A Convex Surrogate Operator for General Non-Modular Loss Functions

Jiaqian B Yu, Matthew Blaschko

A Convex Surrogate Operator for General Non-Modular Loss Functions

Jiaqian Yu JIAQIAN.YU@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay & Inria Saclay, 92295 Châtenay-Malabry, France

Matthew B. Blaschko MATTHEW.BLASCHKO@ESAT.KULEUVEN.BE
Center for Processing Speech and Images, Departement Elektrotechniek, KU Leuven, 3001 Leuven, Belgium

Keywords: convex surrogate function, non-modular loss, submodularity, Sørensen-Dice loss

Abstract

In this work, a novel generic convex surrogate for general non-modular loss functions is introduced, which provides for the first time a tractable solution for loss functions that are neither supermodular nor submodular. This convex surrogate is based on a submodular-supermodular decomposition. It takes the sum of two convex surrogates that separately bound the supermodular component and the submodular component using slack-rescaling and the Lovász hinge, respectively. This surrogate is convex, piecewise linear, an extension of the loss function, and for which subgradient computation is polynomial time. Empirical results are reported on a non-submodular loss based on the Sørensen-Dice difference function demonstrating the improved performance, efficiency, and scalability of the novel convex surrogate. This is a shorten version of a recent published paper (Yu & Blaschko, 2016).

1. Introduction

Following the risk minimization principle (Vapnik, 1995), we may wish to minimize a loss function that more closely reflects the cost of a specific set of predictions. Alternatives to the Hamming loss are frequently employed in the discriminative learning literature(Cheng et al., 2010; Petterson & Caetano, 2011; Doppa et al., 2014; Blaschko & Lampert, 2008; Everingham et al., 2010; Nowozin, 2014; Narasimhan & Bilmes, 2005). Existing polynomial-time convex surrogates exist for supermodular (Tsochantaridis et al., 2005) or submodular losses (Yu & Blaschko, 2015), but not for more general non-modular losses. This has motivated us to study the conditions for a loss function to be tractably upper bounded with a tight convex surrogate. We propose a novel convex surrogate for general non-modular loss functions, which is solvable for the first time for non-supermodular and non-submodular loss functions.

Appearing in *Proceedings of Benelearn 2016*.

2. Method

Any loss function Δ of may be interpreted as a set function where inclusion in a set is defined by a corresponding prediction being incorrect:

$$\Delta(y, \tilde{y}) = l(\{i | y^i \neq \tilde{y}^i\}) \tag{1}$$

for some set function l.

Definition 1 (Submodular function)**.** *A set function $l : \mathcal{P}(V) \mapsto \mathbb{R}$ is submodular iff for all $B \subseteq A \subset V$ and $x \in V \setminus A$, $l(B \cup \{x\}) - l(B) \geq l(A \cup \{x\}) - l(A)$.*

A function is *supermodular* iff its negative is submodular, and a function is modular (e.g. Hamming loss) iff it is both submodular and supermodular.

For all set functions l, there always exists a decomposition into the sum of a submodular function $f \in \mathcal{S}$ and a supermodular function $g \in \mathcal{G}$: $l = f + g$. Moreover, we can always make the supermodular component to be increasing: for an arbitrary decomposition $l = f + g$ where g is not increasing, there exists a modular function m_g s.t. $l = (f - m_g) + (g + m_g)$, with $\tilde{f} := f - m_g \in \mathcal{S}$, and $\tilde{g} := g + m_g \in \mathcal{G}_+$ is increasing.

However, we subsequently demonstrate that decompositions can vary by more than a modular factor:

Proposition 1 (Non-uniqueness of decomposition up to modular transformations.)**.** *For any set function, there exist multiple decompositions into submodular and supermodular components such that these components differ by more than a modular factor:*

$$\exists f_1, f_2 \in \mathcal{S},\ g_1, g_2 \in \mathcal{G}$$
$$(l = f_1 + g_1 = f_2 + g_2) \wedge (g_1 + m_{g_1} \neq g_2 + m_{g_2}) \quad (2)$$

where $\wedge$ denotes "logical and," m_{g_1} and m_{g_2} are modular.

We define $\mathbf{D}$ such that these two sources of non-uniqueness are resolved using a canonical decomposition $l = f^* + g^*$.

Definition 2. *We define an operator $\mathbf{D} : \mathcal{F} \mapsto \mathcal{G}_+$ as*

$$\mathbf{D}l = \arg\min_{g \in \mathcal{G}_+} \sum_{A \subseteq V} g(A), \quad s.t.\ l - g \in \mathcal{S}. \quad (3)$$

We note that minimizing the values of g will simultaneously remove the non-uniqueness due both to the modular non-uniqueness described above, as well as the non-modular non-uniqueness described in Proposition 1. We formally prove this in Proposition 2.

Proposition 2. $\mathbf{D}l$ *is unique for all $l \in \mathcal{F}$ that have a finite base set V.*

We finally denote the resulting decomposition of $\Delta(y, \cdot) = \Delta_{\mathcal{G}}(y, \cdot) + \Delta_{\mathcal{S}}(y, \cdot)$ into its supermodular and submodular components, respectively.

We construct a surrogate $\mathbf{B}$ by taking the sum of two convex surrogates. These surrogates are slack-rescaling (Tsochantaridis et al., 2005) applied to $\Delta_{\mathcal{G}}$:

$$\mathbf{S}\Delta(y, h(x)) := \max_{\tilde{y} \in \mathcal{Y}} \Delta(y, \tilde{y}) \left(1 + \langle h(x), \tilde{y} \rangle - \langle h(x), y \rangle\right), \quad (4)$$

and the Lovász hinge (Yu & Blaschko, 2015) applied to $\Delta_{\mathcal{S}}$:

$$\mathbf{L}\Delta(y, h(x)) :=$$
$$\left(\max_{\pi} \sum_{j=1}^{p} s^{\pi_j} \left(l\left(\{\pi_1, \cdots, \pi_j\}\right) - l\left(\{\pi_1, \cdots, \pi_{j-1}\}\right)\right)\right)_+ \quad (5)$$

where $(\cdot)_+ = \max(\cdot, 0)$, π is a permutation,$s^{\pi_j} = 1 - h^{\pi_j}(x) y^{\pi_j}$, and $h^{\pi_j}(x)$ is the π_jth dimension of $h(x)$.

Definition 3 (General non-modular convex surrogate)**.** *For an arbitrary non-negative loss function Δ, we define*

$$\mathbf{B_D}\Delta := \mathbf{L}\Delta_{\mathcal{S}} + \mathbf{S}\Delta_{\mathcal{G}} \quad (6)$$

where $\mathbf{D}$ is the decomposition of l defined by Def. 2.

We can demonstrate that:

1. If l is supermodular, then f^* is modular.
2. If f^* is increasing, $\mathbf{S}\Delta_{\mathcal{G}} = \mathbf{S}(\Delta - \Delta_{\mathcal{S}}) = \mathbf{S}\Delta - \mathbf{S}\Delta_{\mathcal{S}}$.
3. If $l \in \mathcal{G}_+$, then $\mathbf{B_D}\Delta \geq \mathbf{S}\Delta$ over the unit cube and therefore $\mathbf{B_D}$ is closer to the convex closure of Δ than $\mathbf{S}$.
4. If $l \in \mathcal{S}$, then $\mathbf{B_D}\Delta = \mathbf{L}\Delta$.
5. $\mathbf{B_D}\Delta$ is convex for arbitrary Δ.
6. $\mathbf{B_D}\Delta$ is an extensionof Δ iff $\Delta_{\mathcal{S}}$ is non-negative.
7. The subgradient computation of $\mathbf{B_D}\Delta$ is polynomial time given polynomial time oracle access to f^* and g^*.

3. Sørensen-Dice loss

We introduce the Sørensen-Dice loss based on the Sørensen-Dice coefficient (Dice, 1945; Sørensen, 1948):

$$\Delta_D(y, \tilde{y}) = 1 - \frac{2|y \cap \tilde{y}|}{|y| + |\tilde{y}|}. \quad (7)$$

given a groundtruth y and a predicted output $\tilde{y}$ and $y \subseteq V$ is a set of positive labels, e.g. foreground pixels.

Proposition 3. $\Delta_D(y, \tilde{y})$ *is neither submodular nor supermoduler under the isomorphism $(y^*, \tilde{y}) \rightarrow A := \{i | y_i^* \neq \tilde{y}_i\}$, $\Delta_J(y^*, \tilde{y}) \cong l(A)$.*

We test the proposed surrogate on a binary set prediction problem. Two classes of 2-dimensional data are generated by different Gaussian mixtures. We use the $\mathbf{B_D}$ during training and compare it to slack rescaling $\mathbf{S}$ with an approximate optimization procedure based on greedy maximization. We additionally train an SVM (denoted 0-1 in the results table) for comparison. During test time, we evaluate with Δ_D and with Hamming loss to calculate the empirical error values as shown in Table 1.

	Test	
$p = 6$	Δ_D	0-1
$\mathbf{B_D}$	$\mathbf{0.1121 \pm 0.0040}$	0.6027 ± 0.0125
0-1	0.1497 ± 0.0046	0.5370 ± 0.0114
$\mathbf{S}$	0.3183 ± 0.0148	0.7313 ± 0.0209

Table 1. The cross comparison of average loss values (with standard error), Δ_D is the Dice loss as in Eq. (7).

We can see from the result that training Δ_D with $\mathbf{B_D}$ yields the best result while using Δ_D during test time. $\mathbf{B_D}$ performs better than $\mathbf{S}$ in both cases due to the failure of the approximate maximization procedure necessary to maintain computational feasibility.

Acknowledgments

This work is partially funded by Internal Funds KU Leuven, ERC Grant 259112, and FP7-MC-CIG 334380. The first author is supported by a fellowship from the China Scholarship Council.

References

Blaschko, M. B., & Lampert, C. H. (2008). Learning to localize objects with structured output regression. In D. Forsyth, P. Torr and A. Zisserman (Eds.), *European conference on computer vision*, vol. 5302 of *Lecture Notes in Computer Science*, 2–15. Springer.

Cheng, W., Hüllermeier, E., & Dembczynski, K. J. (2010). Bayes optimal multilabel classification via probabilistic classifier chains. *Proceedings of the International Conference on Machine Learning* (pp. 279–286).

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297–302.

Doppa, J. R., Yu, J., Ma, C., Fern, A., & Tadepalli, P. (2014). HC-search for multi-label prediction: An empirical study. *Proceedings of AAAI Conference on Artificial Intelligence*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*, 303–338.

Narasimhan, M., & Bilmes, J. (2005). A submodular-supermodular procedure with applications to discriminative structure learning. *Uncertainty in Artificial Intelligence (UAI)*.

Nowozin, S. (2014). Optimal decisions from probabilistic models: The intersection-over-union case. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Petterson, J., & Caetano, T. S. (2011). Submodular multi-label learning. *Advances in Neural Information Processing Systems* (pp. 1512–1520).

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, *5*, 1–34.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, *6*, 1453–1484.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.

Yu, J., & Blaschko, M. B. (2015). Learning submodular losses with the Lovász hinge. *Proceedings of the 32nd International Conference on Machine Learning* (pp. 1623–1631). Lille, France.

Yu, J., & Blaschko, M. B. (2016). A convex surrogate operator for general non-modular loss functions. *International Conference on Artificial Intelligence and Statistics* (pp. 1032–1041).