Equivalence of Multicategory SVM and Simplex Cone SVM

Guillaume A. Pouliot

November 4, 2015

Abstract

I show that the multicategory SVM (MSVM) of Lee, Yin and Wahba (2004) is equivalent to the Simplex Cone SVM of Mroueh, Poggio, Rosasco and Slotine (2012). This allows me to provide finite-dimensional kernel asymptotics for MSVM and to partially answer the open question relating to the very competitive performance of the seemingly more naive One-vs-Rest method against MSVM. In particular, I give a Donsker theorem for MSVM along with an asymptotic covariance formula having a sample analog, and I display a case in which the One-vs-Rest procedure is strictly more accurate than MSVM, in expectation, for out-of-sample prediction. Furthermore, I use the obtained covariance formula to develop a variance-weighted classification rule which improves the traditional One-vs-Rest approach.

1 Introduction

Support vector machines (SVM) is a well established algorithm for classification with two categories (Vapnik (1998), Smola and Schölkopf (1998), Steinwart and Christmann (2008), Friedman et al. (2009)). The method, detailed in the appendix, 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 (see appendix).

Given the success of SVM at binary classification, many attempts have been made at extending the methodology to accommodate classification with $K > 2$ categories. Each of these 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 (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 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. The second approach consists in generalizing the standard SVM to develop a single machine which implements multicategory classification solving a single, joint optimization problem. We call such algorithms multicategory support vector machines (MSVM). Many such algorithms have been suggested (Weston and Watkins (1999), Crammer and Singer (2002), Lee et al. (2004)). Intuition would suggest that joint optimization would make for a more statistically efficient procedure, and for superior out-of-sample prediction performance.

Two interesting and important observations emerge from a study of these multicategory classification methods. First, in a quite counterintuitive turn of events, it is 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 MSVM algorithms. This phenomenon is widely acknowledged (Rifkin and Klautau (2004)) but very little theory has been put forth to explain it. Second, in a One-vs-Rest implementation, it is apparent that assessing the "confidence" of classification into a given category (against the the category formed by the union of the remaining categories) using the distance of the test point to the separating hyperplane is inefficient, in the sense that relevant information is foregone. Very intuitively, having a point that is far from a hyperplane, but where the hyperplane is extremely variable, ought to make for a less "confident" classification than having a point which is not as deep within its predicted category, but where the separating hyperplane is very stable.

Amongst the available MSVM agorithms, I focus on the one by Lee et al. (2004). It is a most natural generalization of the standard, binary SVM and it 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. Naturally, it encompasses standard SVM as a special case.

In their article, Lee, Yin, and Wahba (2004) convincingly establish the good properties of their MSVM algorithm. However, the algorithm has not been widely used in application, nor has it been studied from a statistical perspective the way SVM has been. Indeed, 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. The reason for this is that 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 difficult.

The issue is that Lee et al. (2004) encode categories with K-tuples, to give a natural indexing. However, the output space is $K-1$ dimensional, whence the additional linear restriction (see section 2).

This motivates the use of a different encoding of the categories. A desirable encoding does away with the extraneous restriction while remaining interpretable and has an intuitive relationship to the original encoding. Mroueh et al. (2012) propose an interpretable, $K-1$ dimensional encoding and suggest that it yields a loss function which is similar to that of the MSVM of Lee et al. (2004). In this article, I show that both loss functions are in fact exactly equivalent, and that the equivalence of the

two encodings can be intuited through a notion of maximum volume.

With this equivalent reformulation of the MSVM of Lee, Yin, and Wahba (2004), I can overcome the computational and analytical problems encumbering MSVM in its standard encoding. In particular, I give a Donsker theorem for MSVM.

Furthermore, this result in turn makes available analytical tools to tackle the two aforementioned open problems: first, I suggest a better assessment of "confidence" for One-vs-Rest schemes. I give closed form formulae allowing for the computations of a more efficient notion of distance. The idea is to normalize the distance to the hyperplane by the variance of the hyperplane.

The variance formulae I provide are also valid for the special case of binary SVM. To the best of my knowledge, this article is the first to explicitly and correctly[1] give asymptotic covariance formulae for SVM (even for $K = 2$) which have sample analogs. In certain applications, these covariance formulae may nevertheless be challenging to estimate. In separate work (Pouliot (2015)) I give a fast bootstrap algorithm circumventing the estimation difficulties.

Second, I get an analytical characterization of the superior performance of One-vs-Rest. I exhibit a case in which the One-vs-Rest procedure is strictly more statistically efficient than MSVM, thus providing some theoretical explanation for the so far flummoxingly good comparative performance of One-vs-Rest.

The remainder of the paper is divided as follows. Section 2 details the equivalence of MSVM ans SC-SVM. Section 3 explores the implications for statistical asymptotic analysis, and suggests a studentized notion of distance for multicategory classification. Section 4 gives an analytical characterization of the comparative performance of One-vs-Rest with MSVM. Section 5 concludes and discusses related research. The appendix includes background material on SVM as well as deferred mathematical derivations.

2 Equivalence

The multicategory SVM (MSVM) of Lee et al. (2004) is 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 conceptually attractive and is relieved of the linear constraint on the vector argument. However, I believe the simplex encoding is not more widely used because practitioners do not know what standard encoding it corresponds to. This article addresses precisely that issue, making it of practical interest for analysts and researchers using MSVM methods.

With K categories, the 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

[1] (Li et al. (2011)) is the only other paper providing them, albeit incorrectly. The expectation in the third display of p. 17 should be conditional on y. They provide no derivations.

optimization problems that are differently formulated and implemented.

For their multicategory SVM (MSVM), Lee et al. (2004) encode a 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,

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

The loss function they suggest is then based on the *difference* between the decision function and the encoded y. Specifically, in the case of finite-dimensional feature maps, they suggest minimizing

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

where $|||W||| = \text{trace}(W^T W)$, 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. Likewise in the finite-dimensional case, the penalized minimization problem entailed by their loss function is

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

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) $c_i^T c_j = -\frac{1}{K-1}$ 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 $\{x \in \mathbb{R}^K : 1_K^T x = 0\}$.

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 solutions to (1) and (2), respectively, are $\tilde{C}(Wx + b)$ and $C(\tilde{W}x + \tilde{b})$ then solutions to (2) and (1), respectively? I show that this is in fact the case. That is, *both problems are exactly equivalent.*

I now show the problems are equivalent. The equivalence of the loss functions is trivial. Indeed, it is immediate that

$$\sum_{y' \neq y} \left[f_{y'}(x) + \frac{1}{K-1} \right]_+ = \sum_{y' \neq y_i} \left[\left(C\tilde{f}(x) \right)_{y'} + \frac{1}{K-1} \right]_+ = \sum_{y' \neq y_i} \left[\left\langle c_{y'}, \tilde{f}(x) \right\rangle + \frac{1}{K-1} \right]_+, \quad (3)$$

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

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

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 only nontrivial observation of this exercise, which is that $C^T C$ is the diagonal matrix $\frac{K}{K-1} I_{K-1}$. It then immediately follows that

$$\frac{K-1}{K} \text{trace} \left(\tilde{\mathrm{W}}^{\mathrm{T}} \tilde{\mathrm{W}} \right) = \frac{K-1}{K} \text{trace} \left(\mathrm{W}^{\mathrm{T}} \mathrm{C}^{\mathrm{T}} \mathrm{C} \mathrm{W} \right) = \text{trace} \left(\mathrm{W}^{\mathrm{T}} \mathrm{W} \right). \quad (5)$$

In conclusion, we have

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

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

as desired.

We now prove the key linear algebra result. There are other ways (see remarks below) to prove this result. In particular, a streamlined functional analysis proof can be given. However, it is desirable to establish the equivalence between the more practical encoding and the more interpretable one in an intuitive way.[2] 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 [5], vol. 1, p.251) is that

$$V = \sqrt{G}, \quad (7)$$

[2] I would like to thank Lorenzo Rosasco for this motivation.

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

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

which proves the claim.

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

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

□

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

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

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 yieds 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

$$\text{trace}(A^T K A) = \text{trace}(C\tilde{A}^T K \tilde{A} C^T) = \text{trace}(C^T C \tilde{A}^T K \tilde{A}) = \frac{K}{K-1} \text{trace}(\tilde{A}^T K \tilde{A}).$$

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

3 Donsker Theorem

By considering MSVM as penalized M-estimator, one can in principle work out its asymptotic distribution. In (2), under simplex enconding, MSVM is phrased an unconstrained M-estimator, and the asymptotic distribution for the estimated coefficients can be obtained using standard empirical process theory (Van der Vaart, 2000). 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, this article is the only ressource displaying worked out expressions with sample analogs.[3]

One readily obtains a standard central limit theorem result of the form

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

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

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

and the Hessian H_{Multi} is

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

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 deferred to the mathematical appendix.

SVM are most commonly used for classification and prediction tasks. Accordingly, the most immediate practical use for standard errors for the separating hyperplane is to allow for the construction of a better classification function.

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 the category in which it is the "deepest". That is,

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

Pouliot (2015a) argues that studentized distances yield more sensible and reliable classifications by accounting for the comparative uncertainty of the hyperplanes when categorizing a given point. Accordingly, Pouliot (2015a) suggests the categorization rule

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

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 MSVM procedures more efficient.

[3] Koo et al. (2008) and Jiang et al. (2008) do not work them out.

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 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, I tackle this problem by comparing asymptotic covariance matrices, and I display explicitly how one loss function using more information impacts statistical efficiency.

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 $H_{1\text{vsR}}^{-1}\Omega_{1\text{vsR}}H_{1\text{vsR}}^{-1}$.

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

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

$i=1,...,K$, 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 purely for analytical reasons -to make the covariance matrices comparable- and it will be apparent that the analytical conclusion is entirely robust to this aiding modification.

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

$$E\left(c_y^T\cdot\mathbf{1}\{\tilde{a}-\left\langle c_y,\tilde{f}\right\rangle\geq 0\}\right)\left(c_y\cdot\mathbf{1}\{\tilde{a}-\left\langle c_y,\tilde{f}\right\rangle\geq 0\}\right)\otimes(x^T,1)^T(x^T,1)$$

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

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

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

$$E_y \sum_{k=1}^{K} (c_k^T c_k) \left(\mathbf{1}\{y = k\} \cdot p\left(\left\langle c_k, \tilde{b} \right\rangle - \tilde{a} \right) \right) \otimes E\left[(x^T, 1)^T (x^T, 1) \left| \left\langle c_k, \tilde{f} \right\rangle = \tilde{a}, y \right. \right]$$

$$+E_y \sum_{k=1}^{K} (c_k^T c_k) \left(\mathbf{1}\{y \neq k\} \cdot p\left(- \left\langle c_k, \tilde{b} \right\rangle - \tilde{a} \right) \right) \otimes E\left[(x^T, 1)^T (x^T, 1) \left| \left\langle c_k, \tilde{f} \right\rangle = -\tilde{a}, y \right. \right].$$

We get instructive comparisons. For instance, $H_{\text{Multi}} < H_{\text{1vsR}}$. That is, the one-vs-rest problem has more "curvature" than the MSVM. This does not come from the artificial constraint. Indeed, it is clear from inspection 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.

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 $\Omega_{\text{Multi}} = \Omega_{\text{1vsR}}$ and both procedures have the same target hyperplane. That is, *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.*

5. Discussion and conclusion

I established rigorously, and with a proof conveying geometric intuition, the equivalence of MSVM and SC-SVM. I have argued that the unconstrained formulation of the MSVM problem thus obtained can be a useful tool in the analytical study of multicategory classification schemes. I 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. These standard errors allow for the construction of studentized decision functions for MSVM and One-vs-Rest procedures. These make for more reliable classification, especially for extrapolation (Pouliot (2015a)). In separate work (Pouliot (2015b)), I develop a fast bootstrap procedure which can be used when computation of the closed form covariance formulae is problematic. I gave an analytical characterization of the surprisingly good performance of the One-vs-Rest procedure, comparatively to MSVM methods, using the asymptotic distribution of the estimators. I hope this line of inquiry will foster further research.

Appendix

A1 Support Vector Machines

The SVM story begins with supervised binary classification. The algorithm is best understood when built incrementally: first for the separable case, then for the non-separable case, and finally rephrased in its dual formulation to obtain a sparser set of constraints.

The classification task undertaken with support vector machines can be described as follows. We are given a training set $\{(x_1, y_1), ..., (x_n, y_n)\}$, where $(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}\ \forall\ i$, and we are trying to find a hyperplane in $\mathbb{R}^p$ that separates the observations with $y_i = -1$ from those with $y_i = 1$. Furthermore, we want a hyperplane that is optimal in that it maximizes the minimum distance between points and the hyperplane (the motivation for this being that, intuitively, it should make for good out-of-sample performance). If the hyperplane is given by $\omega^T x + b = 0$. You can check that ω must be the normal to the hyperplane. If we require that $\omega^T \omega = 1$, then $\omega^T x$ is the length of the projection of x onto w, and b is the additive inverse of the distance from the origin to the hyperplane (along ω). The sum of the two thus gives the distance from x to the hyperplane, which we call the **geometric margin**[4],

$$f(x) = \omega^T x + b.$$

Note that if x is on the correct side of the classifying hyperplane, we get $f(x) > 0$ and if x is on the wrong side of the hyperplane, we get $f(x) < 0$. Thus $\text{sign}(f(x))$ is our classification rule. Furthermore, $y_i(\omega^T x_i + b)$ will always be positive for correctly classified points.

Suppose, to begin with, that we have a data set which we know is separable by such a hyperplane. This is the **separable case**. Then our optimization problem can be written as

$$\max_{\gamma,\omega,\beta} \gamma$$

subject to

$$\begin{aligned} y_i(\omega^T x + b) &\geq \gamma,\ i = 1, ..., n \\ ||\omega|| &= 1. \end{aligned}$$

This immediately expresses the optimization we wish to carry out, but the normalizing constraint $||\omega|| = 1$ is non-convex (in particular, we shouldn't plug the optimization problem as such in a "standard" optimization software).

We can instead maximize the minimum **functional margin** $\hat{\gamma} = \omega^T x + b$ where we do *not* constrain ω to have norm one. The magnitude of the functional margin is meaningless, but we can normalize it in the objective function to get the geometric margin back, and obtain the reformulated optimization problem

$$\max_{\gamma,\omega,b} \frac{\hat{\gamma}}{||\omega||}$$

subject to

[4]Note that the minimal geometric margin, for a given data set, is also referred to as the geometric margin.

$$y_i(\omega^T x_i + b) \geq \hat{\gamma},\ i = 1, ..., n.$$

This gets us rid of the non-convex constraint, but leaves us with a non-convex objective function $\frac{\hat{\gamma}}{||\omega||}$. A key observation allows us to play a little trick and circumvent this issue. Since the units of the functional margins are arbitrary, we can fix $\hat{\gamma} = 1$, and minimize $||\omega||$ instead of maximizing $\frac{1}{||\omega||}$, thus yielding a well-behaved convex optimization problem[5]

$$\min_{\omega,b} \frac{1}{2}||\omega||$$

subject to

$$y_i(\omega^T x_i + b) \geq 1,\ i = 1, ..., n.$$

The dual of the problem will allow for an even simpler (and computationally easy) constraint, and will inform the design of our algorithm for computing the solution.

The dual problem is

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

subject to

$$\begin{aligned} \alpha_i &\geq 0,\ i = 1, ..., n \\ \sum_{i=1}^{n} \alpha_i y_i &= 0. \end{aligned}$$

Three key points remain to be elucidated: the kernel, the feature map, and how to deal with the non-separable case.

Note that in the dual in the dual formulation, the x's come in only through their inner product. It is thus sufficient to only know the inner products $\langle x_i, x_j \rangle\ ,\ \forall\ i, j$.

As in the case of linear regression, we often want to allow for more flexibility for the fit than is allowed by the **attributes** (the x's) and would like instead to consider **features** $\phi(x)$ (e.g. products and other nonlinear functions of variables). We call ϕ the **feature map**. We can then simply proceed as detailed above replacing x with $\phi(x)$ everywhere.

Once again, only the inner products are necessary for the solution. We thus define the kernel $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. In many cases of interest, the left-hand side is cheaper to compute directly (with a closed form expression for K) than the right-hand side. As in the linear case, we can replace the

[5]The addition of the $\frac{1}{2}$ coefficient to the objective function is to make the Lagrangian of the dual problem prettier.

inner products $\phi(x_i)^T\phi(x_j)$ by the computationally cheaper $K(x_i, x_j)$ everywhere in the optimization problem. That is called the **kernel trick**.

We have presented the kernel trick as afforded by the dual representation, it is however a characteristic of the solution which can be obtained directly through the **representer theorem**.

Furthermore, all ϕ give a kernel, but all positive definite kernels come from some ϕ. In fact, it will often be convenient to work directly with a kernel (as an expression of closeness/angle between two attribute x_i and x_j) without an explicit representation for the corresponding ϕ.

In general, even with a fairly flexible feature map ϕ, the data $\{(x_1, y_1), ..., (x_n, y_n)\}$ may not be perfectly separable[6], i.e. there may not exist a hyperplane separating the observations with $y_i = -1$ from those with $y_i = 1$. This is accommodated, in what is called the **non-separable** case, by introducing slack variables on the functional margin, and regularizing those with an ℓ-1 norm as follows

$$\min_{\xi,\omega,b} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to

$$\begin{aligned} y_i(\omega^T\phi(x_i) + b) &\geq 1 - \xi_i,\ i = 1, ..., n \\ \xi_i &\geq 0,\ i = 1, ..., n. \end{aligned}$$

The dual problem is

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

subject to

$$\begin{aligned} 0 &\leq \alpha_i \leq C,\ i = 1, ..., n \\ \sum_{i=1}^{n}\alpha_i y_i &= 0. \end{aligned}$$

Note that the only difference in the dual problem from adding the regularization term in the primal problem is the upper bound C on the Lagrange multipliers.

[6]Naturally, for a flexible enough ϕ, there will exist such a hyperplane. However, it may have very poor out-of-sample performance.

A2 Mathematical Appendix

I give the closed forms for H_{Multi} and Ω_{Multi} that have estimatable population analogues. Ω_{Multi} is $E\left[\left(\frac{\partial}{\partial\tilde{\Theta}}m_{\text{Multi}}(\tilde{\Theta};x,y)\right)\left(\frac{\partial}{\partial\tilde{\Theta}}m_{\text{Multi}}(\tilde{\Theta};x,y)\right)^T\right]$

$$
\begin{aligned}
&= E\left(\sum_{y'\neq y}\left(c_{y'}\otimes(x^T,1)\right)^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\left(\sum_{y'\neq y}\left(c_{y'}\otimes(x^T,1)\right)^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)^T\\
&= E\left(\left(\sum_{y'\neq y}c_{y'}^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\otimes(x^T,1)^T\right)\left(\left(\sum_{y'\neq y}c_{y'}^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\otimes(x^T,1)^T\right)^T\\
&= E\left(\left(\sum_{y'\neq y}c_{y'}^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\otimes(x^T,1)^T\right)\left(\left(\sum_{y'\neq y}c_{y'}\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\otimes(x^T,1)\right)\\
&= E\left(\sum_{y'\neq y}c_{y'}^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\left(\sum_{y'\neq y}c_{y'}\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}\right)\otimes\left((x^T,1)^T(x^T,1)\right),
\end{aligned}
$$

evaluated at $\tilde{\Theta}^*$.

H_{Multi} is $\frac{\partial^2}{\partial\tilde{\Theta}^2}Em_{\text{Multi}}(\tilde{\Theta};x,y)$

$$
\begin{aligned}
&= \tfrac{\delta}{\delta\theta}E\textstyle\sum_{y'\neq y}\tfrac{\delta}{\delta\theta}\left[\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle+\tfrac{1}{K-1}\right]_+\\
&= E_y\left[\tfrac{\delta}{\delta\theta}E\left[\textstyle\sum_{y'\neq y}\left(c_{y'}\otimes(x^T,1)\right)^T\mathbf{1}\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\geq-\tfrac{1}{K-1}\right\}|y\right]\right]\\
&= E_y\left[E\left[\textstyle\sum_{y'\neq y}\left(c_{y'}\otimes(x^T,1)\right)^T\left(c_{y'}\otimes(x^T,1)\right)\delta\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\tfrac{1}{K-1}\right\}|y\right]\right]\\
&= E_y\left[E_{\langle c_{y'},\tilde{W}x+\tilde{b}\rangle}\left[\textstyle\sum_{y'\neq y}\delta\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\tfrac{1}{K-1}\right\}E\left[\left(c_{y'}\otimes(x^T,1)\right)^T\left(c_{y'}\otimes(x^T,1)\right)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle\right.\right]|y\right]\right]\\
&= E_y\left[\int\textstyle\sum_{y'\neq y}\delta\left\{\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\tfrac{1}{K-1}\right\}E\left[\left(c_{y'}\otimes(x^T,1)\right)^T\left(c_{y'}\otimes(x^T,1)\right)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle,y\right.\right]fd(\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle)\right]\\
&= E_y\left[\textstyle\sum_{y'\neq y}E\left[\left(c_{y'}\otimes(x^T,1)\right)^T\left(c_{y'}\otimes(x^T,1)\right)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\tfrac{1}{K-1},y\right.\right]f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\tfrac{1}{K-1}\right)\right]\\
&= E_y\left[\textstyle\sum_{y'\neq y}E\left[\left(c_{y'}^Tc_{y'}\right)\otimes\left((x^T,1)^T(x^T,1)\right)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\tfrac{1}{K-1},y\right.\right]f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\tfrac{1}{K-1}\right)\right]\\
&= E_y\left[\textstyle\sum_{y'\neq y}\left(\left(c_{y'}^Tc_{y'}\right)f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\tfrac{1}{K-1}\right)\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\tfrac{1}{K-1},y\right.\right]\right)\right],
\end{aligned}
$$

evaluated at $\tilde{\Theta}^*$, where $f=f_{\langle c_{y'},\tilde{W}x+\tilde{b}\rangle|y}$ is the density of $\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle$ conditional on y.

To obtain the asymptotic distribution for the hyperplanes in the original encoding of Lee et al. (2004), on needs the covariance matrix $CH_{\text{Multi}}^{-1}\Omega_{\text{Multi}}H_{\text{Multi}}^{-1}C^T$, which is readily obtained after observing that

$$
\begin{aligned}
& C\left(E_y\left[\sum_{y'\neq y}\left(c_{y'}^T c_{y'} f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\frac{1}{K-1}\right)\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\frac{1}{K-1},y\right.\right]\right)\right]\right)^{-1} \\
= \; & C\left(E_y\left[\sum_{y'\neq y}\left(c_{y'}^T c_{y'} C^T\left(CC^T\right)^{-1} f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\frac{1}{K-1}\right)\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\frac{1}{K-1},y\right.\right]\right)\right]\right) \\
= \; & \left(E_y\left[\sum_{y'\neq y}\left(c_{y'}^T y^T\left(CC^T\right)^{-1} f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\frac{1}{K-1}\right)\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\frac{1}{K-1},y\right.\right]\right)\right]\right)^{-1} \\
= \; & \left(E_y\left[\sum_{y'\neq y}\left(c_{y'}^T e_y f\left(-\left\langle c_{y'},\tilde{b}\right\rangle-\frac{1}{K-1}\right)\otimes E\left[(x^T,1)^T(x^T,1)\left|\left\langle c_{y'},\tilde{W}x+\tilde{b}\right\rangle=-\frac{1}{K-1},y\right.\right]\right)\right]\right)^{-1},
\end{aligned}
$$

where e_i is the i^{th} standard basis vector.

References

A. J. Chan, *Gröbner 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.

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. J*ournal of Machine Learning Research*, 9: 1343–1368, 2008.

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.

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.

G. A. Pouliot. *Probabilistic SVM and Fast MCMC Sampling of the Bootstrap Distribution.* Working document. 2015a.

G. A. Pouliot. *Simplex Monte Carlo and Fast MCMC Sampling of the Bootstrap Distribution.* Working document. 2015b.

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. Schölkopf. *Learning with kernels.* GMD-Forschungszentrum Informationstechnik, 1998.

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

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.