Tight Risk Bounds for Multi-class Learning

Loubna Benabbou benabbou@emi.ac.ma
Department of Industrial Engineering, Ecole Mohammadia d'Ingenieurs, Mohammed V University, Rabat, MA
Pascal. Lang Pascal.Lang@fsa.ulaval.ca
Faculty of Business Administration, Laval University, QC, CA

Abstract

We develop minimal test-set risk bounds of any given classifier in the multi-class setting. Unlike most previous works, which typically reduce multi-class classification to a greedy series of dichotomizations, we consider a simultaneous risk bounds with valued asymmetric loss function reflecting unequal gravity of misclassification. We first observe that on an i.i.d test set, the observed losses follow a multinomial distribution which makes it possible to represent the multi-class classification in a compact form. We then formulate a mathematical program that yields the tightest possible bound. Due to a pseudo-convex constraint, a special method of centers is used to solve this problem.

1 Introduction

We consider a multi-class supervised classification problem where classifiers partition a set of examples into more than two classes. We adopt the PAC setting where data is drawn independent and identically distributed (i.i.d) according to a fixed, but unknown distribution D. A good classifier aims to minimize the generalization errors or true risk. Since the distribution D is unknown, the true risk is an unobservable quantity. In statistical learning theory, finding computable upper bounds for the true risk is a challenging area. Constructing optimal risk bounds in multi-class learning with valued asymmetric loss function is the main object of this paper. The most used approach in multi-class learning is a reduction from multi-class problem into multiple binary classifications with zero-one loss function. The majority of reduction approaches have been integrated under the framework of Error Correction Output Codes (ECOC) (Dietterich and Bakiri 1995, Allwein et al. 2000). The weaknesses of reduction approaches have been appointed in (Daniely et al. 2011, Daniely et al. 2012). A few more direct approaches to treat multiclass learning jointly have been studied in (Vapnik 1998, Weston and Watkins 1999, Fung and Mangasarian 2001, Crammer and Singer 2001, Aiolli et al. (2005), He et al. (2012), Ramaswamy and Agarwal (2016)). In this paper we consider a more realistic multi-class setting with valued loss function and different cost of errors. In the following section we present the multi-class setting with a compact form based on the multinomial distribution of observed losses on an i.i.d test set. Foundations of optimal multi-class test-set bounds are given in section 3. In section 4 we present bounds optimization approaches. We formulate a mathematical program that yields the tightest possible bound. Due to a pseudo-convex constraint, a special method of centers is used to solve this problem. Proofs are provided in Appendix A.

2 Multi-class Setting

We are concerned with a multi-class problem in which each example $\tilde{z} = (\tilde{x}, \tilde{y})$ is constituted from an input-output pair (x, y) where $x \in \mathrm{X}$ and $y \in \mathrm{Y}$; such that $\mathrm{Y} > 2$; a finite set of *observed* classes. Let C a set of *predicted* classes such that $\mathrm{Y} \subseteq \mathrm{C}$. In our multi-class setting, the classification task consists in assigning to each input object x a predicted class c where, possibly, $\mathrm{C} \neq \mathrm{Y}$ if the context so dictates. Many reasons may justify adding to Y new predicted classes like "Unclassified" "hesitation between classes y_1 and y_2", etc. In practice, it's more prudent to not classify an example than to give him a wrong class. We consider a test set $\tilde{S}_n = (\tilde{z}^1, ..., \tilde{z}^n)$ of n *i.i.d.* examples drown from unknown distribution D. We wish to assess a given a classifier $h : \mathrm{X} \to \mathrm{C}$ on the test set S_n. The accuracy of classifier h is measured through a *loss function*. In the binary case, this is usually a zero-one loss function. In our multi-class context, however, different types of errors may deserve different error costs according to their relative gravity. For example in medical diagnosis, the cost of error for classifying a patient with the true class "cancer" in the class "cold" is more than the cost of error to predict the class "unclassified" and ask for more medical examinations. The asymmetric and valued nature of the loss function reflects the uncertainty of the classification decision.

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

We thus posit a more general *valued, cardinal, normalized,* loss function $\mathrm{Q} : \mathrm{C} \times \mathrm{Y} \to [0,1]$, with $\mathrm{Q}(y,y) = 0 \forall y \in \mathrm{Y}$ and $Max_{c,y}\mathrm{Q}(c,y) = 1$. The true error that the classifier h predicted c and the true class is y is defined as the unknown probability: $\pi_{c,y} = Pr\{h(\tilde{x}) = c | \tilde{y} = y\}, (c,y) \in \mathrm{C} \times \mathrm{Y}$. The true risk associated to h is then defined as the expected loss $R_h = E[\mathrm{Q}(h(\tilde{x}), \tilde{y})]$, Since the true risk R_h is unknown, one is interested in upper-bounding this risk. The related empirical value, the empirical risk,R_e , the observed number of errors in each case (c,y).

In order to assess the performance of a multi-class classifier h on a random test set S_n , it is necessary to consider all possible *error cases* (including non-errors)$(c,y) \in \mathrm{C} \times \mathrm{Y}$. Define $\tilde{N}_{cs}$ as the (h-dependent) number of observations in $\tilde{S}_n$ falling into error case (c,y) (so that $\sum_{(c,y) \in \mathrm{C} \times \mathrm{Y}} \tilde{N}_{c,y} \equiv n$). In the simple case of binary classification the probability of observing k errors (heads) out of n examples is a binomial distribution (Langford 2005). In our multi-class setting; by the *i.i.d* assumption, the random array $\tilde{N} = (\tilde{N}_{c,y} | (c,y) \in \mathrm{C} \times \mathrm{Y})$ has another familiar distribution in statistics, the multinomial distribution with unknown probabilities: $\pi_{c,y}$.

Now consider the ordered set of distinct s values that may be incurred of the valued loss function Q(error costs) noted $0 = q_1 < q_2 < ... < q_s = 1$. There is an aggregation function a mapping the set $\mathrm{C} \times \mathrm{Y}$ of error cases into the set $\{1, ..., s\}$ of error costs such that: $\mathrm{Q}_{c,y} = q_i \forall (c,y) \in a^{-1}(i), 1 \leq i \leq s$. Consider the random vector $\tilde{K} = (\tilde{K}_1, ..., \tilde{K}_s)$, where $\tilde{K}_i = \sum_{(c,y) \in a^{-1}(i)} \tilde{N}_{cy}$ is the number of observations from $\tilde{S}_n$ falling into error cost category i, $1 \leq i \leq s$. In the sequel, "multi-class" is taken to mean that the error cost takes on at least one fractional value, reflecting intensity or relative gravity of errors – so that $s > 2$. See example in appendix B for repartition of examples by error cost category for: $|Y| = 3; |C| = 4; s = 3; q_1 = 0; q_2 = 0.5, q_s = 1; n = 30$.

Let $K = \{k \in Z^s_+ \mid e^T k = n\}$ denote the range of $\tilde{K}$, the empirical risk is defined by $\tilde{r} = \frac{1}{n} q^T \tilde{K}$. The following remark will help characterize our multi-class setting:

$\tilde{K} = (\tilde{K}_1, ..., \tilde{K}_s)$ *has a multinomial distribution with probabilities* $p_i = \sum_{(c,y) \in a^{-1}(i)} \pi_{cy}, 1 \leq i \leq s$.

As a consequence, the true risk can equivalently be expressed as: $R_h = q^T p = \sum_{i=1}^{s} q_i p_i$. Let define $K_r = \{k \in K \mid q^T k \leq nr\}$ the set of outcomes for $\tilde{K}$ whose empirical risk does not exceed r. According to the previous remark, the probability of an empirical error less than or equal to r is:

$Pr\{\tilde{r} \leq r \mid p\} = \sum_{k \in K_r} Pr\{\tilde{K} = k \mid p\} = \sum_{k \in K_r} C_k \prod_{i=1}^{s} p_i^{k_i}$ With $C_k = \frac{n!}{\prod_{i=1}^{s} k_i!}$.

3 Optimal Multi-class Test Set Bounds

A *bound* will be meant to provide an upper confidence interval on an unknown true risk. The "test set" context is concerned with the *evaluation* of a *given* classifier *h*. By contrast, the *design* question is one of choosing a particular classifier from a possibly vast family. The fact that the classifier *h* is given does not, however, imply that its inputs $\tilde{x}$ or its inner workings are known. Indeed, the classifier's performance is evaluated only in terms of the observed output pairs $(h(x_j), y_j)$, independently of the underlying classification model. Whereas the multinomial distribution is an exact representation of the error occurrence process, it will later become apparent that approximations are also of interest. Therefore, our framework will encompass more abstract stochastic error models, with the proviso that they are entirely characterized by a probability vector *p*. In this spirit, the following definition is a generalization of Langford's (Langford 2005).

Definition 1 A *tail bound is a function* $\overline{B} : [0,1]^2 \to [0,1]$ *such that* $\forall r \in [0,1], \delta \in (0,1]$, *and* $\forall p \in U$ *such that* $q^T p > \overline{B}(r, \delta) : Pr\{\tilde{r} \leq r | p\} < \delta$.

Where $U = \{x \in \Re^s_+ | e^T x = 1\}$ and $e = (1, 1, ..., 1)$. r is a parametric threshold which will later assume the value of an observed empirical risk. Prior to any observation, given some confidence level δ, the bound is stated as a function of this threshold. The bound's defining property is that whatever the threshold, under a true risk greater than the bound, the probability of observing an empirical risk below the threshold does not exceed δ. A *minimal* bound (one which cannot be tightened) is, of course, unique. The set of pairs (r, δ) over which the bound exists, will depend on the particular probabilistic model considered. Minimal bounds cannot in general be expressed analytically. We will instead seek numerical bounds which (as "best") will entail some optimization. Several mathematical programs, with differing computational advantages, are conceivable to represent definition 1 under specific conditions. This is illustrated below with two families of formulations.

Let consider the function $F : F(p;r) = Pr\{\tilde{r} \leq r \mid p\}$, we will henceforth assume that F is continuous in p. A direct representation of definition 1 is the following mathematical program:

$$B_1(r,\delta) = Sup_p\{q^T p | p \in U, F(p;r) \geq \delta\} \tag{1}$$

This definition implies that for any p such that $q^T p > B_1(r,\delta), F(p;r) < \delta$. It is tempting to replace (1) with the following variant:

$$B_1^{'}(r,\delta) = Sup_p\{q^T p | p \in U, F(p;r) = \delta\}$$

Function F will be called *risk-complete* if $\forall r \in [0,1), Lim_{p \rightarrow e_s} F(p;r) = 0$.

Proposition 1:

- (*i*) B_1 *is a risk bound in the sense of definition* 1.
- (*ii*) *If* F *is risk-complete*, $B_1 = B_1^{'}$.

A second, more indirect computational scheme, is in two steps as follows:

$$V(\beta;r) = Sup_p F(p;r) \tag{2}$$

$$\text{S.t.: } q^T p \geq \beta$$

$$p \in U$$

$$B_2(r,\delta) = Sup\{\beta | V(\beta;r) \geq \delta\} \tag{3}$$

In 2, $\beta \in [0,1]$ is a parameter that will eventually assume the value of the bound. For each possible value of β, we seek the largest possible probability of observing an empirical error equal to *r*, given that the true risk is at least β. Observe that, by a simple relaxation argument,$V(.;r)$ is non-increasing. It follows that the bound B_2, guarantees that for any $\beta > B_2(r,\delta)$ one has $V(\beta;r) < \delta$, which, given the maximality of $V(.;r)$, also holds for the "true" unknown empirical risk distribution.

Proposition 2: $B_2 = B_1$.

We now consider the exact probabilistic model of the multinomial distribution and study the computability of bound $B_1(r,\delta)$. We remind that: $F(p;r) = Pr\{\tilde{r} \leq r \mid p\} = \sum_{k \in K_r} C_k \prod_{i=1}^{s} p_i^{k_i}$. It is readily verified that F is risk-complete. We now establish a second important property that is the key to computing exact bound.

Proposition 3: *F is pseudo-concave in p on* $\Re^s_{++}$.

This property has two important implications: (*i*) any local optimum for problem (1) is also a global optimum, and (*ii*) the Karoush-Kuhn-Tucker (KKT) conditions are necessary and sufficient to characterize optimality (see e.g. Mangasarian 1969. Hence solving problem (1) reduces to finding a KKT point. We thus have an operational criterion to solve problem (1) to optimality when the error occurrence process is modeled as a multinomial distribution (an exact representation). However, any computational scheme to this end will require several evaluations of $F(p;r)$; each such evaluation entails an enumeration of K_r. It can be verified that $|K| = O(s^n)$ and that, for non-trivial values of r, $|K_r|$ grows similarly.

4 Bound Optimization

The best bound is a solution of the mathematical program:

$$B(r,\delta) = Max_p q^T p \tag{4}$$

$$\text{S.t. } F(p;r) \geq \delta \text{ (4.a)}$$

$$e^T p = 1$$

$$p \geq 0$$

where *F* is pseudo-concave in *p*. The only complicating constraint in this problem is (4.a). Non-concavity of *F* precludes constructing inner or outer envelopes on its hypograph. However, the level set: $\Gamma_\delta = \{p \in U | F(p; r) \geq \delta\}$ is convex. Il follows that if a point p' is not in the relative interior of Γ_δ, the inequality $\nabla_p F(p'; r)(p - p') \geq 0$, is valid for Γ_δ. Our solution strategy will thus entail constructing increasingly tight polyhedral relaxations of Γ_δ. Let $P = \{p \in U | a^i p \geq b^i, i \in I\}$ be a polyhedron containing Γ_δ, and $\overline{\beta} = max\{q^T p | p \in P\}$. It is clear that $\overline{\beta}$ is an upper bound on $B(r, \delta)$. Now consider a feasible point $\hat{p} \in \Gamma_\delta$. Let $\underline{\beta} = q^T \hat{p}$, and $P_{\underline{\beta}} = \{p \in P \mid q^T p \geq \underline{\beta}\}$ (a *localization set*). Since $\underline{\beta}$ is a lower bound on $B(r, \delta)$, one has $P_{\underline{\beta}}$ contains all optimal solutions to (4).

Our cut generation mechanism is a variant of the general *Method of centers* (Huard, 1976), which enjoys good stability properties. From a point $\bar{p}$ in the relative interior of *P*, we can measure a weighted distance to each frontier hyperplane of *P*. A center of *P* is defined as a point $\bar{p}$ which maximizes the smallest such (weighted) distance. Since the Euclidian distance from a given point $\bar{p}$ to a hyperplane $\{p | a^T p = b\}$ is $|a^T \bar{p} - b| / \|a\|$, a center $\bar{p}$ of $P_{\underline{\beta}}$ can be found by solving the linear program:

$$Max_{p,z} z \tag{5}$$

$$\begin{gathered} \text{S.t. } a^i p - \|a^i\| z \geq b^i i \in I \\ q^T p - \epsilon \|q\| z \geq \underline{\beta} \\ p \in U \end{gathered}$$

The algorithm generates a sequence $\langle P^t_{\underline{\beta}_t} \rangle$ of localization sets, two sequences of points $\langle x_t \rangle \subset \Gamma_\delta$ and $\langle y_t \rangle \subset P^t_{\underline{\beta}_t} \setminus \Gamma_\delta$ and the associated bounds $\underline{\beta}_t = q^T x_t$, $\bar{\beta}_t = q^T y_t$. At each iteration *t*, a center $\bar{p}_t \in p^t_{\underline{\beta}_t}$ is computed. If $F(\bar{p}_t) > \delta$, this center becomes the next x_{t+1}. Otherwise a cut through $\bar{p}_t$ is generated, and a new upper-bound-yielding y_{t+1} is found. In both cases, we obtain $0 < \bar{\beta}_{t+1} - \underline{\beta}_{t+1} < \bar{\beta}_t - \underline{\beta}_t$ and $p^{t+1}_{\underline{\beta}_{t+1}} \subset p^t_{\underline{\beta}_t}$.

Proposition 4 *Any accumulation point of the sequence $\langle x_t \rangle$ is an optimal solution of (4).*

In practice, the algorithm can be accelerated with approximate line searches. It will be interrupted upon reaching a prescribed degree of accuracy. The current value of $\bar{\beta}_t$ will then serve as a conservative estimate of the test-set bound (note that the $\bar{\beta}_t$'s are tail bounds in the sense of Definition 1).

5 Conclusion

This paper's contribution is focused on multi-class test set bounds optimization. The most disadvantage of test set bound is that data used for testing can not be used for training. Several extensions to training-set and PAC-Bayes bounds are conceivable. Furthermore, the complexity of problem (1) under the multinomial distribution is a strong incentive to seek computationally less demanding approximations. A number of avenues are conceivable toward that end. An obvious candidate is the use of a multivariate normal distribution as an approximation to the multinomial distribution. A sequel to this paper will empirically study the bound's behavior under different sets of parameter values (e.g. sample size, number of classes, class penalties, required confidence level, observed empirical risk) with a view on one hand to assessing the quality of the bound, but foremost to stimulating the development of new approximations.

References

[1] Aiolli, F., Alessandro, S. and Singer, Y.(2005) Multiclass classification with multi-prototype support vector machines. *Journal of Machine Learning Research*, 6:817–850.

[2]Allwein, E.L., Schapire, R.E. and Singer, Y. (2000) Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141.

[3] Crammer, K. and Singer, Y. (2002) On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47:201–233.

[4] Daniely, A., Sabato, S., Ben-david, S. and Shalev-shwartz, S. (2011) Multiclass learnability and the erm principle. In COLT 2011, *The 24th Annual Conference on Learning Theory*.

[5] Daniely, A., Sabato, S., S. and Shalev-shwartz, S. (2012) Multiclass learning approaches: A theoretical comparison with implications. In NIPS 2012 - Advances in Neural Information Processing Systems.

[6] Dietterich, T.G. and Bakiri, G. (1995) Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286.

[7] Fung, G.M. and Mangasarian, O.L. (2001) Multicategory proximal support vector machine classifiers. *Machine Learning*, 59:77–97, 2001.

[8] He, X., Wang, Z., Jin, C., Zheng, Y., and Xue, X.Y. (2012) A simplified multi-class support vector machine with reduced dual optimization. *Pattern Recognition Letters*, 33:71–82.

[9] Huard, P. (1967) Resolution of mathematical programming problems with nonlinear constraints by the method of centers. In *Nonlinear Programming*, pages 206–219. North Holland, Amesterdam.

[10] Langford, J. (2005) Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6, 273-306.

[11] Mangasarian, O. (1969) : *Non-linear programming*. McGraw-Hill.

[12] Ramaswamy, H.G. and Agarwal, S. (2016) Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17:1–45.

[13] Vapnik, V.N. (1998) *Statistical learning theory*. Wiley, New York.

A Appendix A

A.1 Proof of proposition 1(ii)

Let p satisfy $F(p,r) > \delta$. Consider the path

$$\bar{P}(t) = (1-t)p + te_s, t \in [0,1]$$

Since $F(\bar{P}(0);r) > \delta > F(\bar{P}(1^-);r)$ and since $F(.;r)$ is continuous, there is an $\alpha \in (0,1)$ such that $F(\bar{P}(\alpha);r) = \delta$. Since $q^T\bar{P}(t)$ is strictly increasing in t , $q^T\bar{P}(\alpha) > q^Tp$. Therefore, p is not optimal for (1).

A.2 Proof of proposition 2

$$\begin{aligned} B_2(r,\delta) &= Sup_\beta\{\beta | Sup_{p\in V(\beta)}\{F(p;r)\} \geq \delta\} \\ &= Sup_{\beta,p}\{\beta | p \in U, q^Tp \geq \beta, F(p;r) \geq \delta\} \\ &= Sup_p\{q^Tp | p \in U, F(p;r) \geq \delta\} \\ &= B_1(r,\delta) \end{aligned}$$

A.3 Proof of proposition 3

Consider any $p > 0$. Define $D(p) = Diag(p)^{-1}$. Then $\forall k \in K$:

- (*i*) $f(p;k) > 0$
- (*ii*) $\nabla_p f(p;k) = f(p;k)D(p)k$
- (*iii*) $\nabla^2_{pp} f(p;k) = f(p;k)D(p)[-Diag(k) + f(p;k)kk^T]D(p)$

Then, $\forall r \in [0,1]$:

- (*iv*) $F(p;r) > 0$
- (*v*) $g_r(p) = \nabla_p F(p;r) = D(p)\gamma_r(p)$,where $\gamma_r(p) = \sum_{k\in K_r} f(p;k)k$
- (*vi*) $H_r(p) = \nabla^2_{pp} F(p;r) = -A_r(p) + g_r(p)g_r(p)^T$, whith $A_r(p) = D(p)Diag(\gamma_r(p))D(p)$(a diagonal matrix)

Function F is pseudo-concave in p on $\Re^s_{++}$ if and only if for any $p \in \Re^s_{++}$ and for any admissible variation dp (i.e. such that $p + dp \in \Re^s_{++}$ one has

$$\nabla_p F(p;r)^T dp \leq 0 \Longrightarrow F(p + dp;r) \leq F(p;r)$$

Let $s' = max\{i|1 \leq i \leq s, q_i \leq nr\}$. Clearly, $\forall i > s', \forall k \in K_r.k_i = 0-$ and conversely, $i \leq s' \implies \exists k \in K_r : k_i > 0$, which, with (*i*) implies $(\gamma_r(p))_i > 0$. Thus for any admissible variation dp such that $dp_i = 0 \forall i \leq s'$, we have $f(p + dp; k) = f(p; k) \forall k \in K_r$, hence $F(p + dp; r) = F(p; r)$.

Consider now the alternative case of admissible variations dp such that $dp_i \neq 0$ for some $i \leq s'$. Since $F(.; r)$ is twice continuously differentiable,

$$\begin{aligned} F(p + dp; r) &= F(p; r) + g_r(p)^T dp + \tfrac{1}{2} dp^T H_r(p) dp + o(\|dp\|^2) \\ &= F(p; r) + g_r(p)^T dp + \tfrac{1}{2}(g_r(p)^T dp)^2 - \tfrac{1}{2} dp^T A_r(p) dp + o(\|dp\|^2) \end{aligned}$$

From the preceding assumption, $-\frac{1}{2} dp^T A_r(p) dp < 0$. Furthermore, $g_r(p)^T dp + \frac{1}{2}(g_r(p)^T dp)^2 \leq 0$ for any dp such that $0 \leq g_r(p)^T dp \leq 2$. This latter inequality will hold by choosing $\|dp\|$ finitely but sufficiently small. It follows that $F(p + dp; r) \leq F(p; r)$ in some open neighborhood of p. By transitivity, this also holds for any admissible dp .

A.4 Proof of proposition 4

As long as $\overline{\beta}_t > \underline{\beta}_t$, the algorithm generates a new upper bound $\overline{\beta}_{t+1}$ or a new lower bound $\underline{\beta}_{t+1}$ such that $\overline{\beta}_t > \overline{\beta}_{t+1} > \underline{\beta}_t$ or $\overline{\beta}_t > \underline{\beta}_{t+1} > \underline{\beta}_t$. Hence $\langle \overline{\beta}_t \rangle$ and $\langle \underline{\beta}_t \rangle$ converge to a common limit β^* , which is the optimal value of the test-set bound. Since Γ_δ is compact, any accumulation point $x*$ of $\langle x_t \rangle$ is in Γ_δ , yields an objective value of $q^T x* = \beta*$, and is therefore optimal for (4).

B Appendix B

Table 1: Repartition of examples by error cases

$c \setminus y$	1	2	3
1	0	1	0
2	0.5	0	1
3	1	0.5	0
4:Unclassified	1	0.5	1

Error category i	1	2	3
Error cost q_i	0	0.5	1
Error case (c,y): $a^{-1}(i)$	(1,1);(2,2);(3,3);(1;3)	(2,1);(3,2);(4,2)	(1,2);(2,3);(3,1);(4,1);(4,3)
Number of examples $\tilde{N}_{cy}$	20	3	7