Convex Calibration Dimension for General Loss Matrices

Harish G. Ramaswamy HARISH_GURUP@CSA.IISC.ERNET.IN
Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Department of Computer Science and Automation, Indian Institute of Science, Bangalore, 560012.

Abstract

We study consistency properties of surrogate loss functions for general multiclass classification problems, defined by a general loss matrix. We extend the notion of classification calibration, which has been studied for certain specific learning problems, to the general multiclass setting. We then introduce the notion of *convex calibration dimension* of a multiclass loss matrix which is an intrinsic measure of difficulty of the learning problem defined by the loss matrix. We derive both upper and lower bounds on this quantity, use these bounds to analyze various loss matrices and derive interesting results concerning the difficulty of ranking.

1. Introduction

There has been significant interest and progress in recent years in understanding consistency of learning methods for various finite-output learning problems, such as binary classification, multiclass 0-1 classification, and various forms of ranking and multi-label prediction problems (Bartlett et al., 2006; Tewari & Bartlett, 2007; Duchi et al., 2010; Zhang, 2004). Such finite-output problems can all be viewed as instances of a general multiclass learning problem, whose structure is defined by a loss function, or equivalently, by a loss matrix. While the studies above have contributed to the understanding of learning problems corresponding to certain forms of loss matrices, a framework for analyzing consistency properties for a general multiclass learning problem, defined by a general loss matrix, has remained elusive.

In this paper, we analyze consistency of surrogate losses for general multiclass learning problems, building on past results (Tewari & Bartlett, 2007; Zhang, 2004). We start in Section 2 with some background and formalize the notion of calibration with respect to a general loss matrix. Section 3 introduces the notion of *convex calibration (CC) dimension* of a loss matrix, a fundamental quantity that measures the smallest 'size' of a prediction space for which it is possible to design a convex 'calibrated' surrogates. We derive both upper and lower bounds on this quantity, and use these results to analyze various loss matrices. As an application of these bounds, in Section 4, we show that the mean average precision and pairwise disagreement losses used in ranking have large CC-dimension and hence many known algorithms are inconsistent.

2. Preliminaries and Setup

We are given training examples $(X_1, Y_1), \ldots, (X_m, Y_m)$ drawn i.i.d. from a distribution D on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an instance space and $\mathcal{Y} = [n] = \{1, \ldots, n\}$ is a finite set of *class labels*. We are also given a finite set $\mathcal{T} = [k] = \{1, \ldots, k\}$ of *target/prediction labels* in which predictions are to be made, and a *loss function* $\ell : \mathcal{Y} \times \mathcal{T} \to [0, \infty)$, where $\ell(y, t)$ denotes the loss incurred on predicting $t \in \mathcal{T}$ when the label is $y \in \mathcal{Y}$. In many common learning problems, $\mathcal{T} = \mathcal{Y}$, but in general, these could be different (e.g. when there is an'abstain' option available to a classifier, in which case $k = n+1$). We denote by Δ_n, the set $\{\mathbf{p} \in \mathbb{R}^n_+ : \sum_y p_y = 1\}$.

We will find it convenient to view the loss function ℓ as a *loss matrix*. For each $y \in [n], t \in [k]$, we will denote $\ell(y, t)$ by ℓ_{yt} and define $\boldsymbol{\ell}_t$ as $\boldsymbol{\ell}_t = (\ell_{1t}, \ldots, \ell_{nt})^\top \in \mathbb{R}^n$.

The standard ERM algorithm requires us to find a function $h \in \mathcal{H} \subseteq \mathcal{T}^{\mathcal{X}}$ that minimizes

$$\mathrm{er}^\ell[h] = \sum_{i=1}^{m} \ell(Y_i, h(X_i)) \qquad (1)$$

but due to the discrete nature of the above problem one tries to minimize a 'surrogate' loss instead. Let $\widehat{\mathcal{T}} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ and $\psi : \mathcal{Y} \times \widehat{\mathcal{T}} \to \mathbb{R}$. The

Part of this work was presented at NIPS 2012 (Ramaswamy & Agarwal, 2012).

Appearing in Proceedings of the 1st Indian Workshop on Machine Learning, IIT Kanpur, India, 2013.

problem of interest now is to find a function $f \in \mathcal{F} \subseteq \widehat{\mathcal{T}}^{\mathcal{X}}$ that minimizes

$$\mathrm{er}^{\psi}[f] = \sum_{i=1}^{m} \psi(Y_i, f(X_i)) \tag{2}$$

The above problem is often easier because we can choose ψ to be convex in its second argument and $\mathcal{F}$ to be a convex set, in which case the above problem is a convex optimization problem. But in order to solve the original problem given by ℓ, we need to choose ψ appropriately. We now give a desirable property of a surrogate ψ, when our objective is minimizing the ℓ-risk.

Definition 1 (ℓ-calibration)**.** *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. A surrogate loss function $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$ is said to be* ℓ-calibrated *if there exists a function* pred : $\widehat{\mathcal{T}} \to [k]$ *such that $\forall \mathbf{p} \in \Delta_n$*

$$\inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}} : \mathrm{pred}(\hat{\mathbf{t}}) \notin \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \; > \; \inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \, .$$

One can show that if a surrogate ψ is ℓ-calibrated then the ψ-risk minimization algorithm is a consistent procedure for optimizing the ℓ-risk. For more details see (Ramaswamy & Agarwal, 2012).

3. Convex Calibration Dimension

From Equation 2 we see that the optimization variable f consists of d functions from $\mathcal{X}$ to $\mathbb{R}$, and hence d can be seen as a measure of complexity of the surrogate. This observation also gives us a natural definition for an intrinsic measure of difficulty of a loss matrix, which we call the convex calibration dimension.

Definition 2 (Convex calibration dimension)**.** *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. Define the* convex calibration dimension (CC dimension) *of ℓ (denoted by* CCdim(ℓ)*) as the smallest $d \in \mathbb{N}$, such that there exists a convex set $\widehat{\mathcal{T}} \subseteq \mathbb{R}^d$ and function $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$ convex in its second argument and ℓ-calibrated.*

We derive some simple bounds on the above quantity

$$\mu(\ell) \leq \mathrm{CCdim}(\ell) \leq \min(n-1, \mathrm{rank}(\ell)) \tag{3}$$

where rank(ℓ) is the rank of the loss matrix and $\mu(\ell)$ is a geometric property of the loss matrix explained in (Ramaswamy & Agarwal, 2012).

Some losses for which we can exactly determine the CC-dimension are the 0-1 loss ($\ell^{0\text{-}1}(y,t) = \mathbf{1}(y \neq t)$) and the ordinal regression loss ($\ell^{\mathrm{ord}}(y,t) = |y-t|$) both of which have $\mathcal{Y} = \mathcal{T} = [n]$. We get $\mathrm{CCdim}(\ell^{0\text{-}1}) = n-1$ and $\mathrm{CCdim}(\ell^{\mathrm{ord}}) = 1$. We can get tight bounds for some other losses.

Theorem 3. *Let $\ell : [n] \times [k] \to \mathbb{R}$ be such that $\exists \mathbf{p} \in \mathrm{relint}(\Delta_n), c \in \mathbb{R}$ with $\mathbf{p}^\top \boldsymbol{\ell}_t = c$ for all $t \in [k]$. Then*

$$\mathrm{rank}(\ell) - 2 \leq \mathrm{CCdim}(\ell) \leq \mathrm{rank}(\ell)$$

4. Application to Subset Ranking

In particular we can apply the previous theorem to certain ranking losses where the prediction space $\mathcal{T}$ is the set of all permutations of some r objects (say web pages). The losses we consider are the pairwise disagreement (ℓ^{pair}) and mean average precision (ℓ^{MAP}). We refer the reader to (Calauzènes et al., 2012; Ramaswamy & Agarwal, 2012) for details on these loss matrices. We get the following bounds

$$\begin{aligned} \frac{r(r-1)}{2} - 2 \;&\leq\; \mathrm{CCdim}(\ell^{\mathrm{pair}}) \;\leq\; \frac{r(r-1)}{2} \\ \frac{r(r-1)}{2} - 4 \;&\leq\; \mathrm{CCdim}(\ell^{\mathrm{MAP}}) \;\leq\; \frac{r(r+1)}{2} \end{aligned}$$

The lower bound is greater than r for both these losses (if $r \geq 5$) and hence any convex score based surrogate cannot be calibrated with these ranking losses, thus proving the conjecture made in (Duchi et al., 2010).

5. Conclusion and Extensions

The above results can be applied to analyze a variety of loss matrices in a unified framework. We are currently developing methods for constructing explicit low dimensional convex surrogates for certain types of loss matrices with small CC-dimension.

References

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

Calauzènes, Clément, Usunier, Nicolas, and Gallinari, Patrick. On the (non-)existence of convex, calibrated surrogate losses for ranking. In *Neural Information Processing Systems*, 2012.

Duchi, John, Mackey, Lester, and Jordan, Michael. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

Ramaswamy, Harish G. and Agarwal, Shivani. Classification calibration dimension for general multiclass losses. In *Neural Information Processing Systems*, 2012.

Tewari, Ambuj and Bartlett, Peter L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

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