Equivalence of Multicategory SVM and Simplex Cone SVM: Fast Computations and Statistical Theory

Guillaume A. Pouliot [1]

Abstract

The multicategory SVM (MSVM) of Lee et al. (2004) is a natural generalization of the classical, binary support vector machines (SVM). However, its use has been limited by computational difficulties. The simplex-cone SVM (SCSVM) of Mroueh et al. (2012) is a computationally efficient multicategory classifier, but its use has been limited by a seemingly opaque interpretation. We show that MSVM and SCSVM are in fact exactly equivalent, and provide a bijection between their tuning parameters. MSVM may then be entertained as both a natural and computationally efficient multicategory extension of SVM. We further provide a Donsker theorem for finite-dimensional kernel MSVM and partially answer the open question pertaining to the very competitive performance of One-vs-Rest methods against MSVM. Furthermore, we use the derived asymptotic covariance formula to develop an inverse-variance weighted classification rule which improves on the One-vs-Rest approach.

1. Introduction

Support vector machines (SVM) is an established algorithm for classification with two categories (Vapnik, 1998; Smola and Schlkopf, 1998; Steinwart and Christmann, 2008; Friedman et al., 2009). The method finds the maximum margin separating hyperplane; it finds the hyperplane dividing the input space (perhaps after mapping the data to a higher dimensional space) into two categories and maximizing the minimum distance from a point to the hyperplane. SVM can also be adapted to allow for imperfect classification, in which case we speak of soft margin SVM.

Given the success of SVM at binary classification, many attempts have been made at extending the methodology to accommodate classification with $K > 2$ categories (Sun et al., 2017; Dogan et al., 2016; Lopez et al., 2016; Kumar et al., 2017, survey available in Ma and Guo, 2014). Lee, Lin and Wahba (2004) propose what is arguably the natural multicategory generalization of binary SVM. For instance, their multicategory SVM (MSVM) is Fisher consistent (i.e., the classification rule it produces converges to the Bayes rule), which is a key property and motivation for the use standard SVM. Furthermore, it encompasses standard SVM as a special case.

However, the method has not been widely used in application, nor has it been studied from a statistical perspective, the way SVM has been. Amongst the machine learning community, MSVM has not gathered popularity commensurate to that of SVM. Likewise, three major publications (Jiang et al., 2008; Koo et al., 2008; Li et al., 2011) have established Donsker theorems for SVM, and none have done so for MSVM.

Interestingly, computation and statistical analysis of MSVM are hindered by the same obstable. The optimization problem which MSVM consists of is done under a sum-to-zero constraint on the vector argument. This makes both the numerical optimization task and the statistical asymptotic analysis of the estimator more challenging. The numerical optimization is substantially slowed down by the equality constraint[1] as detailed in Table 1. Likewise, standard methods for deriving Donsker theorems and limit distribution theory do not apply to such constrained vectors of random variables.[2]

In a separate strain of literature, Mroueh, Poggio, Rosasco and Slotine (2012) have proposed the simplex-cone SVM (SC-SVM), a multicategory classifier developped within the vector reproducing kernel Hilbert space set-up. The SC-SVM optimization program is computationally tractable, and in particular does away with the equality constraint slowing down the primal and dual formulations of MSVM (Lee

[1]Harris School of Public Policy, University of Chicago, Chicago, IL, USA. Correspondence to: Guillaume A. Pouliot <guillaumepouliot@uchicago.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]In fact, a "hack" sometimes used is to ignore the equality constraint in the primal or dual formulation. This can result in arbitrarily large distorsions of the optimal solution.

[2]For instance, the covariance matrix of a vector of random variables constrained to sum to zero is not positive definite.

et al., 2004). Nevertheless, its use has remained marginal, arguably due to more limited interpretability, e.g., the notion of distance is captured via angles and the nesting of binary SVM as a special case is not entirely straightforward.

As our main contribution, we show that MSVM and SC-SVM are in fact exactly equivalent. As a direct consequence, we deliver faster computations for MSVM. Simulations such as those presented in Table 1 display speed gains or an order of magnitude. Furthermore, the equivalence with an unconstrained estimator allows for statistical analysis of MSVM. As a second contribution, we deliver a Donsker theorem for MSVM, as well as an asymptotic covariance formula with sample analog. A third contribution is to use the asymptotic analysis result to propose a statistically efficient, inverse-variance weighted modification of One-vs-Rest. Finally, as a fourth contribution, we show that the asymptotic analysis allows us to partially answer, with analytic characterizations, the open question relating to the very competitive performance of the seemingly more naive One-vs-Rest method against MSVM.

The fourth contribution is important because it provides analytical substance to a long-standing open question. To be sure, the different attempts at developping a multicategory generalization of binary SVM can be understood as subscribing to one of two broad approaches. The first approach consists in doing multicategory classification using the standard, binary SVM. For instance, the popular One-vs-Rest approach works as follows: to predict the category of a point in a test set[3] (i.e. out of sample), run K binary SVMs where the first category is one of the original K categories, and the second category is the union of the remaining $K-1$ categories. The predicted category is the one that was picked against all others with the greatest "confidence". In practice, the confidence criteria used is the distance of the test point to the separating hyperplane (we show in Subsection 3.1 that even this can be improved according to statistical considerations). The second approach consists in generalizing the standard SVM to develop a single machine which implements multicategory classification solving a single, joint optimization problem. Many such algorithms have been suggested (Weston and Watkins, 1999; Crammer and Singer, 2002; Lee et al., 2004). Intuition would suggest that joint optimization makes for a more statistically efficient procedure, and for superior out-of-sample prediction performance. However, in a quite counterintuitive turn of events, it has been widely observed in practice that multicategory classification with binary machines offers a performance (for instance, in out-of-sample classification) which is competitive with, and sometimes superior to, that of single-machine multicategory SVM algorithms. This phenomenon is widely acknowledged (Rifkin and Klautau, 2004) but very little theory has been put forth to explain it.

We make some progress towards an analytical characterization of the comparative performance, and are able to suggest an explanation as to the competitive, and sometimes superior, empirical performance of One-vs-Rest compared to MSVM. We argue that, in some respect, One-vs-Rest makes a more efficient use of the information contained in the data.

The remainder of the paper is organized as follows. Section 2 defines both MSVM and SC-SVM, and contains the proof of the equivalence between the two methods. Section 3 gives the Donsker theorem for MSVM, and describes how the asymptotic distribution may be used for more efficient classification. Section 4 suggests an analytical explanation for the surprisingly competitive performance of One-vs-Rest classifiers versus MSVM. Section 5 discusses and concludes.

2. Equivalence

The multicategory SVM (MSVM) of Lee et al. (2004) is arguably the more elegant and natural generalization of SVM to multicategory data. However, its implementation, even for moderate size data sets, is complicated by the presence of a sum constraint on the vector argument.

The simplex encoding of Mroueh et al. (2012) is relieved of the linear constraint on the vector argument. However, we believe the simplex encoding is not more widely used because it is not known what standard encoding it corresponds to, making it challenging for practitioners to carry out interpretable classification analysis. The following result resolves both issues, making it of practical interest for analysts and researchers using multicategory classification methods.

We define MSVM and SC-SVM, and establish their equivalence. The presentation is done with finite-dimensional kernels for ease of exposition. Remark 3 details the generalization to infinite-dimensional kernels in reproducing kernel Hilbert spaces.

With K categories, data is of the form $(x_i, y_i) \in \mathbb{R}^p \times \{1, ..., K\}$, $i = 1, ..., N$. When carrying out multicategory classification, different choices of *encodings* of the category variables y_i lead to optimization problems that are differently formulated and implemented.

For their multicategory SVM (MSVM), Lee et al. (2004) encode y_i associated with category $k \in \{1, ..., K\}$ as a K-tuple with 1 in the k^{th} entry and $\frac{-1}{K-1}$ in every other entry. For instance,

$$\text{"}y_i \text{ in category 2"} \Leftrightarrow y_i = \left(\frac{-1}{K-1}, 1, \frac{-1}{K-1}, \cdots, \frac{-1}{K-1}\right).$$

The loss function they suggest is then based on the *difference* between the decision function and the encoded y_i's.

[3] Or the fitted category of a point in the training set.

Specifically, in the case of finite-dimensional feature maps, they suggest minimizing

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i)\cdot[Wx_i+b-y_i]_+ + \frac{\lambda}{2}|||W|||, \qquad (1)$$

where $|||W||| = \mathrm{trace}(W^TW)$, and $L(y_i) = 1_K - e_{y_i}$ is a vector that has 0 in the k^{th} entry when y_i designates category k, and a 1 in every other entry. Importantly, the decision function is constrained to sum to zero, i.e. $1_k^T(Wx+b) = 0,\ \forall\ x$. The function $[\cdot]_+$applies pointwise to its vector argument.

Mroueh et al. (2012) preconize an encoding that does away with the sum-to-zero constraint. The loss function they suggest is based on the *inner product* between the decision function and their encoding of y_i's. Likewise in the finite-dimensional case, the penalized minimization problem entailed by their loss function is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{y'\neq y}\left[\frac{1}{K-1}+\left\langle c_{y'},\tilde{W}x_i+\tilde{b}\right\rangle\right]_+ + \frac{\tilde{\lambda}}{2}|||\tilde{W}|||, \qquad (2)$$

where c_y is a unit vector in $\mathbb{R}^{K-1}$ which encodes the response; it is a row of a simplex coding matrix, which is the key building block of their construction.

A simplex coding matrix (Mroueh et al., 2012; Pires et al., 2013) is a matrix $C\in\mathbb{R}^{K\times(K-1)}$ such that its rows c_k satisfy $(i)\ \|c_k\|_2^2 = 1$; $(ii)\ c_i^Tc_j = -\frac{1}{K-1}$ for $i\neq j$; and $(iii)\ \sum_{k=1}^{K} c_k = 0_{K-1}$. It encodes the responses as unit vectors in $\mathbb{R}^{K-1}$ having maximal equal angle with each other. Further note that, because its domain is a $(K-1)$-dimensional subspace of $\mathbb{R}^K$, any given C has a unique inverse operator $\tilde{C}$ defined on the image $\{x\in\mathbb{R}^K : 1_K^Tx = 0\}$.

For a given choice of simplex encoding defined by C, the operator $C:\mathbb{R}^{K-1}\to\mathbb{R}^K$ can be thought of as mapping decision functions and encoded y's from the unrestricted simplex encoding space to the standard, restricted encoding space used by Lee et al. (2004).

A natural question is then: if $f(x) = Wx+b$ and $\tilde{f}(x) = \tilde{W}x+\tilde{b}$ are optimal solutions to (1) and (2), respectively, are $\tilde{C}(Wx+b)$ and $C(\tilde{W}x+\tilde{b})$ then optimal solutions to (2) and (1), respectively? We show that this is in fact the case. That is, *both problems are exactly equivalent*.

We now show the problems are equivalent. The equivalence of the loss functions is straighforward. Indeed,

$$\begin{aligned}\sum_{y'\neq y}\left[f_{y'}(x)+\tfrac{1}{K-1}\right]_+ &= \sum_{y'\neq y_i}\left[\left(C\tilde{f}(x)\right)_{y'}+\tfrac{1}{K-1}\right]_+\\ &= \sum_{y'\neq y_i}\left[\left\langle c_{y'},\tilde{f}(x)\right\rangle+\tfrac{1}{K-1}\right]_+,\end{aligned} \qquad (3)$$

which is exactly the SC-SVM loss of Mroueh et al. (2004). Writing out f and $\tilde{f}$ as linear functions, the identity becomes

$$\begin{aligned}&\sum_{y'\neq y}\left[\omega_{y'}x+b_{y'}+\tfrac{1}{K-1}\right]_+\\ =\ &\sum_{y'\neq y_i}\left[\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle+\tfrac{1}{K-1}\right]_+\end{aligned} \qquad (4)$$

with $f(x) = Wx+b$ and $\tilde{f}(x) = \tilde{W}x+\tilde{b}$, and $\omega_{y'}$ is the $(y')^{\text{th}}$ row of W.

Equality (up to a change of tuning parameter) of the penalty relies on the key observation of this exercise, which is that C^TC is the diagonal matrix $\frac{K}{K-1}I_{K-1}$. It then immediately follows that

$$\begin{aligned}\tfrac{K-1}{K}\mathrm{trace}\left(\tilde{\mathrm{W}}^{\mathrm{T}}\tilde{\mathrm{W}}\right) &= \tfrac{K-1}{K}\mathrm{trace}\left(\mathrm{W}^{\mathrm{T}}\mathrm{C}^{\mathrm{T}}\mathrm{CW}\right)\\ &= \mathrm{trace}\left(\mathrm{W}^{\mathrm{T}}\mathrm{W}\right).\end{aligned} \qquad (5)$$

In conclusion, we have

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i)\cdot[Wx_i+b-y_i]_+ + \frac{\lambda}{2}|||W|||$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{y'\neq y}\left[\frac{1}{K-1}+\left\langle c'_y,\tilde{W}x+\tilde{b}\right\rangle\right]_+ + \frac{\lambda(K-1)}{2K}|||\tilde{W}|||, \qquad (6)$$

as desired.

We now prove the key linear algebra result. There are other ways (see remarks below) to prove this result. However, it is desirable to establish the equivalence between the more practical encoding and the more interpretable one in an intuitive way. The geometric proof given below accomplishes this by establishing the equivalence through a volume preservation argument.

PROPOSITION

Let $C\in\mathbb{R}^{K\times(K-1)}$ be a simplex coding matrix. Then its columns are orthogonal and have norm $\sqrt{\frac{K}{K-1}}$.

Proof

The key observation (Gantmacher, 1959, vol. 1, p.251) is that

$$V = \sqrt{G}, \tag{7}$$

where $V = V(C)$ is the volume of the parallelipiped spanned by the columns of C, and $G = G(C)$ is the Grammian of C. The Grammian (defined below) extends the notion of volume to objects determined by more vectors than the space they are embedded in has dimensions.

Let $C_{\cdot i}$ denote the i^{th} column of C, and recall that $|||C|||$ denotes the sum of the squared entries of C. Note that $V \leq \|C_{\cdot 1}\| \cdot \cdots \cdot \left\|C_{\cdot(K-1)}\right\|$, which holds with equality if and only if all columns are mutually orthogonal. Further note that

$$\|C_{\cdot 1}\| \cdot \cdots \cdot \left\|C_{\cdot(K-1)}\right\| \leq \left(\sqrt{\frac{|||C|||}{K-1}}\right)^{K-1} = \left(\frac{K}{K-1}\right)^{\frac{K-1}{2}}$$

which holds with equality if and only if $\|C_{\cdot i}\| = \sqrt{\frac{K}{K-1}},\ i = 1, ..., K-1$. Hence, if $G = \left(\frac{K}{K-1}\right)^{K-1}$, it must be that the statement of the proposition is true.

We compute the Grammian. By Gantmacher (1959),

$$G(C) = \sum_{i=1}^{K} \det{}^2\left(C_{-i\cdot}\right), \tag{8}$$

where $C_{-i\cdot}$ is C with the i^{th} row removed. Noting that $C_{-i\cdot}C_{-i\cdot}^T$ is a circulant matrix and using the relevant determinant formula, we find that

$$\begin{aligned}
& \textstyle\sum_{i=1}^{K} \det\left(C_{-i\cdot}C_{-i\cdot}^T\right) \\
= \quad & K \cdot \det \begin{pmatrix} 1 & & -\frac{1}{K-1} \\ & \ddots & \\ -\frac{1}{K-1} & & 1 \end{pmatrix} \\
= \quad & K \cdot \textstyle\prod_{j=0}^{K-2}\left(1 - \frac{1}{K-1}\sum_{m=1}^{K-2}\left(e^{\frac{2\pi ij}{K-1}}\right)^m\right) \\
= \quad & K \cdot \left(1 - \frac{K-2}{K-1}\right) \cdot \textstyle\prod_{j=1}^{K-2}\left(1 - \frac{1}{K-1}\sum_{m=1}^{K-2}\left(e^{\frac{2\pi ij}{K-1}}\right)^m\right) \\
= \quad & K \cdot \left(\frac{1}{K-1}\right) \cdot \textstyle\prod_{j=1}^{K-2}\left(1 + \frac{1}{K-1}\right) \\
= \quad & K \cdot \left(\frac{1}{K-1}\right) \cdot \left(\frac{K}{K-1}\right)^{K-2} \\
= \quad & \left(\frac{K}{K-1}\right)^{K-1},
\end{aligned}$$

which proves the claim.

Note that we have used the orthogonality of the complex exponential basis,

$$\sum_{m=0}^{n-1} e^{\frac{2\pi ijm}{n}} = \begin{cases} n, & j \mod n = 0 \\ 0, & \text{o.w.} \end{cases}.$$

□

The immediate implication of the above argument and proposition is that we may compute MSVM using the equivalent, unconstrained representation of SC-SVM. In Table 1, we display "clock-on-the-wall" computation times. Collected simulations suggest gains of an order of magnitude.

Remark 1 The result of the Proposition holds for a more general simplex matrix $C \in \mathbb{R}^{K\times D}$, $0 < D < K$, having rows of equal norm and maximal equal angle between them.

Remark 2 A different argument of a more algebraic geometry flavor can be given, which suggests the choice of a canonical C. Given K, there exists a simplex coding matrix C such that pairwise coordinate projections (i.e. projections on a plane spanned by two distinct standard basis vectors) yield equidistant points around a circle ("a pie with equal sized slices"). This is trivial for $K = 3$, and geometrically obvious for $K = 4$. Call such a simplex coding matrix a *canonical coding matrix.* From this geometric observation, the orthogonality of the columns readily follows: for any two disctinct columns of C, say $C_{\cdot i}, C_{\cdot j}$, $i \neq j$, we have that

$$\begin{aligned}
\langle C_{\cdot i}, C_{\cdot j}\rangle &= \sum_{t=1}^{K} \cos\left(\frac{t\pi}{K/2}\right)\sin\left(\frac{t\pi}{K/2}\right) \\
&= \frac{1}{2}\sum_{t=1}^{K}\sin\left(\frac{t\pi}{K/4}\right) = 0.
\end{aligned}$$

The length of the columns can be established as in the proof of the Proposition. Furthermore, and somewhat surprisingly, we can go the other way and construct C from the condition on its pairwise coordinate projections (Chan, 2013).

Remark 3 The equivalence of MSVM and SC-SVM immediately generalizes to the infinite-dimensional kernel case. The representer theorem yields that $f_j(x) = b_j + \sum_{i=1}^{n} a_{ij}K(x_i, x)$ for $j = 1, ..., K$ with sum-to-zero constraint. Then (3) holds in the same notation. Letting A denote the matrix with (i, j) entry a_{ij} and K the matrix with (i, j) entry $K(x_i, x_j)$, the penalty equivalence follows from observing that

$$\begin{aligned}
&\text{trace}(A^TKA) = \text{trace}(C\tilde{A}^TK\tilde{A}C^T) \\
&= \text{trace}(C^TC\tilde{A}^TK\tilde{A}) = \frac{K}{K-1}\text{trace}(\tilde{A}^TK\tilde{A}).
\end{aligned}$$

We then get, again, equality of the objective functions up to the tuning parameter.

Table 1. Computation times in seconds. Primal and dual MSVMs are implemented as described in Lee et al. (2004). All simulations are for $K = 3$ balanced categories. Computations were done in Gurobi on a Macbook Pro with a 3.1 GHz Intel Core i7 processor.

FORMULATION	$n = 200$	$n = 1000$	$n = 5000$
PRIMAL MSVM	5.5	81.4	9740.4
DUAL MSVM	0.4	8.2	887.5
SC-SVM	0.1	0.4	7.7

3. Donsker Theorem

By considering MSVM as penalized M-estimator, one can in principle work out its asymptotic distribution. In (2), under simplex encoding, MSVM is phrased as an unconstrained M-estimator, and the asymptotic distribution for the estimated parameters –and thus separating hyperplane– can be obtained using standard empirical process theory. The expression for the covariance matrices presented below are novel and of practical use. To the best of my knowledge, if a practitioner wants to compute the asymptotic covariance matrix of SVM or MSVM –which is essential in order to know where extrapolation is reliable– this article is the only resource displaying worked out expressions with sample analogs.[4]

One readily obtains (Van der Vaart, 2008) a standard central limit theorem result of the form

$$\sqrt{n}\left(\hat{\tilde{\Theta}}_n - \tilde{\Theta}^*\right) \xrightarrow{d} N(0, H^{-1}_{\text{Multi}}\Omega_{\text{Multi}}H^{-1}_{\text{Multi}}), \qquad (9)$$

where $\tilde{\Theta} = (\text{vec}(\tilde{W})^T, \tilde{b})^T$, the information matrix Ω_{Multi} is

$$E\left(\sum_{y'\neq y} c^T_{y'}\mathbf{1}\left\{\left\langle c_{y'}, \tilde{f}\right\rangle \geq -\tilde{a}\right\}\right)\left(\sum_{y'\neq y} c_{y'}\mathbf{1}\left\{\left\langle c_{y'}, \tilde{f}\right\rangle \geq -\tilde{a}\right\}\right) \otimes \left((x^T,1)^T(x^T,1)\right),$$

and the Hessian H_{Multi} is

$$E_y\left[\sum_{y'\neq y}\left(c^T_{y'}c_{y'}\right)p\left(-\left\langle c_{y'}, \tilde{b}\right\rangle - \tilde{a}\right) \otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'}, \tilde{f}\right\rangle = -\tilde{a}, y\right.\right]\right].$$

Both are evaluated at $\tilde{\Theta}^*$, $\tilde{f} = \tilde{f}(x)$, and $\tilde{a} = \frac{1}{K-1}$, and $p = p_{\langle c_{y'}, \tilde{W}x+\tilde{b}\rangle|y}$ is the density of $\left\langle c_{y'}, \tilde{f}\right\rangle$ conditional on y. Derivations are given in the online appendix.

[4] Koo et al. (2008) and Jiang et al. (2008) do not provide expressions with sample analogs.

3.1. Efficient Classifiers

SVM are most commonly used for classification and prediction tasks. Accordingly, the most immediate practical use for an estimate of the variance of the separating hyperplane is the construction of a more accurate classifier.

Consider the One-vs-Rest method, for instance. The One-vs-Rest method fits K hyperplanes, which in the linear case are defined by $(\omega_i, b_i) \in \mathbb{R}^{p+1}$, and categorizes a point by attributing it to the category in which it is the "deepest". That is,

$$\hat{y}_{\text{new}} = \arg\max_k \left\{\hat{\omega}_k^T x_{\text{new}} + \hat{b}_k\right\}.$$

However, studentized distances yield more sensible and reliable classifications by accounting for the comparative uncertainty of the hyperplanes when categorizing a given point. Naturally, a point being "deeper" with respect to a classifying hyperplane –in terms of the length of the line from the point to the hyperplane and normal to the hyperplane– should make one more confident in the classification if it occurs in a section of the space where the hyperplane has lower variance. In sections with high variance, the distance could be much smaller in resamplings of the data. Accordingly, we suggest the following efficient categorization rule

$$\hat{y}^*_{\text{new}} = \arg\max_k \left\{\frac{\hat{\omega}_k^T x_{\text{new}} + \hat{b}_k}{\sqrt{(x^T_{\text{new}}, 1)\Sigma_k(x^T_{\text{new}}, 1)}}\right\},$$

where Σ_k is the asymptotic variance of $(\hat{\omega}_k, \hat{b}_k)$, or a consistent estimate. An analog modification can be applied to make the MSVM procedure more efficient.

4. Efficiency of One-vs-Rest

Explaining the surprisingly competitive performance of the naive One-vs-Rest approach, comparatively to the more sophisticated MSVM approach, is an important open question. The phenomenon is detailed and documented empirically in Rifkin and Klautau (2004) and is well established in the machine learning folklore. However, there are practically no theoretical results in the way of an explanation. In this section, we consider this question from the asymptotic statistics perspective and argue that the competitive performance of One-vs-Rest may be explained by a more efficient use of information.

The idea is to consider the full One-vs-Rest method as a single M-estimator and to artificially impose a sum-to-zero constraint on the decision function. I can use the simplex encoding and obtain the (joint) asymptotic variance of the K separating hyperplanes in the form $H^{-1}_{1\text{vsR}}\Omega_{1\text{vsR}}H^{-1}_{1\text{vsR}}$.

Note that I pick the geometric margin to be $\frac{1}{K-1}$, rather than 1 in the standard form for binary (and thus One-vs-

Rest) SVM. The loss function for One-vs-Rest in simplex encoding is

$$\sum_{k=1}^{K}\left(\mathbf{1}\{y=k\}\cdot\left[\frac{1}{K-1}-\left\langle c_k,\tilde{W}x+\tilde{b}\right\rangle\right]_+ \right. \tag{10}$$
$$\left.+\mathbf{1}\{y\neq k\}\cdot\left[\frac{1}{K-1}+\left\langle c_k,\tilde{W}x+\tilde{b}\right\rangle\right]_+\right)$$

which is minimized in $\tilde{W}$ and $\tilde{b}$. The first summand penalizes classification for which the point x is not sufficiently far from the hyperplane within the true category. This is where we speak of using the information from a point's "own" category. The second summand penalizes classifications for which the point x is not sufficiently far from the hyperplane away from the wrong category. This is where we speak of using the information from "other" categories.

The sum-to-zero constraint is added for analytical reasons; we need it to make the covariance matrices comparable. It will be apparent that the analytical conclusion is robust to this modification.

The information matrix $\Omega_{1\text{vsR}}$ is

$$E\left(c_y^T\mathbf{1}\{\tilde{a}-\left\langle c_y,\tilde{f}\right\rangle\geq 0\}\right)\left(c_y\mathbf{1}\{\tilde{a}-\left\langle c_y,\tilde{f}\right\rangle\geq 0\}\right)$$
$$\otimes(x^T,1)^T(x^T,1)$$
$$-2E(c_y^T\mathbf{1}\{\tilde{a}-\left\langle c_y,\tilde{f}\right\rangle\geq 0\})(\sum_{y'\neq y}c_{y'}\mathbf{1}\{\tilde{a}+\left\langle c_{k'},\tilde{f}\right\rangle\geq 0\})$$
$$\otimes(x^T,1)^T(x^T,1)$$
$$+E(\sum_{y'\neq y}c_{y'}^T\mathbf{1}\{\tilde{a}+\left\langle c_k,\tilde{f}\right\rangle\geq 0\})(\sum_{y'\neq y}c_{y'}\mathbf{1}\{\tilde{a}+\left\langle c_{k'},\tilde{f}\right\rangle\geq 0\})$$
$$\otimes(x^T,1)^T(x^T,1),$$

and the Hessian $H_{1\text{vsR}}$ is

$$E_y\left[(c_y^Tc_y)\left(p\left(\left\langle c_y,\tilde{b}\right\rangle-\tilde{a}\right)\right)\right.$$
$$\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_k,\tilde{f}\right\rangle=\tilde{a},y\right.\right]\Big]$$
$$+E_y\left[\sum_{y'\neq y}(c_{y'}^Tc_{y'})\left(p\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\tilde{a}\right)\right)\right.$$
$$\left.\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'},\tilde{f}\right\rangle=-\tilde{a},y\right.\right]\right].$$

We get instructive comparisons. First of all, $H_{\text{Multi}} < H_{1\text{vsR}}$. That is, the one-vs-rest problem has more "curvature" than the MSVM. Indeed, it is clear from inspection[5] that this comes from the one-vs-rest procedure using information from the "own category", while MSVM doesn't as it only uses information with respect to "other" categories.

It is clear from the comparison of the loss functions (2) and (10), corresponding to SC-SVM and the simplex encoding of One-vs-Rest, respectively, that both penalize for an observation that falls within the half-space assigned to an "other" category, but only One-vs-Rest rewards for points falling within their true, "own" category. It was not clear, however, if the rewarding a point for being in its "own" category is still informative and not redundant when it is already penalized if it is in any "other" category. However, in spite of imposing an additional constraint on the solution space of the One-vs-Rest problem, we do find from inspection of the Hessian that the additional information from rewarding classification of points within their "own" category is informative and not redundant. Although this was not obvious a priori, it is revealed by the statistical asymptotic analysis.

Furthermore, in the special case of a *separable* data generating process (DGP), that is in the case in which $\mathbf{1}\{\tilde{a}-\left\langle c_y,\tilde{f}\right\rangle\geq 0\}=0$ a.s., we get that $\Omega_{\text{Multi}}=\Omega_{1\text{vsR}}$ and both procedures have the same target hyperplane. Therefore, One-vs-Rest is strictly more statistically efficient than multicategory when the DGP is separable. In this specific case, this translates into smaller expected prediction error. We have displayed a case, that of perfect seprarability, where One-vs-Rest (with simplex encoding) provably dominates MSVM.

In non-separable cases, this dominance may not hold. In fact, we expect MSVM to outperform One-vs-Rest in some cases due to the more efficient gathering, by joint optimization, of the information from the "other" categories.

5. Discussion and Conclusion

We established rigorously, and with a proof conveying geometric intuition, the equivalence of MSVM and SC-SVM. This provides a formulation of the optimization problem for computing MSVM which is relieved of the sum-to-zero constraint that bogged down computations in the implementations as suggested in Lee et al. (2004). Our hope is that availablity of faster computations for MSVM will encourage applied researchers and analysts to employ MSVM in multicategory classification tasks. We gave the first central limit theorem for MSVM, along with an asymptotic covariance formula having a sample analog, which is a new result even for binary SVM. The variance formula allows for the construction of studentized decision functions for One-vs-Rest procedures, improving their accuracy and statistical efficiency. These make for more reliable classification, especially for extrapolation. We gave an analytical characterization of the surprisingly good performance of the

[5] Note the addition of a $y=y'$ positive summand.

One-vs-Rest procedure, comparatively to MSVM, using the asymptotic distribution of estimators. We hope this line of study fosters further research.

Acknowledgements

I would like to thank Lorenzo Rosasco for introducing me to the material studied in this article and for his support throughout this project. Different angles for studying SVM methods as M-estimators arose in stimulating conversation with Isaiah Andrews. Jann Spiess read an early draft of the article and made helpful comments. Jules Marchand-Gagnon collaborated on a cousin project and contributed insights which this article bears the mark of.

Appendix: Construction of the Simplex Encoding Matrix C

It is straightforward to build a function that takes K and outputs the simplex encoding matrix C. We give intuitive pseudocode for the general K case. A code file is available in the online appendix.

The construction relies on mapping vectors in spherical coordinates to their representation in Cartesian coordinates. The function $\mathrm{StoC} : \mathbb{R}^{K-2} \to \mathbb{R}^{K-1}$ does just that.

$$
\begin{aligned}
&\mathrm{StoC} \leftarrow \mathrm{function}(\phi_1, \phi_2, ..., \phi_{K-2})\{ \\
&\quad v_1 = \cos(\phi_1) \\
&\quad v_2 = \sin(\phi_1)\cos(\phi_2) \\
&\quad v_3 = \sin(\phi_1)\sin(\phi_2)\cos(\phi_3) \\
&\quad \vdots \\
&\quad v_{K-2} = \sin(\phi_1)\cdots\sin(\phi_{K-3})\cos(\phi_{K-2}) \\
&\quad v_{K-1} = \sin(\phi_1)\cdots\sin(\phi_{K-3})\sin(\phi_{K-2}) \\
&\quad \\
&\quad v = (v_1, ..., v_{K-1}) \\
&\quad \\
&\quad v \\
&\}
\end{aligned}
$$

Using the StoC function, it is now easy to construct the simplex encoding matrix. This can be done with the function $\mathrm{C} : K \mapsto \mathbb{R}^{K\times(K-1)}$, which we now describe.

$$
\begin{aligned}
&\mathrm{C} \leftarrow \mathrm{function}(K)\{ \\
&\quad C_{1,\cdot} = \mathrm{StoC}(0, 0, ..., 0) \\
&\quad C_{2,\cdot} = \mathrm{StoC}\left(\mathrm{acos}\left(\frac{-1}{K-1}\right), 0, ..., 0\right) \\
&\quad C_{3,\cdot} = \mathrm{StoC}\left(\mathrm{acos}\left(\frac{-1}{K-1}\right), \mathrm{acos}\left(\frac{-1}{K-2}\right), 0, ..., 0\right) \\
&\quad \vdots \\
&\quad C_{K-1,\cdot} = \mathrm{StoC}\left(\mathrm{acos}\left(\frac{-1}{K-1}\right), \mathrm{acos}\left(\frac{-1}{K-2}\right),\right. \\
&\quad \left. ..., \mathrm{acos}\left(\frac{-1}{3}\right), \mathrm{acos}\left(\frac{-1}{2}\right)\right) \\
&\quad C_{K,\cdot} = \mathrm{StoC}\left(\mathrm{acos}\left(\frac{-1}{K-1}\right), \mathrm{acos}\left(\frac{-1}{K-2}\right),\right. \\
&\quad \left. ..., \mathrm{acos}\left(\frac{-1}{3}\right), 2\cdot\mathrm{acos}\left(\frac{-1}{2}\right)\right) \\
&\quad \\
&\quad C \\
&\}
\end{aligned}
$$

References

A. J. Chan, *Grobner Bases over fields with Valuations and Tropical Curves by Coordinate Projections*, PhD Thesis, University of Warwick, 2013.

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

Dogan, Urn, Tobias Glasmachers, and Christian Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research* 17, no. 45 (2016): 1-32.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, Berlin: Springer series in statistics, 2009.

F. R. Gantmacher. *The Theory of Matrices*, Chelsea, New York, 1959.

B. Jiang, X. Zhang, and T. Cai, Estimating the Confidence Interval for Prediction Errors of Support Vector Machine Classifiers. *Journal of Machine Learning Research*, 9: 521–540, 2008.

J. Koo, Y. Lee, Y. Kim, and C. Park, A Bahadur representation of the linear support vector machine. *Journal of*

Machine Learning Research, 9: 1343–1368, 2008.

Kumar, Deepak, and Manoj Thakur. All-in-one multicategory least squares nonparallel hyperplanes support vector machine. *Pattern Recognition Letters* (2017).

Y. Lee, L. Yin, 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*, 99:67-81, 2004.

B. B. Li, A. Artemiou and L. Li, Principal Support Vector Machines for Linear and Nonlinear Sufficient Dimension Reduction. *The Annals of Statistics*, 39: 3182-3210, 2011.

Lopez, Julio, Sebastin Maldonado, and Miguel Carrasco. A novel multi-class SVM model using second-order cone constraints. *Applied Intelligence* 44, no. 2 (2016): 457-469.

Ma, Yunqian, and Guodong Guo, eds. *Support vector machines applications*. Switzerland: Springer, 2014.

Y. Mroueh, T. Poggio, L. Rosasco, J.-J. E. Slotine, Multiclass Learning with Simplex Coding. *Advances in Neural Information Processing Systems, NIPS* 2012.

B. A. Pires, C. Szepesvari, M. Ghavamzadeh, Cost-sensitive multiclass classification risk bounds. *Proceedings of The 30th International Conference on Machine Learning*. 2013.

R. Rifkin, and A. Klautau, In defense of one-vs-all classification. *The Journal of Machine Learning Research* 5, 101-141, 2004.

A. J. Smola and B. Schlkopf. *Learning with kernels.* GMD-Forschungszentrum Informationstechnik, 1998.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Sun, Hui, Bruce A. Craig, and Lingsong Zhang. Angle-based multicategory distance-weighted SVM. *The Journal of Machine Learning Research* 18, no. 1 (2017): 2981-3001.

A. W. Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

V. N. Vapnik. *Statistical learning theory*. Vol. 1. New York: Wiley, 1998.

J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *ESANN*, vol. 99, pp. 219-224. 1999.