Classification Calibration Dimension for General Multiclass Losses

Authored by:

Shivani Agarwal
Harish G. Ramaswamy

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 binary and multiclass 0-1 classification problems (and for certain other specific learning problems), to the general multiclass setting, and derive necessary and sufficient conditions for a surrogate loss to be classification calibrated with respect to a loss matrix in this setting. We then introduce the notion of emph{classification calibration dimension} of a multiclass loss matrix, which measures the smallest 'size' of a prediction space for which it is possible to design a convex surrogate that is classification calibrated with respect to the loss matrix. We derive both upper and lower bounds on this quantity, and use these results to analyze various loss matrices. In particular, as one application, we provide a different route from the recent result of Duchi et al. (2010) for analyzing the difficulty of designing 'low-dimensional' convex surrogates that are consistent with respect to pairwise subset ranking losses. We anticipate the classification calibration dimension may prove to be a useful tool in the study and design of surrogate losses for general multiclass learning problems.

1 Paper Body

There has been signi?cant interest and progress in recent years in understanding consistency of learning methods for various ?nite-output learning problems, such as binary classi?cation, multiclass 0-1 classi?cation, and various forms of ranking and multi-label prediction problems [1?15]. Such ?nite-output problems can all be viewed as instances of a general multiclass learning problem, whose structure is de?ned 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, de?ned by a general loss matrix, has remained elusive. In this paper, we analyze consistency of surrogate losses for general multiclass learning problems, building on the results of [3, 5?7] and others. We start in Section 2 with some background and examples that will be used as running examples to illustrate concepts throughout the paper, and formalize the notion of classi?cation calibration with respect to a general loss matrix. In Section 3, we derive both necessary and suf?cient conditions for classi?cation calibration with respect to general multiclass losses; these are both of independent interest and useful in our later results. Section 4 introduces the notion of classi?cation calibration 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 surrogate that is classi?cation calibrated with respect to the loss matrix. We derive both upper and lower bounds on this quantity, and use these results to analyze various loss matrices. As one application, in Section 5, we provide a different route from the recent result of Duchi et al. [10] for analyzing the dif?culty of designing ?low-dimensional? convex surrogates that are consistent with respect to certain pairwise subset ranking losses. We conclude in Section 6 with some future directions. 1

2

Preliminaries, Examples, and Background

Setup. We are given training examples (X1 , Y1), . . . , (Xm , Ym) drawn i.i.d. from a distribution D on X ? Y, where X is an instance space and Y = [n] = {1, . . . , n} is a ?nite set of class labels. We are also given a ?nite set T = [k] = {1, . . . , k} of target labels in which predictions are to be made, and a loss function ? : Y ? T ?[0, ?), where ?(y, t) denotes the loss incurred on predicting t ? T when the label is y ? Y. In many common learning problems, T = Y, but in general, these could be different (e.g. when there is an?abstain? option available to a classi?er, in which case k = n + 1). We will ?nd it convenient to represent the loss function ? as a loss matrix L ? Rn?k (here R+ = + [0, ?)), and for each y ? [n], t ? [k], will denote by ?yt the (y, t)-th element of L, ?yt = (L)yt = ?(y, t), and by ?t the t-th column of L, ?t = (?1t , . . . , ?nt)? ? Rn . Some examples follow:

Example 1 (0-1 loss). Here Y = T = [n], and the loss incurred is 1 if the predicted label t is different from the actual class label y, and 0 otherwise: ?0-1 (y, t) = 1(t ?= y) , where 1(?) is 1 if the argument is true and 0 otherwise. The loss matrix L0-1 for n = 3 is shown in Figure 1(a). Example 2 (Ordinal regression loss). Here Y = T = [n], and predictions t farther away from the actual class label y are penalized more heavily, e.g. using absolute distance: ?ord (y, t) = —t ? y— . The loss matrix Lord for n = 3 is shown in Figure 1(b). Example 3 (Hamming loss). Here Y = T = [2r] for some r ? N, and the loss incurred on predicting t when the actual class label is y is the number ?r of bit-positions in which the r-bit binary representations of t ? 1 and y ? 1 differ: ?Ham (y, t) = i=1 1((t ? 1)i ?= (y ? 1)i) , where for any z ? {0, . . . , 2r ? 1}, zi ? {0, 1} denotes the i-th bit in the r-bit binary representation of z. The loss matrix LHam for r = 2 is shown in Figure 1(c). This loss is used in sequence labeling tasks [16]. Example 4 (?Abstain? loss). Here Y = [n] and T = [n+1], where t = n+1 denotes ?abstain?. One possible loss function in this setting assigns a loss of 1 to incorrect predictions in [n], 0 to correct predictions, and 12 for abstaining: ?(?) (y, t) = 1(t ?= y) 1(t ? [n]) + 12 1(t = n + 1) . The loss matrix L(?) for n = 3 is shown in Figure 1(d). The goal in the above setting is to learn from the training examples a function h : X ?[k] with low expected loss on a new example drawn from D, which we will refer to as the ?-risk of h: n ? ? er?D [h] = E(X,Y)?D ?(Y, h(X)) = EX py (X)?(y, h(X)) = EX p(X)? ?h(X) , (1) y=1

where py (x) = P(Y = y — X = x) under D, and p(x) = (p1 (x), . . . , pn (x))? ? Rn denotes the conditional probability vector at x. In particular, the goal is to learn a function with ?-risk close to the optimal ?-risk, de?ned as ?

er?,? D =

inf

h:X ?[k]

er?D [h] =

inf

h:X ?[k]

EX p(X)? ?h(X) = EX min p(X)? ?t . t?[k]

(2)

Minimizing the discrete ?-risk directly is typically dif?cult computationally; consequently, one usually employs a surrogate loss function ? : Y ? T? ?R+ operating on a surrogate target space T? ? Rd for some appropriate d ? N,1 and minimizes (approximately, based on the training sample) the ?-risk instead, de?ned for a (vector) function f : X ?T? as n ? ? er? [f] = E ?(Y, f (X)) = E py (X)?(y, f (X)) . (3) X (X,Y)?D D y=1

The learned function f : X ?T? is then used to make predictions in [k] via some transformation pred : T? ?[k]: the prediction on a new instance x ? X is given by pred(f (x)), and the ?-risk incurred is er?D [pred ? f]. As an example, several algorithms for multiclass classi?cation with respect to 0-1 loss learn a function of the form f : X ?Rn and predict according to pred(f (x)) = argmaxt?[n] ft (x).

Below we will ?nd it useful to represent the surrogate loss function ? via n real-valued functions ?y : T? ?R+ de?ned as ?y (?t) = ?(y, ?t) for y ? [n], or equivalently, as a vector-valued function ? : T? ?Rn+ de?ned as ?(?t) = (?1 (?t), . . . , ?n (?t))? . We will also de?ne the sets ? ? ? ? R? = ?(?t) : ?t ? T? and S? = conv(R?) , (4) where for any A ? Rn , conv(A) denotes the convex hull of A. 1

? + , where R ? + = R+ ? {?} and ?(y, ?t) = ? ??t ? Equivalently, one can de?ne ? : Y ? Rd ?R / T? .

2

?

0 1 1

1 0 1

1 1 0

?

?

0 1 2

(a)

1 0 1
2 1 0
?
?
0 ? 1 ? 1 2
(b)
1 1 0 2 2 0 1 1 (c)
? 2 1 ? 1 ? 0
?
0 1 1 ? 1 0 1 1 1 0
1 2 1 2 1 2
(d)
? ?

Figure 1: Loss matrices corresponding to Examples 1-4: (a) L0-1 for n = 3; (b) Lord for n = 3; (c) LHam for r = 2 (n = 4); (d) L(?) for n = 3. Under suitable conditions, algorithms that approximately minimize the ?-risk based on a training sample are known to be consistent with respect to the ?-risk, i.e. to converge (in probability) to the optimal ?-risk, de?ned as ?

er?,? D =
inf er? D [f] =
f :X ?T?
inf EX p(X)? ?(f (X)) = EX inf p(X)? z = EX inf p(X)? z . z?R?
f :X ?T?
z?S?

(5) This raises the natural question of whether, for a given loss ?, there are surrogate losses ? for which consistency with respect to the ?-risk also guarantees consistency with respect to the ?-risk, i.e. guarantees convergence (in probability) to the optimal ?-risk (de?ned in Eq. (2)). This question has been studied in detail for the 0-1 loss, and for square losses of the form ?(y, t) = ay 1(t ?= y), which can be analyzed similarly to the 0-1 loss [6, 7]. In this paper, we consider this question for general multiclass losses ? : [n] ? [k]?R+ , including rectangular losses with k ?= n. The only assumption we make on ? is that for each t ? [k], ?p ? ?n such that argmint? ?[k] p? ?t? = {t} (otherwise the label t never needs to be predicted and can simply be ignored).2 De?nitions and Results. We will need the following de?nitions and basic results, generalizing those of [5?7]. The notion of classi?cation calibration will be central to our study; as Theorem 3 below shows, classi?cation calibration of a surrogate loss ? w.r.t. ? corresponds to the property that consistency w.r.t. ?-risk implies consistency w.r.t. ?-risk. Proofs of these results are straightforward generalizations of those in [6, 7] and are omitted. De?nition 1 (Classi?cation calibration). A surrogate loss function ? : [n] ? T? ?R+ is said to be classi?cation calibrated with respect to a loss function ? : [n] ? [k]?R+ over P ? ?n if there exists a function pred : T? ?[k] such that p? ?(?t) ¿ inf p? ?(?t) . ?p ? P : inf ? t?T?

? ? t?T? :pred(? t)?argmin / t p ?t

Lemma 2. Let ? : [n] ? [k]?R+ and ? : [n] ? T? ?R+ . Then ? is classi?cation calibrated with respect to ? over P ? ?n iff there exists a function pred? : S? ?[k] such that ?p ? P :

inf

? z?S? :pred? (z)?argmin / t p ?t

p? z ¿

inf p? z .

z?S?

Theorem 3. Let ? : [n] ? [k]?R+ and ? : [n] ? T? ?R+ . Then ? is classi?cation calibrated with respect to ? over ?n iff ? a function pred : T? ?[k] such that for all distributions D on X ? [n] and all sequences of random (vector) functions fm : X ?T? (depending on (X1 , Y1), . . . , (Xm , Ym)),3 P

P

er? ? er?,? D [fm] ? D

implies er?D [pred ? fm] ? ? er?,? D . ? De?nition 4 (Positive normals). Let ? : [n] ? T ?R+ . For each point z ? S? , the set of positive normals at z is de?ned as4 ? ? ? NS? (z) = p ? ?n : p? (z ? z?) ? 0 ?z? ? S? . De?nition 5 (Trigger probabilities). Let ? : [n] ? [k]?R+ . For each t ? [k], the set of trigger probabilities of t with respect to ? is de?ned as ? ? ? ? ? Q?t = p ? ?n : p? (?t ? ?t?) ? 0 ?t? ? [k] = p ? ?n : t ? argmint? ?[k] p? ?t? . Examples of trigger probability sets for various losses are shown in Figure 2. 2

Here ?n denotes the probability simplex in Rn , ?n = {p ? Rn : pi ? 0 ? i ? [n], P 3 Here ? ? denotes convergence in probability. 4 The set of positive normals is non-empty only at points z in the boundary of S? .

3

?n

i=1

pi = 1}.

Q10-1 = {p ? ?3 : p1 ? max(p2 , p3)} Qord = {p ? ?3 : p1 ? 1 = {p ? ?3 : p2 ? max(p1 , p3)} Qord Q0-1 = {p ? ?3 : p1 ? 2 2 = {p ? ?3 : p3 ? max(p1 , p2)} Qord Q0-1 = {p ? ?3 : p3 ? 3 3

1 2} 1 2 , p3 1 2}

?

(a)

1 2}

(?) Q1 (?) Q2 (?) Q3 (?) Q4

= {p ? ?3 : p1 ? = {p ? ?3 : p2 ? = {p ? ?3 : p3 ?

1 2} 1 2} 1 2}

= {p ? ?3 : max(p1 , p2 , p3) ?

1 2}

(b) (c) Figure 2: Trigger probability sets for (a) 0-1 loss ?0-1 ; (b) ordinal regression loss ?ord ; and (c) ?abstain? loss ?(?) ; all for n = 3, for which the probability simplex can be visualized easily. Calculations of these sets can be found in the appendix. We note that such sets have also been studied in [17, 18].

3

Necessary and Suf?cient Conditions for Classi?cation Calibration

We start by giving a necessary condition for classi?cation calibration of a surrogate loss ? with respect to any multiclass loss ? over ?n , which requires the positive normals of all points z ? S? to be ?well-behaved? w.r.t. ? and generalizes the ?admissibility? condition used for 0-1 loss in [7]. All proofs not included in the main text can be found in the appendix. Theorem 6. Let ? : [n] ? T? ?R+ be classi?cation calibrated with respect to ? : [n] ? [k]?R+ over ?n . Then for all z ? S? , there exists some t ? [k] such that NS? (z) ? Q?t .

We note that, as in [7], it is possible to give a necessary and suf?cient condition for classi?cation calibration in terms of a similar property holding for positive normals associated with projections of S? in lower dimensions. Instead, below we give a different suf?cient condition that will be helpful in showing classi?cation calibration of certain surrogates. In particular, we show that for a surrogate loss ? to be classi?cation calibrated with respect to ? over ?n , it is suf?cient for the above property of positive normals to hold only at a ?nite number of points in R? , as long as their positive normal sets jointly cover ?n : ? Theorem 7. Let ?r? : [n]?[k]?R+ and ? : [n]? T ?R+ . Suppose there exist r ? N and z1 , . . . , zr ? R? such that j=1 NS? (zj) = ?n and for each j ? [r], ?t ? [k] such that NS? (zj) ? Q?t . Then ? is classi?cation calibrated with respect to ? over ?n . Computation of NS? (z). The conditions in the above results both involve the sets of positive normals NS? (z) at various points z ? S? . Thus in order to use the above results to show that a surrogate ? is (or is not) classi?cation calibrated with respect to a loss ?, one needs to be able to compute or characterize the sets NS? (z). Here we give a method for computing these sets for certain surrogate losses ? and points z ? S? . Lemma 8. Let T? ? Rd be a convex set and let ? : T? ?Rn+ be convex.5 Let z = ?(?t) for some ?t ? T? such that for each y ? [n], the subdifferential of ?y at ?t can be written as ??y (?t) = ?n conv({w1y , . . . , wsyy }) for some sy ? N and w1y , . . . , wsyy ? Rd .6 Let s = y=1 sy , and let ? ? B = [byj] ? Rn?s , A = w11 . . . ws11 w12 . . . ws22 . . . . . . w1n . . . wsnn ? Rd?s ; where byj is 1 if the j-th column of A came from {w1y , . . . , wsyy } and 0 otherwise. Then ? ? NS? (z) = p ? ?n : p = Bq for some q ? Null(A) ? ?s , where Null(A) ? Rs denotes the null space of the matrix A. 5

A vector function is convex if all its component functions are convex. ? + at a point u0 ? Rd is de?ned as Recall ?that the subdifferential of a convex function ? : Rd ??R ??(u0) = w ? Rd : ?(u) ? ?(u0) ? w? (u ? u0) ?u ? Rd and is a convex set in Rd (e.g. see [19]). 6

4

We give an example illustrating the use of Theorem 7 and Lemma 8 to show classi?cation calibration of a certain surrogate loss with respect to the ordinal regression loss ?ord de?ned in Example 2: Example 5 (Classi?cation calibrated surrogate for ordinal regression loss). Consider the ordinal regression loss ?ord de?ned in Example 2 for n = 3. Let T? = R, and let ? : {1, 2, 3} ? R?R+ be de?ned as (see Figure 3) ?(y, t?) = —t? ? y— ?y ? {1, 2, 3}, t? ? R . (6) ? ? ?? ? ? ? ? ? ? Thus R? = ?(t) = —t ? 1—, —t ? 2—, —t ? 3— : t ? R . We will show there are 3 points in R? satisfying the conditions of Theorem 7.

Speci?cally, consider t?1 = 1, t?2 = 2, and t?3 = 3, giving z1 = ?(t?1) = (0, 1, 2)? , z2 = ?(t?2) = (1, 0, 1)? , and z3 = ?(t?3) = (2, 1, 0)? in R? . Observe that T? here is a convex set and ? : T? ?R3 is a convex function. Moreover, for t?1 = 1, we have ??1 (1) ??2 (1) ??3 (1)

= = =

[?1, 1] = conv({+1, ?1}) ; {?1} = conv({?1}) ; {?1} = conv({?1}) .

Therefore, we can use Lemma 8 to compute NS? (z1). Here s = 4, and ? ? 1 1 0 0 A = [+1 ?1 ?1 ?1] ; B= 0 0 1 0 . 0 0 0 1 This gives NS? (z1)

= = =

Figure 3: The surrogate ? ? p ? ?3 : p = (q1 + q2 , q3 , q4) for some q ? ?4 , q1 ? q2 ? q3 ? q4 = 0 ? ? p ? ?3 : p = (q1 + q2 , q3 , q4) for some q ? ?4 , q1 = 12 ? ? p ? ?3 : p1 ? 12 ?

= Qord 1 . ord A similar procedure yields NS? (z2) = Qord 2 and NS? (z3) = Q3 . Thus, by Theorem 7, we get that ord ? is classi?cation calibrated with respect to ? over ?3 . We note that in general, computational procedures such as Fourier-Motzkin elimination [20] can be helpful in computing NS? (z) via Lemma 8.

4

Classi?cation Calibration Dimension

We now turn to the study of a fundamental quantity associated with the property of classi?cation calibration with respect to a general multiclass loss ?. Speci?cally, in the above example, we saw that to develop a classi?cation calibrated surrogate loss w.r.t. the ordinal regression loss for n = 3, it was suf?cient to consider a surrogate target space T? = R, with dimension d = 1; in addition, this yielded a convex surrogate ? : R?R3+ which can be used in developing computationally ef?cient algorithms. In fact the same surrogate target space with d = 1 can be used to develop a similar convex, classi?cation calibrated surrogate loss w.r.t. the ordinal regression loss for any n ? N. However not all losses ? have such ?low-dimensional? surrogates. This raises the natural question of what is the smallest dimension d that supports a convex classi?cation calibrated surrogate for a given multiclass loss ?, and leads us to the following de?nition: De?nition 9 (Classi?cation calibration dimension). Let ? : [n] ? [k]?R+ . De?ne the classi?cation calibration dimension (CC dimension) of ? as ? ? CCdim(?) = min d ? N : ? a convex set T? ? Rd and a convex surrogate ? : T? ?Rn+ ? that is classi?cation calibrated w.r.t. ? over ?n , if the above set is non-empty, and CCdim(?) = ? otherwise.

From the above discussion, CCdim(?ord) = 1 for all n. In the following, we will be interested in developing an understanding of the CC dimension for general losses ?, and in particular in deriving upper and lower bounds on this.

5

4.1

Upper Bounds on the Classi?cation Calibration Dimension

We start with a simple result that establishes that the CC dimension of any multiclass loss ? is ?nite, and in fact is strictly smaller than the number of class labels n. ? ? ?n?1 Lemma 10. Let ? : [n] ? [k]?R+ . Let T? = ?t ? Rn?1 : j=1 t?j ? 1 , and for each y ? [n], let + ?y : T? ?R+ be given by ? ?y (?t) = 1(y ?= n) (t?y ? 1)2 + t?j 2 . j?[n?1],j?=y

Then ? is classi?cation calibrated with respect to ? over ?n . In particular, since ? is convex, CCdim(?) ? n ? 1.

It may appear surprising that the convex surrogate ? in the above lemma is classi?cation calibrated with respect to all multiclass losses ? on n classes. However this makes intuitive sense, since in principle, for any multiclass problem, if one can estimate the conditional probabilities of the n classes accurately (which requires estimating n?1 real-valued functions on X), then one can predict a target label that minimizes the expected loss according to these probabilities. Minimizing the above surrogate effectively corresponds to such class probability estimation. Indeed, the above lemma can be shown to hold for any surrogate that is a strictly proper composite multiclass loss [21]. In practice, when the number of class labels n is large (such as in a sequence labeling task, where n is exponential in the length of the input sequence), the above result is not very helpful; in such cases, it is of interest to develop algorithms operating on a surrogate target space in a lower-dimensional space. Next we give a different upper bound on the CC dimension that depends on the loss ?, and for certain losses, can be signi?cantly tighter than the general bound above. Theorem 11. Let ? : [n] ? [k]?R+ . Then CCdim(?) ? rank(L), the rank of the loss matrix L. Proof. Let rank(L) = d. We will construct a convex classi?cation calibrated surrogate loss ? for ? with surrogate target space T? ? Rd .

Let ?t1 , . . . , ?td be linearly independent columns of L. Let {e1 , . . . , ed } denote the standard basis ? : Rd ?Rn by in Rd . We can de?ne a linear function ? ? j) = ?t ?j ? [d] . ?(e j

? = z. Then for each z in the column space of L, there exists a unique vector u ? Rd such that ?(u) ? t) = ?t . In particular, there exist unique vectors u1 , . . . , uk ? Rd such that for each t ? [k], ?(u Let T? = conv({u1 , . . . , uk }), and de?ne ? : T? ?Rn+ as ? ?t) ; ?(?t) = ?(?k we note that the resulting vectors are always in Rn+ , since by de?nition, for any ?t = t=1 ?t ut for ?k ? ? ?k , ?(?t) = t=1 ?t ?t , and ?t ? Rn+ ?t ? [k]. The function ? is clearly convex. To show ? is classi?cation calibrated w.r.t. ? over ?n , we will use Theorem 7. Speci?cally, consider the k points zt = ?(ut) = ?t ? R? for t ? [k]. By de?nition of ?, we have S? = conv({?1 , . . . , ?k }); from the de?nitions of positive normals and trigger probabilities, it then follows that NS? (zt) = NS? (?t) = Q?t for all t ? [k]. Thus by Theorem 7, ? is classi?cation calibrated w.r.t. ? over ?n . Example 6 (CC dimension of Hamming loss). Consider the Hamming loss ?Ham de?ned in Example 3, for n = 2r . For each i ? [r], de?ne ? i ? Rn as ? +1 if (y ? 1)i , the i-th bit in the r-bit binary representation of (y ? 1), is 1 ?iy = ?1 otherwise. Then the loss matrix LHam satis?es

r

LHam =

r ? 1? ee ? ?i ?i ? , 2 2 i=1

where e is the n ? 1 all ones vector. Thus rank(LHam) ? r + 1, giving us CCdim(?Ham) ? r + 1. For r ? 3, this is a signi?cantly tighter upper bound than the bound of 2r ? 1 given by Lemma 10. 6

We note that the upper bound of Theorem 11 need not always be tight: for example, for the ordinal regression loss, for which we already know CCdim(?ord) = 1, the theorem actually gives an upper bound of n, which is even weaker than that implied by Lemma 10. 4.2

Lower Bound on the Classi?cation Calibration Dimension

In this section we give a lower bound on the CC dimension of a loss function ? and illustrate it by using it to calculate the CC dimension of the 0-1 loss. Section 5 we will explore consequences of the lower bound for classi?cation calibrated surrogates for certain types of ranking losses. We will need the following de?nition: De?nition 12. The feasible subspace dimension of a convex set C at p ? C, denoted by ?C (p), is de?ned as the dimension of the subspace FC (p) ? (?FC (p)), where FC (p) is the cone of feasible directions of C at p.7 The following gives a lower bound on the CC dimension of a loss ? in terms of the feasible subspace dimension of the trigger probability sets Q?t at certain points p ? Q?t :

Theorem 13. Let ? : [n] ? [k]?R+ . Then for all p ? relint(?n) and t ? arg mint? p? ?t? (i.e. such that p ? Q?t): 8 CCdim(?) ? n ? ?Q?t (p) ? 1 . The proof requires extensions of the de?nition of positive normals and the necessary condition of Theorem 6 to sequences of points in S? and is quite technical. In the appendix, we provide a proof in the special case when p ? relint(?n) is such that inf z?S? p? z is achieved in S? , which does not require these extensions. Full proof details will be provided in a longer version of the paper. Both the proof of the lower bound and its applications make use of the following lemma, which gives a method to calculate the feasible subspace dimension for certain convex sets C and points p ? C: ? ? Lemma 14. Let C = u ? Rn : A1 u ? b1 , A2 u ? b2 , A3 u = b3 . Let p ? C be such that ? 1? ?? 1 ?? A A , the dimension of the null space of . A1 p = b1 , A2 p ¡ b2 . Then ?C (p) = nullity 3 A A3 The above lower bound allows us to calculate precisely the CC dimension of the 0-1 loss: Example 7 (CC dimension of 0-1 loss). Consider the 0-1 loss ?0-1 de?ned in Example 1. Take p = (n1 , . . . , n1)? ? relint(?n). Then p ? Q0-1 t for all t ? [k] = [n] (see Figure 2); in particular, 0-1 we have p ? Q0-1 . Now Q can be written as 1 1 ? ? 0-1 Q1 = q ? ?n : q1 ? qy ?y ? {2, . . . , n} ? ? ? = q ? Rn : ?en?1 In?1 q ? 0, ?q ? 0, e? n q = 1} ,

where en?1 , en denote the (n ? 1) ? 1 and n ? 1 all ones and In?1 denotes ? ? vectors, respectively, the (n ? 1) ? (n ? 1) identity matrix. Moreover, we have ?en?1 In?1 p = 0, ?p ¡ 0. Therefore, by Lemma 14, we have ?? ?? ?1 1 0 . . . 0 ???1 0 1 . . . 0?? ?? ?? ?? ?? ?en?1 In?1 .. ? ?? = 0 . ?Q0-1 (p) = nullity = nullity ? ? . ? ? ?? 1 en ???1 0 0 . . . 1?? 1 1 1 ... 1 Thus by Theorem 13, we get CCdim(?0-1) ? n ? 1. Combined with the upper bound of Lemma 10, this gives CCdim(?0-1) = n ? 1. 7 For a set C ? Rn and point p ? C, the cone of feasible directions of C at p is de?ned as FA (p) = {v ? Rn : ??0 ¿ 0 such that p + ?v ? C ?? ? (0, ?0)}. 8 Here relint(?n) denotes the relative interior of ?n : relint(?n) = {p ? ?n : py ¿ 0 ?y ? [n]}.

7

5

Application to Pairwise Subset Ranking

We consider an application of the above framework to analyzing certain types of subset ranking problems, where each instance x ? X consists of a query together with a set of r documents (for simplicity, r ? N here is ?xed), and the goal is to learn a predictor which given such an instance will return a ranking (permutation) of the r documents [8]. Duchi et al. [10] showed recently that for certain pairwise subset ranking losses ?, ?nding a predictor that minimizes the ?-risk is an NP-hard problem. They also showed that several common pairwise convex surrogate losses that operate on T? = Rr (and are used to learn scores for the r documents) fail to be classi?cation calibrated with respect to such losses ?, even under some low-noise conditions on the distribution, and proposed an alternative convex surrogate, also operating on T? = Rr , that is classi?cation calibrated under certain conditions on the distribution (i.e. over a strict subset of the associated probability simplex). Here we provide an alternative route to analyzing the dif?culty of obtaining consistent surrogates for such pairwise subset ranking problems using the classi?cation calibration dimension. Speci?cally, we will show that even for a simple setting of such problems, the classi?cation calibration dimension of the underlying loss ? is greater than r, and therefore no convex surrogate operating on T? ? Rr can be classi?cation calibrated w.r.t. such a loss over the full probability simplex. Formally, we will identify the set of class labels Y with a set G of ?preference graphs?, which are simply directed acyclic graphs (DAGs) over r vertices; for each directed edge (i, j) in a preference graph g ? G associated with an instance x ? X , the i-th document in the document set in x is preferred over the j-th document. Here we will consider a simple setting where each preference graph has exactly one edge, so that —Y— = —G— = r(r ? 1); in this setting, we can associate each g ? G with the edge (i, j) it contains, which we will write as g(i,j) . The target labels consist of permutations over r objects, so that T = Sr with —T — = r!. Consider now the following simple pairwise loss ?pair : Y ? T ?R+ : ? ? ?pair (g(i,j) , ?) = 1 ?(i) ¿ ?(j) . (7) 1 1 Let p = (r(r?1) , . . . , r(r?1))? ? relint(?r(r?1)), and observe that p? ?pair = ?

Thus p

?

(?pair ?

?

?pair ??)

= 0 ??, ? ? T , and so p ? ?

Qpair ?

?? ? T .

1 2

for all ? ? T .

Let (?1 , . . . , ?r!) be any ?xed ordering of the permutations in T , and consider Qpair ?1 , de?ned by pair ? ?) ? 0 for t = 2, . . . , r! and the intersection of r! ? 1 half-spaces of the form q? (?pair ?1 ?t the simplex constraints q ? ?r(r?1) . Moreover, from the above observation, p ? Qpair ?1 satis?es pair ? ?) = 0 ?t = 2, . . . , r!. Therefore, by Lemma 14, we get p?

(?pair ?1 ?t ?? ?? ? pair pair pair ?Qpair (p) = nullity (?pair , ?1 ? ??2), . . . , (??1 ? ??r!), e ? 1

(8)

T } spans a where e is the r(r ? 1) ? 1 all ones vector. It is not hard to see that the set {?pair ? : ? ? ? r(r?1) ? r(r?1) dimensional space, and hence the nullity of the above matrix is at most r(r?1)? ?1 . 2 2 ? ? r(r?1) + 1 ? 1 = ? 2 . In Thus by Theorem 13, we get that CCdim(?pair) ? r(r ? 1) ? r(r?1) 2 2 pair particular, for r ? 5, this gives CCdim(?) ¿ r, and therefore establishes that no convex surrogate ? operating on a surrogate target space T? ? Rr can be classi?cation calibrated with respect to ?pair on the full probability simplex ?r(r?1) .

6

Conclusion

We developed a framework for analyzing consistency for general multiclass learning problems de?ned by a general loss matrix, introduced the notion of classi?cation calibration dimension of a multiclass loss, and used this to analyze consistency properties of surrogate losses for various general multiclass problems. An interesting direction would be to develop a generic procedure for designing consistent convex surrogates operating on a ?minimal? surrogate target space according to the classi?cation calibration dimension of the loss matrix. It would also be of interest to extend the results here to account for noise conditions as in [9, 10]. 8

Acknowledgments We would like to thank the anonymous reviewers for helpful comments. HG thanks Microsoft Research India for a travel grant. This research is supported in part by a Ramanujan Fellowship to SA from DST and an Indo-US Joint Center Award from the Indo-US Science & Technology Forum.

2 References

[1] G?abor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. Annals of Statistics, 32(1):30?55, 2004. [2] Wenxin Jiang. Process consistency for AdaBoost. Annals of Statistics, 32(1):13?29, 2004. [3] Tong Zhang. Statistical behavior and consistency of classi?cation methods based on convex risk minimization. Annals of Statistics, 32(1):56?134, 2004. [4] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classi?ers. IEEE Transactions on Information Theory, 51(1):128?142, 2005. [5] Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classi?cation and risk bounds. Journal of the American Statistical Association, 101(473):138?156, 2006. [6] Tong Zhang. Statistical analysis of some multi-category large margin classi?cation methods. Journal of Machine Learning Research, 5:1225?1251, 2004. [7] Ambuj Tewari and Peter Bartlett. On the consistency of multiclass classi?cation methods. Journal of Machine Learning Research, 8:1007?1025, 2007. [8] David Cossock and Tong Zhang. Statistical analysis of bayes optimal subset ranking. IEEE Transactions on Information Theory, 54(11):5140?5154, 2008. [9] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li.

Listwise approach to learning to rank: Theory and algorithm. In International Conference on Machine Learning, 2008. [10] John Duchi, Lester Mackey, and Michael Jordan. On the consistency of ranking algorithms. In International Conference on Machine Learning, 2010. [11] Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In International Conference on Arti?cial Intelligence and Statistics(AISTATS), volume 15. JMLR: W&CP, 2011. [12] David Buffoni, Cl?ement Calauz‘enes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In International Conference on Machine Learning, 2011. [13] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In Conference on Learning Theory, 2011. [14] Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In International Conference on Machine Learning, 2011. [15] Ingo Steinwart. How to compare different loss functions and their risks. Constructive Approximation, 26:225?287, 2007. [16] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. In Neural Information Processing Systems, 2003. [17] Deirdre O?Brien, Maya Gupta, and Robert Gray. Cost-sensitive multi-class classi?cation from probability estimates. In International Conference on Machine Learning, 2008. [18] Nicolas Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In ACM Conference on Electronic Commerce, 2009. [19] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. Convex Analysis and Optimization. Athena Scienti?c, 2003. [20] Jean Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, Voronoi diagrams and Delaunay triangulations. Technical report, Department of Computer and Information Science, University of Pennsylvania, 2009. [21] Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. In Neural Information Processing Systems, 2011. 9