E0 370 Statistical Learning Theory **Lecture 15 (Oct 1, 2013)**

Consistency of surrogate risk minimization methods for multiclass 0-1 classification

Lecturer: Shivani Agarwal *Scribe: Debarghya Ghoshdastidar*

1 Introduction

In the previous two lectures, we learned about consistency of surrogate risk minimization methods for binary classification. In this lecture, we will study how the previous results extend to a multiclass setting. As examples, we will study consistency of some multiclass SVM formulations.

We begin by describing how the surrogate risk minimization framework changes for multiclass classification. This is shown in Table 1. For the purpose of this lecture, we will restrict ourselves to the multiclass 0-1 loss.

	Binary classification	Multiclass classification
Label space, $\mathcal{Y}$ and Prediction space, $\mathcal{T}$	$\{\pm 1\}$	$[n]$
Target 0-1 loss $\boldsymbol{\ell}_{0\text{-}1} : \mathcal{Y} \times \mathcal{T} \mapsto \mathbb{R}_+$	$\boldsymbol{\ell}_{0\text{-}1} : \{\pm 1\} \times \{\pm 1\} \mapsto \mathbb{R}_+$ $\boldsymbol{\ell}_{0\text{-}1}(y,t) = \mathbf{1}(t \neq y)$	$\boldsymbol{\ell}_{0\text{-}1} : [n] \times [n] \mapsto \mathbb{R}_+$ $\boldsymbol{\ell}_{0\text{-}1}(y,t) = \mathbf{1}(t \neq y)$
Surrogate loss $\boldsymbol{\psi} : \mathcal{Y} \times \mathcal{C} \mapsto \mathbb{R}_+$	$\boldsymbol{\psi} : \{\pm 1\} \times \mathcal{C} \mapsto \mathbb{R}_+$ where $\mathcal{C} \subseteq \mathbb{R}$	$\boldsymbol{\psi} : [n] \times \mathcal{C} \mapsto \mathbb{R}_+$ where $\mathcal{C} \subseteq \mathbb{R}^n$
'pred' function pred : $\mathcal{C} \mapsto \mathcal{T}$	pred : $\mathcal{C} \mapsto \{\pm 1\}$ pred$(\alpha) = \text{sign}(\alpha)$	pred : $\mathcal{C} \mapsto [n]$ pred$(\mathbf{u}) = \arg\max_{y \in [n]} u_y$
Conditional $\boldsymbol{\psi}$-risk	$L_{\boldsymbol{\psi}} : [0,1] \times \mathcal{C} \mapsto \mathbb{R}_+$ $L_{\boldsymbol{\psi}}(\eta, \alpha) = \eta \boldsymbol{\psi}(1,\alpha) + (1-\eta)\boldsymbol{\psi}(-1,\alpha)$	$L_{\boldsymbol{\psi}} : \Delta_n \times \mathcal{C} \mapsto \mathbb{R}_+$ $L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) = \sum_{y=1}^{n} p_y \boldsymbol{\psi}(y, \mathbf{u})$
Conditional Bayes $\boldsymbol{\psi}$-risk	$H_{\boldsymbol{\psi}} : [0,1] \mapsto \mathbb{R}_+$ $H_{\boldsymbol{\psi}}(\eta) = \inf_{\alpha \in \mathcal{C}} L_{\boldsymbol{\psi}}(\eta, \alpha)$	$H_{\boldsymbol{\psi}} : \Delta_n \mapsto \mathbb{R}_+$ $H_{\boldsymbol{\psi}}(\mathbf{p}) = \inf_{\mathbf{u} \in \mathcal{C}} L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u})$
Conditional $\boldsymbol{\psi}$-regret	$R_{\boldsymbol{\psi}} : [0,1] \times \mathcal{C} \mapsto \mathbb{R}_+$ $R_{\boldsymbol{\psi}}(\eta, \alpha) = L_{\boldsymbol{\psi}}(\eta, \alpha) - H_{\boldsymbol{\psi}}(\eta)$	$R_{\boldsymbol{\psi}} : \Delta_n \times \mathcal{C} \mapsto \mathbb{R}_+$ $R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) = L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) - H_{\boldsymbol{\psi}}(\mathbf{p})$

Table 1: Surrogate risk minimization framework for binary and multiclass classification. We denote the y^{th} coordinate of vectors $\mathbf{u}$ and $\mathbf{p}$ by u_y and p_y, respectively. Here, $p_y = \mathbf{P}(x \in \text{Class } y | x)$.

Example 1 (Multiclass logistic regression)**.** The loss function in multiclass logistic regression is given by the surrogate loss $\boldsymbol{\psi}_{\log} : [n] \times \mathbb{R}^n \mapsto \mathbb{R}_+$ defined as

$$\boldsymbol{\psi}_{\log}(y, \mathbf{u}) = -\ln \left(\frac{e^{-u_y}}{\sum_{y'=1}^{n} e^{-u_{y'}}} \right) \qquad \forall y \in [n] \quad \forall \mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n.$$

One can verify that minimizing $\boldsymbol{\psi}_{\log}$ is equivalent to maximizing the softmax function, which can be used as parametric form for conditional class probability. It is also known that $\boldsymbol{\psi}_{\log}$ is a multiclass proper composite loss [4].

2 Classification calibration in multiclass setting

We will now see how the definition of $\boldsymbol{\ell}_{0\text{-}1}$-calibrated loss function extends to the multiclass framework. We can also state a result on multiclass $\boldsymbol{\ell}_{0\text{-}1}$-calibrated losses similar to the one in previous lecture. This is given in Theorem 2.1.

Definition. For $\mathcal{C} \subseteq \mathbb{R}^n$, a surrogate loss function $\boldsymbol{\psi} : [n] \times \mathcal{C} \mapsto \mathbb{R}^n$ is said to be $\boldsymbol{\ell}_{0\text{-}1}$-calibrated if

$$\forall \mathbf{p} \in \Delta^n, \qquad \inf_{\mathbf{u} \in \mathcal{C} : \text{pred}(\mathbf{u}) \notin \arg\min_{t \in [n]} L_{0\text{-}1}(\mathbf{p}, t)} L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) > \inf_{\mathbf{u} \in \mathcal{C}} L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}),$$

or equivalently,

$$\forall \mathbf{p} \in \Delta^n, \qquad \inf_{\mathbf{u} \in \mathcal{C} : R_{0\text{-}1}(\mathbf{p}, \text{pred}(\mathbf{u})) > 0} R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) > 0.$$

Theorem 2.1. Let $\mathcal{C} \subseteq \mathbb{R}^n$. Let $\boldsymbol{\psi} : [n] \times \mathcal{C} \mapsto \mathbb{R}_+$ be such that $\boldsymbol{\psi}_y(\mathbf{u}) := \boldsymbol{\psi}(y, \mathbf{u})$ is continuous $\forall y \in [n]$. Then the following are equivalent:

1. $\boldsymbol{\psi}$ is $\boldsymbol{\ell}_{0\text{-}1}$-calibrated.
2. for all distributions D on $\mathcal{X} \times [n]$,

$$\text{regret}_D^{\boldsymbol{\psi}}[f_S] \xrightarrow{p} 0 \quad \Longrightarrow \quad \text{regret}_D^{0\text{-}1}[\text{pred} \circ f_S] \xrightarrow{p} 0.$$

3. $\exists g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that g is continuous at 0, $g(0) = 0$ and $\forall D$ on $\mathcal{X} \times [n]$, $\forall f : \mathcal{X} \mapsto \mathcal{C}$,

$$\text{regret}_D^{0\text{-}1}[\text{pred} \circ f] \leqslant g\left(\text{regret}_D^{\boldsymbol{\psi}}[f]\right).$$

Proof. As in the previous lecture, we will skip the proof for (2) $\Rightarrow$ (1), whereas (3) $\Rightarrow$ (2) is obvious. For the case of (1) $\Rightarrow$ (3), we note that the proof is similar to the corresponding proof in previous lecture. The key step in that proof was a result on positivity of the uniform calibration function. We will state and prove a generalized version of that lemma, which can be directly applied to prove that (1) $\Rightarrow$ (3) as in previous lecture. □

We restate the definition of calibration function in the multiclass setting.

Definition. The calibration function at $\mathbf{p} \in \Delta_n$ is the function $\delta_{\mathbf{p}} : [0, 1] \mapsto \mathbb{R}_+$ defined as

$$\delta_{\mathbf{p}}(\epsilon) = \inf_{\mathbf{u} \in \mathcal{C} : R_{0\text{-}1}(\mathbf{p}, \text{pred}(\mathbf{u})) \geqslant \epsilon} R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) \ .$$

The uniform calibration function $\delta : [0, 1] \mapsto \mathbb{R}_+$ is given by $\delta(\epsilon) = \inf_{\mathbf{p} \in \Delta_n} \delta_{\mathbf{p}}(\epsilon)$.

Note that $\delta_{\mathbf{p}}(0) = 0 \ \forall \mathbf{p} \in \Delta_n$, and hence, $\delta(0) = 0$. Also, from the definition of classification calibration, we have that

$$\boldsymbol{\psi} \text{ is } \boldsymbol{\ell}_{0\text{-}1}\text{-calibrated} \quad \Longleftrightarrow \quad \delta_{\mathbf{p}}(\epsilon) > 0 \quad \forall \epsilon > 0, \forall \mathbf{p} \in \Delta_n.$$

Hence, we can say that $\delta(\epsilon) \geqslant 0 \quad \forall \epsilon > 0$. We now state the result by Zhang [6], which guarantees positivity of $\delta(\epsilon) \ \forall \epsilon > 0$. Before proving this result, recall the following property of conditional Bayes risk.

Exercise. The function $H_{\boldsymbol{\psi}}(\mathbf{p})$ is a continuous function of $\mathbf{p}$ on Δ_n.

Lemma 2.2. Let $\boldsymbol{\psi} : [n] \times \mathcal{C} \mapsto \mathbb{R}_+$ be such that $\boldsymbol{\psi}_y(\mathbf{u})$ is continuous $\forall y \in [n]$. If $\boldsymbol{\psi}$ is $\boldsymbol{\ell}_{0\text{-}1}$-calibrated, then $\delta(\epsilon) > 0 \ \forall \epsilon > 0$.

Proof. Let $\boldsymbol{\psi}$ be $\boldsymbol{\ell}_{0\text{-}1}$-calibrated. We will prove the claim by contradiction.
Let, if possible, $\exists \epsilon > 0$ such that $\delta(\epsilon) = 0$, *i.e.*,

$$\inf_{\mathbf{p} \in \Delta_n} \inf_{\mathbf{u} \in \mathcal{C} : R_{0\text{-}1}(\mathbf{p}, \text{pred}(\mathbf{u})) \geqslant \epsilon} R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) = 0.$$

Note that $R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u})$ is continuous both in $\mathbf{p}$ and $\mathbf{u}$. Hence, there exists a sequence $\left(\mathbf{p}^{(m)}, \mathbf{u}^{(m)}\right)_{m\in\mathbb{N}}$ in $\Delta_n \times \mathcal{C}$ such that

$$R_{0\text{-}1}(\mathbf{p}^{(m)}, \text{pred}(\mathbf{u}^{(m)})) \geqslant \epsilon \quad \forall m \tag{1}$$

$$\text{and } \lim_{m\to\infty} R_{\boldsymbol{\psi}}(\mathbf{p}^{(m)}, \mathbf{u}^{(m)}) = 0, \textit{ i.e., } \lim_{m\to\infty} \left(L_{\boldsymbol{\psi}}(\mathbf{p}^{(m)}, \mathbf{u}^{(m)}) - H_{\boldsymbol{\psi}}(\mathbf{p}^{(m)})\right) = 0. \tag{2}$$

Since Δ_n is compact, $\exists$ a convergent subsequence of $\left(\mathbf{p}^{(m)}\right)$. Let $\mathbf{p}$ be the limit of that subsequence. We will further work only with this subsequence, and so, for convenience, we denote the subsequence also by the same index m, *i.e.*, $\mathbf{p}^{(m)} \to \mathbf{p} \in \Delta_n$.
By continuity of $H_{\boldsymbol{\psi}}$, we have $H_{\boldsymbol{\psi}}(\mathbf{p}^{(m)}) \to H_{\boldsymbol{\psi}}(\mathbf{p})$ and so,

$$\lim_{m\to\infty} L_{\boldsymbol{\psi}}(\mathbf{p}^{(m)}, \mathbf{u}^{(m)}) = H_{\boldsymbol{\psi}}(\mathbf{p}). \tag{3}$$

Assume without loss of generality that $\mathbf{p} = (p_1, \ldots, p_n)$ is such that $p_y = 0$ for $y = 1, \ldots, k$ and $p_y > 0$ for $y = k+1, \ldots, n$. We have

$$\begin{aligned}
\limsup_{m\to\infty} L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}^{(m)}) &= \limsup_{m\to\infty} \sum_{y=k+1}^{n} p_y \boldsymbol{\psi}_y(\mathbf{u}^{(m)}) \\
&= \limsup_{m\to\infty} \sum_{y=k+1}^{n} p_y^{(m)} \boldsymbol{\psi}_y(\mathbf{u}^{(m)}) && \text{(since } \mathbf{p}^{(m)} \to \mathbf{p}\text{)} \\
&\leqslant \limsup_{m\to\infty} \sum_{y=1}^{n} p_y^{(m)} \boldsymbol{\psi}_y(\mathbf{u}^{(m)}) \\
&= \lim_{m\to\infty} L_{\boldsymbol{\psi}}(\mathbf{p}^{(m)}, \mathbf{u}^{(m)}) = H_{\boldsymbol{\psi}}(\mathbf{p}) && \text{(from Eq. (3)).}
\end{aligned}$$

On the other hand, from definition of $H_{\boldsymbol{\psi}}(\mathbf{p})$, it follows that

$$\liminf_{m\to\infty} L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}^{(m)}) \geqslant H_{\boldsymbol{\psi}}(\mathbf{p}).$$

Hence, $\lim_{m\to\infty} L_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}^{(m)}) = H_{\boldsymbol{\psi}}(\mathbf{p})$, *i.e.*, $\lim_{m\to\infty} R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}^{(m)}) = 0$.
But from Eq. (1) and the fact $\mathbf{p}^{(m)} \to \mathbf{p}$, we can derive that

$$R_{0\text{-}1}(\mathbf{p}, \text{pred}(\mathbf{u}^{(m)})) \geqslant \epsilon \quad \forall m.$$

Thus, we have $\delta_{\mathbf{p}}(\epsilon) = \inf_{\mathbf{u}\in\mathcal{C}: R_{0\text{-}1}(\mathbf{p}, \text{pred}(\mathbf{u})) \geqslant \epsilon} R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) = 0$, which contradicts the fact that $\epsilon > 0$.
Hence, the claim is proved. □

In the above results (Theorem 2.1 and Lemma 2.2), the assumption on continuity of $\boldsymbol{\psi}_y$ will not be required if the loss is a margin based binary loss. This follows from the exact expression that can be obtained for margin based losses (as discussed in the previous lecture).

Exercise. We have proved Theorem 2.1 and Lemma 2.2 based on the assumption that $\mathcal{Y} = \mathcal{T} = [n]$ is finite. Identify which of the statements in the proofs make use of this assumption. Can we modify the proofs such that the results hold for $\mathcal{Y} = \mathcal{T} = [0, 1]$?

3 Consistency of some multiclass SVM formulations

Theorem 2.1 can be used to study consistency of different surrogate risk minimization methods for multiclass 0-1 classification. In particular, we may study consistency of various multiclass formulations for SVM that have been proposed in the literature [5, 1, 2]. This has been studied in [6]. In all the multiclass SVM formulations, a multiclass extension of the hinge loss is minimized. However, one may consider any margin function ϕ in general as used below. In particular, $\phi(z) = (1 - z)_+$ would correspond to the algorithms in [5, 1, 2].

3.1 Formulation by Weston & Watkins [5]

$\mathcal{C} = \mathbb{R}^n$ and surrogate loss $\boldsymbol{\psi}_{WW} : [n] \times \mathbb{R}^n \mapsto \mathbb{R}_+$ is

$$\boldsymbol{\psi}_{WW}(y, \mathbf{u}) = \sum_{y' \neq y} \phi(u_y - u_{y'}) \qquad \text{for some } \phi : \mathbb{R} \mapsto \mathbb{R}_+$$

Theorem 3.1. Let ϕ be a decreasing function that is differentiable everywhere and $\phi'(0) < 0$. Then $\boldsymbol{\psi}_{WW}$ is $\boldsymbol{\ell}_{0\text{-}1}$-calibrated.

Though the above includes exponential and logistic losses, it does not include the hinge loss. In fact Zhang [6] showed a counter-example in case of hinge loss. Thus $\boldsymbol{\psi}_{WW}$ with the hinge loss is not $\boldsymbol{\ell}_{0\text{-}1}$-calibrated.

3.2 Formulation by Crammer & Singer [1]

$\mathcal{C} = \mathbb{R}^n$ and surrogate loss $\boldsymbol{\psi}_{CS} : [n] \times \mathbb{R}^n \mapsto \mathbb{R}_+$ is

$$\boldsymbol{\psi}_{CS}(y, \mathbf{u}) = \phi\left(u_y - \max_{y' \neq y} u_{y'}\right) \qquad \text{for some } \phi : \mathbb{R} \mapsto \mathbb{R}_+$$

This formulation is popularly used in structured prediction. However, for convex ϕ, in general $\boldsymbol{\psi}_{CS}$ is not $\boldsymbol{\ell}_{0\text{-}1}$-calibrated.

3.3 Formulation by Lee, Lin & Wahba [2]

$\mathcal{C} = \left\{\mathbf{u} \in \mathbb{R}^n : \sum_{y=1}^n u_y = 0\right\}$ and surrogate loss $\boldsymbol{\psi}_{LLW} : [n] \times \mathcal{C} \mapsto \mathbb{R}_+$ is

$$\boldsymbol{\psi}_{LLW}(y, \mathbf{u}) = \sum_{y' \neq y} \phi(-u_{y'}) \qquad \text{for some } \phi : \mathbb{R} \mapsto \mathbb{R}_+$$

Theorem 3.2. Let ϕ be a convex function that is differentiable on $(-\infty, 0]$ and $\phi'(0) < 0$. Then $\boldsymbol{\psi}_{LLW}$ is $\boldsymbol{\ell}_{0\text{-}1}$-calibrated.

This result includes the hinge loss, and therefore this yields a universally consistent formulation of multiclass SVMs.

3.4 Formulation using one vs. all approach

$\mathcal{C} = \mathbb{R}^n$ and surrogate loss $\boldsymbol{\psi}_{OvA} : [n] \times \mathbb{R}^n \mapsto \mathbb{R}_+$ is

$$\boldsymbol{\psi}_{OvA}(y, \mathbf{u}) = \phi(u_y) + \sum_{y' \neq y} \phi(-u_{y'}) \qquad \text{for some } \phi : \mathbb{R} \mapsto \mathbb{R}_+$$

Exercise. Show that one vs. all approach effectively minimizes the above loss on the training sample.

Theorem 3.3. Let ϕ be a convex function that is differentiable everywhere and $\phi(z) < \phi(-z)\ \forall z > 0$. Then $\boldsymbol{\psi}_{OvA}$ is $\boldsymbol{\ell}_{0\text{-}1}$-calibrated.

This result also does not hold for the hinge loss.

4 Consistency over a class of distributions

The above discussion shows that most of the multiclass SVM formulations (in Sections 3.1,3.2 and 3.4) are not consistent. However, in practice, they are found to perform quite well. This is due to the fact that these losses are $\boldsymbol{\ell}_{0\text{-}1}$-calibrated under certain conditions on the underlying distributions. This can formalized through the following definition.

Definition. For $\mathcal{P} \subseteq \Delta_n$ and $\mathcal{C} \subseteq \mathbb{R}^n$, a surrogate loss function $\boldsymbol{\psi} : [n] \times \mathcal{C} \mapsto \mathbb{R}^n$ is said to be $\boldsymbol{\ell}_{0\text{-}1}$-calibrated over $\mathcal{P}$ if

$$\forall \mathbf{p} \in \mathcal{P}, \qquad \inf_{\mathbf{u} \in \mathcal{C}: R_{0\text{-}1}(\mathbf{p}, \text{pred}(\mathbf{u})) > 0} R_{\boldsymbol{\psi}}(\mathbf{p}, \mathbf{u}) > 0.$$

Based on this definition, we can redefine the calibration and uniform calibration functions. The statement of Theorem 2.1 holds for all distributions D on $\mathcal{X} \times [n]$ such that $D_{y|x} \in \mathcal{P}$, *i.e.*, we do not get universal consistency, but consistency on $\mathcal{P}$. In Sections 3.1,3.2 and 3.4, the formulation with hinge loss is consistent over $\mathcal{P} = \left\{ \mathbf{p} \in \Delta_n : \max_{y \in [n]} p_y > \frac{1}{2} \right\}$, *i.e.*, when one of the classes strictly dominates over others.

5 Additional pointers

Most of the results discussed in this lecture can be found in [6]. An alternative approach was studied in [3] using geometric interpretations. The authors of this paper also study if consistency is possible when the 'pred' function is not of 'argmax' type. One of their observations is that if $\boldsymbol{\psi} : [n] \times \mathcal{C} \mapsto \mathbb{R}_+$ is a symmetric surrogate loss, then it suffices to show only the 0-1 calibration for argmax pred function. Here, the term 'symmetric' loss means that for all $\mathbf{u} \in C$ and permutations $\sigma : [n] \to [n]$, $\exists \mathbf{u}' \in C$ such that $\psi(\sigma(y), \mathbf{u}) = \psi(y, \mathbf{u}') \forall y \in [n]$.

We had seen in last class that better bounds are achieved using classification calibration than using proper losses. This apparently seems counter-intuitive as stronger assumptions are used in the latter case. However, in case of proper losses, we tried to estimate distribution (given by η) which implied that we tried to learn more than just the classes of data. Hence, the bounds were looser than the case of classification calibration, where we try to achieve consistency only with respect to problem of learning the classes. In fact, it is known that the classification calibration based bounds are the tightest in the distribution-free setting.

References

[1] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[2] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[3] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[4] E. Vernet, R. C. Williamson, and M. Reid. Composite multiclass losses. In *Advances in Neural Information Processing Systems*, 2011.

[5] J. Weston and C. Watkins. Multi-class support vector machines. Technical report, CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, 1998.

[6] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.