin{cases} 
1 & \text{if } z < -\delta \\
0 & \text{if } z \geq \delta \\
d & \text{if } |z| \leq \delta 
\end{cases} \]

for some $d \in [0, 1/2]$, $\delta \in (0, 1)$

- if $d \leq 0$ always reject, if $d > 1/2$ never reject
- intuition: $d$ is necessary amount of confidence for classifying
minimize $L_{d,\delta}(f) = \mathbb{E}\{\ell_{d,\delta}(Yf(X))\}$

$$f_d^*(X) = \begin{cases} +1 & \text{if } \eta(X) > 1 - d \\ 0 & \text{if } d \leq \eta(X) \leq 1 - d \\ -1 & \text{if } \eta(X) < d \end{cases}$$

where $\eta(X) = P(Y = +1|X)$

follows from conditional expectation:

$$\mathbb{E}_{Y|X}\{\ell_{d,\delta}(Yf(X))\} = \eta(X) \ell_{d,\delta}(f(X)) + (1 - \eta(X)) \ell_{d,\delta}(-f(X))$$
Introduction – Bayes rule

minimize \( L_{d, \delta}(f) = \mathbb{E}\{\ell_{d, \delta}(Yf(X))\} \)

\[
f_d^*(X) = \begin{cases} 
+1 & \text{if } \eta(X) > 1 - d \\
0 & \text{if } d \leq \eta(X) \leq 1 - d \\
-1 & \text{if } \eta(X) < d
\end{cases}
\]

where \( \eta(X) = P(Y = +1 | X) \)

in practice, minimize empirical risk:

\[
P_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_{d, \delta}(Y_if(X_i))
\]

minimizing \( P_n \) is NP hard
General “spirit”

▶ propose convex surrogate loss $\phi_d$

consider $\phi_d$ risk: $Q(f) = \mathbb{E}\{\phi_d(Yf(X))\}$

minimize empirical risk: $Q_n(f)$

let $\hat{f}_n = \arg\min_{f \in F} Q_n$

▶ study the performance of $\phi_d$

.. consistency

.. bounding excess risk

.. convergence rates for empirical maximizer
General “spirit”

. consistency
  .. \( f^*_d = \arg\min_f Q(f) \)
  .. asymptotic minimizer of \( \phi_d \) risk is same as for \( \ell_{d,\delta} \) risk
  .. able to recover \( f^*_d \)

. bounding excess risk
  .. want some non-decreasing function \( \rho(\cdot) \) such that:

\[
L_{d,\delta}(f) - L_{d,\delta}(f^*_d) \leq \rho(Q(f) - Q(f^*_d))
\]
\[
\Delta L_{d,\delta}(f) \leq \rho(\Delta Q(f))
\]
  .. guarantee performance of \( f \)

. convergence rate for empirical maximizer
  .. compute rate for \( \Delta Q(\hat{f}_n) \to 0 \)
  .. combine with bound on excess risk
Generalized hinge loss

convex surrogate of the form:

\[ \phi_d(z) = \begin{cases} 
1 - \left(\frac{1-d}{d}\right)z & \text{if } z \leq 0 \\
1 - z & \text{if } 0 < z \leq 1 \\
0 & \text{if } z > 1 
\end{cases} \]

consistency

**Proposition 1** The Bayes discriminant function \( f_d^* \) minimizes risk \( Q \) with:

\[ dQ(f_d^*) = L_{d,\delta}(f_d^*) \]
Generalized hinge loss

convex surrogate of the form:

\[
\phi_d(z) = \begin{cases} 
1 - \left(\frac{1-d}{d}\right)z & \text{if } z \leq 0 \\
1 - z & \text{if } 0 < z \leq 1 \\
0 & \text{if } z > 1 
\end{cases}
\]

excess risk bounds

note: for \( \delta \leq 1 - d \), \( \ell_{d,\phi}(X) \leq \phi_d(X) \)

Theorem 2 for fixed \( f \), \( d \in [0, 1/2) \):

for all \( \delta \in (0, 1/2] \),

\[
\Delta L_{d,\delta}(f) \leq \frac{d}{\delta} \Delta Q(f)
\]

for all \( \delta \in [1/2, 1 - d] \),

\[
\Delta L_{d,\delta}(f) \leq \Delta Q(f)
\]

recommend using \( \delta = 1/2 \)
Generalized hinge loss

convex surrogate of the form:

\[ \phi_d(z) = \begin{cases} 
1 - \left(\frac{1-d}{d}\right)z & \text{if } z \leq 0 \\
1 - z & \text{if } 0 < z \leq 1 \\
0 & \text{if } z > 1 
\end{cases} \]

excess risk bounds

note: for \( \delta \leq 1 - d \), \( \ell_{d,\phi}(X) \leq \phi_d(X) \)

Theorem 2 for fixed \( f \), \( (\delta, d) = (0, 1/2) \):

\[ \Delta L(f) \leq \Delta Q(f) \]

risk bounds for usual classification with hinge loss
Generalized hinge loss

convex surrogate of the form:

\[ \phi_d(z) = \begin{cases} 
1 - \left( \frac{1-d}{d} \right) z & \text{if } z \leq 0 \\
1 - z & \text{if } 0 < z \leq 1 \\
0 & \text{if } z > 1 
\end{cases} \]

**convergence rates**

bounds of the form:

\[ \mathbb{E} Q(\hat{f}_n) - Q(f_d^*) \leq 2 \inf_{f \in \mathcal{F}} (Q(f) - Q(f_d^*)) + \epsilon_n \]

- **approximation error** for space \( \mathcal{F} \)
  - method of sieves
- **estimation error** for finite sample \( n \)
  \( \rightarrow \) authors’ results
Generalized hinge loss

convergence rates

Margin condition: we say that $\eta$ satisfies the margin condition at $d$ with exponent $\alpha > 0$ if there is a $c \geq 1$ such that for all $t > 0$,

$$\mathbb{P}\{|\eta(X) - d| \leq t\} \leq ct^\alpha$$
$$\mathbb{P}\{|\eta(X) - (1 - d)| \leq t\} \leq ct^\alpha$$

Bernstein class: we say that $\mathcal{G} \subset L_2(P)$ is a $(\beta, B)$-Bernstein class with respect to the probability measure $P$, with $\beta \in (0, 1]$ and $B \geq 1$ if every $g \in \mathcal{G}$ satisfies

$$Pg^2 \leq B(Pg)^\beta$$
Generalized hinge loss

convergence rates

**Lemma 8** if $\eta$ satisfies the margin condition at $d$ with exponent $\alpha$, then for any class $\mathcal{F}$ of measurable uniformly bounded functions, the class $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$ has a *Bernstein* exponent $\beta = \frac{\alpha}{1+\alpha}$

where $g_f(x, y) = \phi_d(yf(x)) - \phi_d(yf_d^*(x))$

- the excess $\phi_d$ loss

- $\mathbb{E}(g_f(X, Y)) = Q(f)$
Generalized hinge loss

convergence rates

Theorem 12 if $\eta$ satisfies the margin condition at $d$ with exponent $\alpha$, $\mathcal{F}$ is a countable class of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $\|f\|_{\infty} \leq B$, and $\mathcal{F}$ satisfies

$$\log N(\epsilon, L_{\infty}, \mathcal{F}) \leq C\epsilon^{-p}$$

for all $\epsilon > 0$ and some $0 \leq p \leq 2$, then there exists a constant $C'$ independent of $n$, such that

$$\mathbb{E} Q(\hat{f}_n) - Q(f_d^*) \leq 2 \inf_{f \in \mathcal{F}} (Q(f) - Q(f_d^*)) + C'n^{-\frac{1+\alpha}{2+p+\alpha+p\alpha}}$$
Generalized hinge loss - QCQP

can be solved as the following quadratic constraint quadratic programming problem (QCQP)

**Theorem 5** for any \( x_1, \ldots, x_n \in \mathcal{X} \) and \( y_1, \ldots, y_n \in \{ \pm 1 \} \), let \( \hat{f} \in \mathcal{H} \) be the solution to

\[
\minimize \quad f \mapsto \sum_{i=1}^{n} \phi_d(y_i f(x_i)) \\
\text{such that} \quad \|f\|^2 \leq r^2
\]

where \( r > 0 \)
can be solved as the following quadratic constraint quadratic programming problem (QCQP)

**Theorem 5**

*Then we can represent \( \hat{f} \) as the finite sum\[
\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x_i, x),
\]
where \( \hat{\alpha}_1, \ldots, \hat{\alpha}_n \) is the solution to the following QCQP:

\[
\begin{align*}
\min_{\alpha_i, \xi_i, \gamma_i} & \quad \frac{1}{n} \sum_{i=1}^{n} \left( \xi_i + \frac{1 - 2d}{d} \gamma_i \right) \\
\text{such that} & \quad \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \leq r^2 \\
& \quad \xi_i \geq 0, \quad \xi_i \geq 1 - y_i \sum_{j=1}^{n} \alpha_j K(x_i, x_j) \\
& \quad \gamma_i \geq 0, \quad \gamma_i \geq -y_i \sum_{j=1}^{n} \alpha_j K(x_i, x_j)
\end{align*}
\]
Generalization

generalize to the following loss

\[
\ell[g(X), Y] = \begin{cases} 
1 & \text{if } g(X) \neq Y \text{ and } g(X) \neq 0 \\
1 & \text{if } g(X) = 0 \\
0 & \text{if } g(X) = Y 
\end{cases}
\]

with classification rule

\[
C(f(X), \delta) = \begin{cases} 
+1 & \text{if } f(X) > \delta \\
0 & \text{if } |f(X)| \leq \delta \\
-1 & \text{if } f(X) < -\delta 
\end{cases}
\]

equivalence to previous case can be seen by:

\[
\ell[C(f(X), \delta), Y] \equiv \ell_{d,\delta}(Yf(X))
\]

since looking at a family of surrogates we want to separate \( \delta \) from the loss function
Generalization

**consistency**

**Theorem 1** assume for convex $\phi$, the classification rule $C(f_\phi^*, \delta)$ for some $\delta > 0$ is infinite sample consistent if and only if

1. $\phi'(\delta)$ and $\phi'(-\delta)$ both exist
2. $\phi'(0) < 0$
3. $\frac{\phi'(\delta)}{\phi'(\delta) + \phi'(-\delta)} = d$

- for strictly convex $\phi$, at most one value of $\delta$
- for generalized hinge loss, holds for $\delta \in (0, 1)$
Theorem 9 assume \( \phi \) is a convex infinite sample consistent surrogate and suppose that there exists constants \( C > 0 \) and \( s \geq 1 \) such that

\[
|\eta - d|^s \leq C^s \Delta Q_\eta(-\delta)
\]

\[
|(1 - \eta) - d|^s \leq C^s \Delta Q_\eta(\delta)
\]

then,

\[
\Delta R[C(f, \delta)] \leq 2C[\Delta Q(f)]^{1/s}
\]

where \( Q_\eta(z) = \eta(X)\phi(z) + (1 - \eta(X))\phi(-z) \)

excess risk bounds
Generalization

**excess risk bounds**

**Theorem 10** further assume

\[ 2^s(\eta - 1/2)^s_+ \leq C^s\Delta Q_\eta(-\delta) \]
\[ 2^s((1 - \eta) - 1/2)^s_+ \leq C^s\Delta Q_\eta(\delta) \]

then,

\[ \Delta R[C(f, \delta)] \leq C[\Delta Q(f)]^{1/s} \]

**Theorem 11** if margin condition holds for some \( c \geq 1 \) and \( \alpha \geq 0 \),

then,

\[ \Delta R[C(f, \delta)] \leq K[\Delta Q(f)]^{1/(s+\beta-\beta s)} \]

where \( K \) depends on \( c \) and \( \alpha \), and \( \beta = \alpha/(1 + \alpha) \)
Generalization

convergence rates

**Theorem 18** assume that $|f| \leq B$ for all $f \in \mathcal{F}$ and let $\gamma \in (0, 1)$. with probability at least $1 - \gamma$

$$Q(\hat{f}_n) \leq \inf_{f \in \mathcal{F}} Q(f) + \frac{3L}{n} + 8 \left( \frac{L^2}{2c} + \frac{B}{6} \right) \frac{\log(N_n/\gamma)}{n}$$

where $N_n = N(1/n, L_\infty, \mathcal{F})$

relies on:

- Lipschitz continuity of $\phi$
- constraint on modulus of convexity of $Q$
Corollary 19 under the assumptions of Theorems 9 and 18, we have, with probability at least $1 - \gamma$

$$\Delta R(C(\hat{f}_n, \delta)) \leq 2C\left\{ \inf_{f \in F} \Delta Q(f) + \frac{3L}{n} + 8\left(\frac{L^2}{2c} + \frac{LB}{3}\right)\frac{\log(N_n/\gamma)}{n}\right\}^{1/s}$$

if the margin condition also holds, with probability at least $1 - \gamma$

$$\Delta R(C(\hat{f}_n, \delta)) \leq K\left\{ \inf_{f \in F} \Delta Q(f) + \frac{3L}{n} + 8\left(\frac{L^2}{2c} + \frac{LB}{3}\right)\frac{\log(N_n/\gamma)}{n}\right\}^{1/(s\beta - \beta s)}$$
Miscellanea

other losses considered:

- least squares
- exponential
- logistic
- squared hinge
- DWD

further extensions:

- asymmetric loss
- hinge loss with $L_1$ regularization – YW Bernoulli, 2011