E0 370 Statistical Learning Theory **Lecture 14 (Sep 26, 2013)**

Consistency of Surrogate Risk Minimization Methods for Binary Classification using Classification Calibrated Losses

Lecturer: Shivani Agarwal *Scribe: Gaurav Pandey*

1 Introduction

In the previous lecture, we saw that for a $\lambda-$strongly proper composite loss ψ, it is possible to bound the $0-1$ regret in terms of its $\psi-$regret. Hence, for $\lambda-$strongly proper composite loss ψ, if we have a $\psi-$ consistent algorithm, we can use it to obtain a $0-1$ consistent algorithm. However, not all loss functions used as surrogates in binary classification are proper, the hinge loss being one such example.

In this lecture, we extend the family of loss functions ψ for which $\psi-$ consistency implies $0-1$ consistency. In particular, we show that (under certain continuity assumptions), it is necessary and sufficient for the loss function to be classification calibrated in order to satisfy this property.

2 Classification Calibrated Loss Functions

From the previous lecture, we recall that for a given sample S, finding a prediction model $h_S : \mathcal{X} \to \mathcal{T}$ that minimizes the target loss $l : \mathcal{Y} \times \mathcal{T} \to \mathbb{R}_+$ on S, when the target space is finite, is, in general, computationally hard. Hence, a commonly used approach is to learn a mapping $f_S : \mathcal{X} \to \mathcal{C}$, where $\mathcal{C}$ is some continuous space, that minimizes the surrogate loss $\psi : \mathcal{Y} \times \mathcal{C} \to \mathbb{R}_+$ on S, instead. The prediction model is then given by $h_S = \text{pred} \circ f_S$, where $\text{pred} : \mathcal{C} \to \mathcal{T}$. For binary classification, $\mathcal{Y} = \mathcal{T} = \{\pm 1\}$, we often assume $\mathcal{C} = \mathbb{R}$ and pred = sign, i.e., $h_S(x) = \text{sign}(f_S(x))$.

For a given instance x, let η be the true probability of its label being 1. Furthermore, for a given loss function $l : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$, let L_l and R_l denote the conditional $l-$risk and $l-$regret respectively. Let $S_\eta = \{\alpha \in \mathbb{R} : \text{sign}(\alpha) \notin \arg\min_{\hat{y} \in \{\pm 1\}} L_{0-1}(\eta, \hat{y})\}$. We can rewrite the set as

$$\begin{aligned} S_\eta &= \{\alpha \in \mathbb{R} : L_{0-1}(\eta, sign(\alpha)) > \min_{\hat{y} \in \{\pm 1\}} L_{0,1}(\eta, \hat{y})\} \\ &= \{\alpha \in \mathbb{R} : L_{0-1}(\eta, sign(\alpha)) - \min_{\hat{y} \in \{\pm 1\}} L_{0,1}(\eta, \hat{y}) > 0\} \\ &= \{\alpha \in \mathbb{R} : R_{0-1}(\eta, sign(\alpha)) > 0\} \end{aligned} \tag{1}$$

Definition. A loss function $\psi : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ is called classification calibrated if $\forall \eta \in [0, 1]$

$$\inf_{\alpha \in S_\eta} L_\psi(\eta, \alpha) > \inf_{\alpha \in \mathbb{R}} L_\psi(\eta, \alpha) \tag{2}$$

The RHS in the above equation is the condtional Bayes $\psi-$risk. Hence, by plugging in (1) we can restate the above condition as

$$\inf_{\alpha \in \mathbb{R} : R_{0-1}(\eta, sign(\alpha)) > 0} R_\psi(\eta, \alpha) > 0 \tag{3}$$

If $\eta > .5$, the conditional $0-1$ risk L_{0-1} is optimized by assigining $\hat{y} = +1$. Hence, treating predictions sign(0) for $\eta \neq \frac{1}{2}$ as error, we have

$$\alpha \in S_\eta \Longleftrightarrow \text{sign}(\alpha) \notin \arg\min_{\hat{y}\in\{\pm 1\}} L_{0-1}(\eta, \hat{y}) \iff \alpha \leq 0.$$

Similarly, if $\eta < .5$ (again treating predictions sign(0) for $\eta \neq \frac{1}{2}$ as errors):

$$\alpha \in S_\eta \iff \text{sign}(\alpha) \notin \arg\min_{\hat{y}\in\{\pm 1\}} L_{0-1}(\eta, \hat{y}) \iff \alpha \geq 0.$$

If $\eta = .5$, the set S_η is empty. Combining the above statements, we get

$$S_\eta = \{\alpha \in \mathbb{R} : \alpha(\eta - .5) \leq 0\} \tag{4}$$

Example 1. Logistic loss: $\psi_{\log}(y, \alpha) = \ln(1 + \exp(-y\alpha))$
In the previous lecture, we have already seen that the logistic loss is $\lambda-$strictly proper composite with $\lambda = 4$. Here, we show that the logistic loss is also classification calibrated. In order to prove it, we first notice that conditional $\psi-$risk for the logistic loss given by

$$L_\psi(\eta, \alpha) = \eta \ln(1 + \exp(-\alpha)) + (1 - \eta) \ln(1 + \exp(\alpha)),$$

is strictly convex in α. Hence, it has a unique minimizer α^* which can be obtained by differentiating (since the ψ risk is differentiable in α) the above function and equating it to 0. The corresponding minimizer is given by

$$\alpha^* = \ln\left(\frac{\eta}{1-\eta}\right)$$

From the above equation, if $\eta > .5$, $\alpha^* > 0$ and vice-versa. Hence, from (4), $\alpha^* \notin S_\eta$. Since α^* is the unique minimizer, the minimum value of the $\psi-$ risk over the set S_η will be strictly greater than the value corresponding to α^*. Hence, from (2), the loss functions is classification calibrated.

In fact, it turns out that the above property holds in general for all strictly proper losses when composed with a link function that satisfies certain properties as mentioned in the next theorem.

Theorem 2.1. For any strictly proper loss $l : \{\pm 1\} \times [0, 1] \to \mathbb{R}_+$ and strictly increasing link $\gamma : [0, 1] \to \mathbb{R}$ with $\gamma(\frac{1}{2}) = 0$, the proper composite loss $\psi(y, \alpha) = l(y, \gamma^{-1}(\alpha))$ is classification calibrated.

Proof. We have,

$$\begin{aligned} L_\psi(\eta, \alpha) &= \eta\psi(1, \alpha) + (1 - \eta)\psi(-1, \alpha) \\ &= \eta l(1, \gamma^{-1}(\alpha)) + (1 - \eta) l(-1, \gamma^{-1}(\alpha)) \\ &= L_l(\eta, \gamma^{-1}(\alpha)) \end{aligned} \tag{5}$$

Since, the loss is strictly proper, for fixed η, conditional $l-$risk $L_l(\eta, \hat{\eta})$ is minimized uniquely at $\hat{\eta} = \eta$. Hence, from (5), $\alpha^* = \gamma(\eta)$ is the unique minimizer of conditional $\psi-$risk.

Furthermore, since γ is a strictly increasing function with $\gamma(\frac{1}{2}) = 0$, we have $\eta > \frac{1}{2} \Rightarrow \alpha^* = \gamma(\eta) > 0$ and $\eta < \frac{1}{2} \Rightarrow \alpha^* = \gamma(\eta) < 0$. Hence, $\alpha^*(\eta - \frac{1}{2}) > 0$. Since α^* is also the unique minimizer, we have,

$$\inf_{\alpha\in\mathbb{R}:\alpha(\eta-\frac{1}{2})\leq 0} L_\psi(\eta, \alpha) > \inf_{\alpha\in\mathbb{R}} L_\psi(\eta, \alpha)$$

Hence, ψ is classification calibrated. □

In general, if the link function does not satisfy the property $\gamma(\frac{1}{2}) = 0$, we can define a new link function $\bar{\gamma} : [0,1] \to \mathbb{R}$, such that $\bar{\gamma}(\eta) = \gamma(\eta) - \gamma(\frac{1}{2})$. The strictly proper composite loss obtained by composing the strictly proper loss function with this new link function, will then be classification calibrated.

Next, we show that the reverse is not true, i.e., there exist surrogate loss functions $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$, which are not strictly proper composite, but classification calibrated.

Example 2 (Hinge loss)**.** $\psi_{hinge}(y, \alpha) = (1 - y\alpha)_+$

We have,

$$\begin{aligned} L_{hinge}(\eta, \alpha) &= \eta(1-\alpha)_+ + (1-\eta)(1+\alpha)_+ \\ &= \begin{cases} \eta(1-\alpha), & \text{if } \alpha \leq -1. \\ \eta(1-\alpha) + (1-\eta)(1+\alpha), & \text{if } -1 < \alpha < 1. \\ (1-\eta)(1+\alpha), & \text{if } \alpha \geq 1. \end{cases} \end{aligned} \tag{6}$$

If $\eta < \frac{1}{2}$, the infimum over the individual ranges in (6) is given by $2\eta, 2\eta, 2(1-\eta)$, with corresponding α values being $-1, -1$ and 1, respectively. The infimum over the entire range of α is achieved at $\alpha^* = -1$.

If $\eta > \frac{1}{2}$, the infimum over the individual ranges in (6) is given by $2\eta, 2(1-\eta), 2(1-\eta)$, with corresponding α values being $-1, 1$ and 1, respectively. The infimum over the entire range of α is achieved at $\alpha^* = 1$.

In either case $\alpha^*(\eta - \frac{1}{2}) > 0$, and hence from (4), $\alpha^* \notin S_\eta$. Since, α^* is also the unique minimizer, the conditional $\psi-$risk is strictly greater than conditional Bayes $\psi-$risk, and hence the loss function is classification calibrated.

3 Surrogate regret bounds for classification calibrated loss functions

Theorem 3.1 (Proved for margin based losses by Bartlett et al, 2006 [1] and in more general settings by Zhang, 2004 [3] and Steinwart, 2007 [2])**.** Let $\psi : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ be such that $\psi_y(.) = \psi(y, .)$ is continuous $\forall y \in \{\pm 1\}$. The follwoing are equivalent:

1. ψ is classification calibrated.
2. $\forall$ distributions $D, \text{regret}_D^{\psi}[f_S] \xrightarrow{P} 0 \Rightarrow \text{regret}_D^{0-1}[\text{sign} \circ f_S] \xrightarrow{P} 0$.
3. There exists an increasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ continuous at 0, with $g(0) = 0$, such that for all distributions D, $f : \mathcal{X} \to \mathbb{R}$, $\text{regret}_D^{0-1}[\text{sign} \circ f_S] \leq g(\text{regret}_D^{\psi}[f_S])$

Here, we just prove that (1) $\Rightarrow$ (3). (3) $\Rightarrow$ (2) is obvious. We will not prove (2) $\Rightarrow$ (1) here. In order to prove the above theorem, we need to define a few more quantities.

- **Calibration function (δ_η):** For a given surrogate loss function and fixed η, the calibration function $\delta_\eta : [0,1] \to \mathbb{R}_+$ is defined as

$$\delta_\eta(\epsilon) = \inf_{\alpha \in \mathbb{R} : R_{0-1}(\eta, \text{sign}(\alpha)) \geq \epsilon} R_\psi(\eta, \alpha) \tag{7}$$

  It is easy to observe that $\delta_\eta(0) = 0$. Furthermore, for a classification calibrated loss, from (3), $\delta_\eta(\epsilon) > 0$, for all $\epsilon > 0$. Moreover, as ϵ increases, the infimum has to be taken over a smaller set. Hence, $\delta_\eta(\epsilon)$ is a non-decreasing function in ϵ.

- **Uniform calibration function (δ):** The uniform calibration function $\delta : [0,1] \to \mathbb{R}_+$ is defined as

$$\delta(\epsilon) = \inf_{\eta \in [0,1]} \delta_\eta(\epsilon) \tag{8}$$

  Since $\delta_\eta(0) = 0$ for all $\eta \in [0,1]$, $\delta(0) = 0$. Furthermore, for $\epsilon > 0$, $\delta(\epsilon) \geq 0$.

- **Fenchel Legendre biconjugate of δ (δ^{**}):** Fenchel Legendre biconjugate of δ is defined as the function whose epigraph is the closed convex hull of epigraph of δ. In other words, δ^{**} is the largest lower semicontinuous convex function that satisfies $\delta^{**} \leq \delta$.

Theorem 3.2. For all distributions D, and for all functions $f : \mathcal{X} \to \mathbb{R}$

$$\delta^{**}(\text{regret}_D^{0-1}[\text{sign} \circ f]) \leq \text{regret}_D^{\psi}[f]$$

Proof.

$$\begin{aligned} LHS &= \delta^{**}(\mathbb{E}_X[R_{0,1}(\eta(x), sign(f(x)))]) \\ &\leq \mathbb{E}_X[\delta^{**}(R_{0-1}(\eta(x), sign(f(x))))] && \text{(By convexity of } \delta^{**}) \\ &\leq \mathbb{E}_X[\delta(R_{0-1}(\eta(x), sign(f(x))))] && (\because \delta^{**} \leq \delta) \end{aligned}$$

From (7) and (8), $\delta(\epsilon) \leq R_\psi(\eta, \alpha)$, for all $\epsilon \leq R_{0-1}(\eta, \text{sign}(\alpha))$. In particular $\delta(R_{0-1}(\eta, sign(\alpha))) \leq R_\psi(\eta, \alpha)$. Hence,

$$\mathbb{E}_X[\delta(R_{0-1}(\eta(x), sign(f(x))))] \leq \mathbb{E}_X[R_\psi(\eta(x), f(x))] = \text{regret}_D^{\psi}[f]$$

Hence, $\delta^{**}(\mathbb{E}_X[R_{0,1}(\eta(x), sign(f(x)))]) \leq \text{regret}_D^{\psi}[f]$. □

Lemma 3.3. If ψ is l_{0-1} calibrated and $\psi_y(.)$ is continuous $\forall y \in \{\pm 1\}$, then $\delta(\epsilon) > 0$ for all $\epsilon > 0$.

Now, we are in a position to prove Theorem 3.1.

Proof of Theorem 3.1 $1 \Rightarrow 3$: From Lemma 3.3, $\delta(\epsilon) > 0$ for all $\epsilon > 0$. Hence, all points in the epigraph of δ in the interval $(0, 1]$ must lie above the X-axis. Since, epigraph of δ^{**} is the set of all convex combination of points in the epigraph of δ, the same must hold true for epigraph of δ^{**}. Hence, δ^{**} must be greater than 0 in the range $(0, 1]$.

Furthermore since, $\delta(0) = 0$, $\delta^{**}(0) = 0$. Hence, the function is strictly increasing at 0. Thus, it has a positive slope at $x = 0$. Furthermore, since δ^{**} is convex, its slope cannot decrease. Hence, it must have positive slope in the entire range $[0, 1]$. Hence, the function is strictly increasing in $[0, 1]$. Hence, it is invertible. Combining this result with the result of Theorem 3.2, we get

$$\text{regret}_D^{0-1}[\text{sign} \circ f] \leq (\delta^{**})^{-1}(\text{regret}_D^{\psi}[f])$$

4 Margin Based Losses

Recall from previous lectures that a loss function $\psi(y, \alpha)$ that can be written in the form of $\phi(y\alpha)$ for some $\phi : \mathbb{R} \to \mathbb{R}_+$, is termed as margin based loss.

Example 3. Margin based losses:

- **Logistic loss:** $\psi_{log}(y, \alpha) = \ln(1 + \exp(-y\alpha))$
- **Exponential loss:** $\psi_{exp}(y, \alpha) = \exp(-y\alpha)$
- **Squared loss:** $\psi_{sq}(y, \alpha) = (1 - y\alpha)^2$
- **Hinge loss:** $\psi_{hinge}(y\alpha) = (1 - y\alpha)_+$

Theorem 4.1 (Bartlett et al, 2006 [1])**.** For any margin based loss $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$, uniform calibration function can be written as

$$\delta(\epsilon) = \inf_{\alpha \in \mathbb{R}: \alpha \leq 0} R_\psi\left(\frac{1+\epsilon}{2}, \alpha\right)$$

Note that, in this case, classification calibration implies $\delta(\epsilon) > 0 \ \forall \epsilon > 0$, and hence, for margin based losses, classification calibration always implies a surrogate regret bound for $0-1$ loss.

If $\psi(y, \alpha) = \phi(y\alpha)$, with ϕ convex, then:

1. ψ is l_{0-1} calibrated $\iff$ ϕ is differentiable and $\phi'(0) < 0$
2. ψ is l_{0-1} calibrated $\Rightarrow \delta^{**}(\epsilon) = \phi(0) - H_\psi(\frac{1+\epsilon}{2})$, where H_ψ is the conditional Bayes $\psi-$risk.

Example 4 (Squared loss)**.** $\phi(u) = (1-u)^2$. The conditional Bayes $\psi-$risk is given by $H_{sq}(\eta) = 4\eta(1-\eta)$. From Theorem 4.1, $\delta^{**}(\epsilon) = 1 - (1-\epsilon)(1+\epsilon) = \epsilon^2$. By applying Theorem 3.2, we get

$$\text{regret}_D^{0-1}[\text{sign} \circ f] \leq \sqrt{\text{regret}_D^{sq}[f]}$$

Since, squared loss is also a $\lambda-$proper composite loss with $\lambda = 8$, we get from last lecture's results:

$$\begin{aligned}\text{regret}_D^{0-1}[\text{sign} \circ f] &\leq 2\sqrt{\frac{2}{\lambda}\text{regret}_D^{sq}[f]} \\ &= \sqrt{\text{regret}_D^{sq}[f]}\end{aligned}$$

In this case, both bounds turn out to be the same.

Example 5 (Exponential loss)**.** $\phi(u) = \exp(-u)$. The conditional Bayes $\psi-$risk is given by $H_{exp}(\eta) = 2\sqrt{\eta(1-\eta)}$. From Theorem 4.1, $\delta^{**}(\epsilon) = 1 - \sqrt{(1-\epsilon)(1+\epsilon)}$. Hence, $(\delta^{**})^{-1}(\bar{\epsilon}) = \sqrt{2\bar{\epsilon} - \bar{\epsilon}^2}$. By applying Theorem 3.2, we get from last lecture's results:

$$\text{regret}_D^{0-1}[\text{sign} \circ f] \leq \sqrt{2\,\text{regret}_D^{\exp}[f] - (\text{regret}_D^{\exp}[f])^2}$$

Since, exponential loss is also a $\lambda-$proper composite loss with $\lambda = 4$, we get

$$\text{regret}_D^{0-1}[\text{sign} \circ f] \leq \sqrt{2\,\text{regret}_D^{\exp}[f]}$$

In this case, we are able to obtain a tighter bound compared to the bound obtained on using the property that the loss is strongly proper composite.

Example 6 (Hinge loss)**.** $\phi(u) = (1-u)_+$. The conditional Bayes $\psi-$risk is given by $H_{hinge}(\eta) = 2\min(\eta, 1-\eta)$. From Theorem 4.1, $\delta^{**}(\epsilon) = 1 - 2\min\{\frac{1-\epsilon}{2}, \frac{1+\epsilon}{2}\} = \epsilon$. By applying Theorem 3.2, we get

$$\text{regret}_D^{0-1}[\text{sign} \circ f] \leq \text{regret}_D^{hinge}[f]$$

5 Next Lecture

In the last two lectures, we studied the consistency of surrogate risk minimization for binary classification problems. In next lecture, we will study the cosistency of surrogate risk minimization for multiclass learning problems.

References

[1] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[2] Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

[3] Tong Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–134, 2004.