Journal of Machine Learning Research ? (?) Submitted 7/15; Published ?/??

A Unified View on Multi-class Support Vector Classification Supplement

Ürün Doğan UDOGAN@MICROSOFT.COM
Microsoft Research

Tobias Glasmachers TOBIAS.GLASMACHERS@INI.RUB.DE
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany

Christian Igel IGEL@DIKU.DK
Department of Computer Science
University of Copenhagen, Denmark

Editor: Ingo Steinwart

A. Aggregation Operators As Linear Programs

Aggregation operators, which combine the d margin violations into a single cost value, can be understood as computing the value of the linear program

$$\begin{aligned} \Delta(v, y) \quad = \quad & \min_{\xi} \quad \sum_{r \in R_y} \xi_r \\ & \text{s.t.} \quad \forall p \in P_y : \ \ \xi_{s_y(p)} \geq v_p(f(x), y) \end{aligned}$$

for the variables $\xi = (\xi_r)_{r \in R_y}$, where $P_y \subset Y$, R_y is an index set, and $s_y : P_y \to R_y$ is surjective. The set P_y lists all margin violations that enter the loss, R_y lists the slack variables, and s_y assigns slack variables to margin components, depending on the particular loss in use. Table 1 lists the configurations of the linear programs corresponding to the different aggregation operators.

	linear program definition		
aggregation operator	P_y	R_y	s_y
Δ^{self}	$\{y\}$	$\{*\}$	$p \mapsto *$
$\Delta^{\text{o-max}}$	$Y \setminus \{y\}$	$\{*\}$	$p \mapsto *$
$\Delta^{\text{t-max}}$	Y	$\{*\}$	$p \mapsto *$
$\Delta^{\text{t-sum}}$	Y	Y	id
$\Delta^{\text{o-sum}}$	$Y \setminus \{y\}$	$Y \setminus \{y\}$	id

Table 1: Aggregation operators and the corresponding linear programs, expressed in terms of the sets P_y and R_y, and the assignment $s_y : P_y \to R_y$, for each $y \in Y$.

As for the margin function definition based on the sparse coefficients $\nu_{y,p,m}$, the true underlying degrees of freedom for aggregation operators are far more restricted than it

seems, in particular if classes are treated symmetrically. For symmetry reasons, the sets P_y can take only the three values $\{y\}$, Y, and $Y \setminus \{y\}$, since all classes $c \neq y$ are to be treated the same way. The same argument implies that s_y either has to be injective or constant, restricted to the atomic invariant subsets $\{y\}$ and $Y \setminus \{y\}$. This again leaves only few choices for R_y under the restriction that s_y is surjective.

The hinge loss $L^{\text{hinge}}(\mu) = \max\{0, 1-\mu\}$ can also be expressed as a linear program, namely

$$\begin{aligned} L^{\text{hinge}}(\mu) \quad = \quad & \min_{u} \quad u \\ & \text{s.t.} \quad u \geq 1 - \mu \\ & \qquad\; u \geq 0 \ . \end{aligned}$$

The two linear programs can be combined into one:

$$\begin{aligned} L(f(x), y) \quad = \quad & \min_{\xi} \quad \sum_{r \in R_y} \xi_r \\ & \text{s.t.} \quad \forall p \in P_y : \;\; \xi_{s_y(p)} \geq 1 - \mu_p(f(x), y) \\ & \qquad\; \forall r \in R_y : \;\; \xi_r \geq 0 \end{aligned}$$

The first constraint can be rewritten as

$$\mu_p(f(x), y) = \sum_m \nu_{y,p,m} \cdot f_m(x) \geq 1 - \xi_{s_y(p)} \ .$$

Thus, the decision function values enter a multi-class loss based on the hinge loss as parameters of a linear program.

B. Deriving the Uniform Dual Problems

For deriving the dual problem from the primal, we introduce Lagrange multipliers $\alpha_{i,p} \geq 0$, $\beta_{i,r} \geq 0$, $\eta \in \mathcal{H}$, and $\tau \in \mathbb{R}$ corresponding to the constraints of the primal problem, and compute the Lagrangian

$$\begin{aligned} \mathcal{L} = & \frac{1}{2} \sum_c \|w_c\|^2 + C \cdot \sum_{i,r} \xi_{i,r} \\ & + \sum_{i,p} \alpha_{i,p} \left[\gamma_{y_i,p} - \sum_c \nu_{y_i,p,c} \big(\langle w_c, \phi(x_i) \rangle + b_c \big) - \xi_{i,s_{y_i}(p)} \right] - \sum_{i,r} \beta_{i,r} \xi_{i,r} \\ & + \left\langle \eta, \sum_c w_c \right\rangle + \tau \sum_c b_c \ , \end{aligned}$$

with derivatives

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_c} &= w_c - \sum_{i,p} \alpha_{i,p} \nu_{y_i,p,c} \phi(x_i) + \eta = 0 \Rightarrow w_c = \sum_{i,p} \alpha_{i,p} \nu_{y_i,p,c} \phi(x_i) - \eta \qquad (1)\\
\frac{\partial \mathcal{L}}{\partial b_c} &= -\sum_{i,p} \alpha_{i,p} \nu_{y_i,p,c} + \tau = 0 \Rightarrow \sum_{i,p} \alpha_{i,p} \nu_{y_i,p,c} = \tau \\
\frac{\partial \mathcal{L}}{\partial \xi_{i,r}} &= C - \left(\sum_{p \in P_y^r} \alpha_{i,p} \right) - \beta_{i,r} = 0 \Rightarrow \sum_{p \in P_y^r} \alpha_{i,p} \leq C \ .
\end{aligned}$$

The sets P_y^r, $r \in R_y$, are defined as $P_y^r = s_y^{-1}(\{r\}) = \{p \in P_y \,|\, s_y(p) = r\}$. They form a partition of the set P_y of constraints.

To derive the dual in the absence of the sum-to-zero constraint we just set the dual variables η and τ to zero. Then the first derivative above gives us an expression of w_c in terms of α.

In the case with sum-to-zero constraint we get

$$0 = \sum_c w_c = \sum_c \left[\sum_{i,p} \alpha_{i,p} \nu_{y_i,p,c} \phi(x_i) - \eta \right] \Rightarrow \eta = \sum_{i,p} \alpha_{i,p} \left(\frac{1}{d} \sum_c \nu_{y_i,p,c} \right) \phi(x_i)$$

and thus

$$w_c = \sum_{i,p} \alpha_{i,p} \left[\sum_m \left(\delta_{m,c} - \frac{1}{d} \right) \nu_{y_i,p,m} \right] \phi(x_i) \ .$$

To get to the dual problem, we plug this expression into the Lagrangian using the identity

$$\sum_c \left(\delta_{m,c} - \frac{1}{d} \right) \left(\delta_{n,c} - \frac{1}{d} \right) = \left(\delta_{m,n} - \frac{1}{d} \right) \ .$$

C. Proof of Theorem 5

In the following, we outline a proof of Theorem 5. Let $L(f(x), y)$ denote either the loss function used by the AMO machine or the loss function used by the ATM machine, that is, the loss resulting from application of either the max-over-others or the total-max operator to absolute margins:

$$L(f(x), y) = \max_{c \in Y \setminus \{y\}} \left\{ v_c^{\text{abs}}(f(x), y) \right\} = \left[1 + \max_{c \in Y \setminus \{y\}} \left\{ f_c(x) \right\} \right]_+ \qquad \text{(AMO)}$$

or

$$L(f(x), y) = \max_{c \in Y} \left\{ v_c^{\text{abs}}(f(x), y) \right\} = \max \left\{ \left[1 + \max_{c \in Y \setminus \{y\}} \left\{ f_c(x) \right\} \right]_+ , \left[1 - f_y(x) \right]_+ \right\} \ . \qquad \text{(ATM)}$$

Then Theorem 5 states that the minimizer f^* of the corresponding risk $\mathcal{R} = \mathbb{E}[L(f(x), y)]$, subject to the sum-to-zero constraint $\sum_{c \in Y} f_c(x) = 0$, satisfies:

- If there exists a *majority class* $y \in Y$ such that $P_y > (d-1)/d$, then $f_y^*(x) = d-1$ and $f_c^*(x) = -1$ for all $c \in Y \setminus \{y\}$.

- If $P_y < (d-1)/d$ for all $y \in Y$, then $f^*(x) = 0$.

Proof We demonstrate the proof for the AMO loss function. Following Liu (2007), we argue that $f_c^*(x) \geq -1$ for all $c \in Y$. Suppose $f_c(x) < -1$, then it is easy to see that $\tilde{f}$ defined as $\tilde{f}_c(x) = -1$ and $\tilde{f}_e(x) = f_e(x) + (f_c(x)+1)/(d-1)$ fulfills $\mathcal{R}_x(\tilde{f}) \leq \mathcal{R}_x(f)$ contradicting the optimality of f^*. Restricting the solution space to $f_c(x) \geq -1$ allows us to write the point-wise risk as

$$\mathcal{R}_x = \sum_{y \in Y} P_y \cdot \Big(1 + \max\big\{f_c(x) \,\big|\, c \in Y \setminus \{y\}\big\}\Big) \ .$$

Now we pick $y \in \arg\max\{f_c(x) \,|\, c \in Y\}$ and treat the value $f_y(x) \geq 0$ (which is non-negative because of the sum-to-zero constraint) as fixed from now on. We write the point-wise risk as

$$\mathcal{R}_x = P_y \cdot \Big(1 + \max\big\{f_c(x) \,\big|\, c \in Y \setminus \{y\}\big\}\Big) + \sum_{c \neq y} P_c \cdot \big(1 + f_y(x)\big) \ .$$

The best we can do to keep this risk low is to set all components $f_c(x)$, $c \neq y$, to the same value: $f_c(x) = \sum_{c \neq y} f_c(x)/(d-1) = -f_y(x)/(d-1)$ for all $c \in Y \setminus \{y\}$. It holds

$$\begin{aligned}\mathcal{R}_x &= P_y \cdot \left(1 - \frac{f_y(c)}{d-1}\right) + \sum_{c \neq y} P_c \cdot \big(1 + f_y(x)\big) = P_y \cdot \left(1 - \frac{f_y(c)}{d-1}\right) + (1 - P_y) \cdot \big(1 + f_y(x)\big) \\ &= 1 - P_y \cdot \frac{f_y(c)}{d-1} + (1 - P_y) \cdot f_y(x) = 1 + \left(1 - \frac{d}{d-1} P_y\right) f_y(x) \ .\end{aligned}$$

For $P_y > (d-1)/d$ this expression is a decreasing function of $f_y^*(x)$, resulting in the optimum $f_y^*(x) = d-1$ and $f_c^*(x) = -1$ for $c \neq y$, which maximizes $f_y(x)$ under the constraints $\sum_x f_c(x) = 0$ and $\forall c : f_c(x) \geq -1$. In contrast, for $P_y < (d-1)/d$ the risk is lower bounded by one. In this case $f^*(x) = 0$ minimizes the expression yielding $\mathcal{R}_x = 1$.

The analogous result for the ATM loss function can be proven with exactly the same arguments. ■

D. Data Sets

The descriptive statistics of the 12 UCI data sets used in both the linear as well as non-linear SVM experiments are given in Table 2. The additional data sets used in the linear SVM experiments are described in Table 3.

Data set	d	ℓ_{train}	ℓ_{test}	p
Abalone	27	3133	1044	10
Car	4	1209	519	6
Glass	6	149	65	9
Iris	3	105	45	4
Opt. digits	10	3823	1797	64
Page blocks	5	3831	1642	10
Sat	7	4435	2000	36
Segment	7	1617	693	19
Soy bean	19	214	93	35
Vehicle	4	592	254	18
Red wine	10	1120	479	11
White wine	10	3428	1470	11

Table 2: Descriptive statistics of the 12 UCI data sets used in the non-linear SVM study. The columns d, ℓ_{train}, ℓ_{test}, and p contain the number of classes, the number of training examples, the number of test examples, and the input space dimension (number of features), respectively.

Data set	d	ℓ_{train}	ℓ_{test}	p
Covertype	7	406,707	174,305	54
Letter	26	15,000	5,000	16
News-20	20	15,935	3,993	62,061
Sector	105	6,412	3,207	55,197
Usps	10	7,291	2,007	256

Table 3: Descriptive statistics of the additional data sets used in the linear SVM experiments. The columns d, ℓ_{train}, ℓ_{test}, and p contain the number of classes, the number of training examples, the number of test examples, and the input space dimension (number of features), respectively.

E. Model Selection Results

The best parameter configurations (C, γ) for the non-linear SVMs are found in Table 4. The values of the parameter C for the linear SVM experiments are listed in Table 5.

	OVA		**MMR**		**WW**		**CS**		**LLW**		**AMO**		**ATS**		**ATM**		**RM**	
	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$	$\log_2(C)$	$\log_2(\gamma)$
Abalone	-6	1	-3	0	3	-4	0	3	0	0	9	3	0	0	8	4	-1	0
Car	3	-1	2	0	4	-1	4	-1	6	-1	8	-1	5	-1	6	-1	6	-2
Glass	0	-2	0	-1	-2	-1	2	-2	2	0	0	3	4	-3	-3	3	2	-2
Iris	7	-6	1	-2	10	-9	1	-3	7	-5	1	-2	13	-9	1	-2	9	-7
Opt. digits	3	-6	1	-3	3	-5	3	-5	7	-6	-3	-2	7	-6	9	-6	6	-6
Page blocks	1	-1	-2	2	5	-3	7	-4	11	-4	4	-1	10	-4	4	-1	9	-4
Sat	2	-2	1	0	3	-2	3	-2	5	-2	6	-2	4	-2	6	-2	3	-2
Segment	6	0	1	1	7	0	9	-5	9	0	12	-4	9	0	9	-2	8	0
Soy bean	2	-5	1	-3	1	-5	3	-5	6	-6	9	-5	8	-7	9	-5	8	-7
Vehicle	12	-7	0	-2	11	-7	11	-8	14	-7	17	-8	16	-8	17	-8	16	-9
Red wine	-1	0	-1	0	-4	0	-1	0	2	0	3	0	2	0	2	0	1	0
White wine	5	0	1	0	-1	0	6	0	1	1	4	0	3	0	4	0	5	0

Table 4: Best hyperparameter values $(\log_2(C), \log_2(\gamma))$ found by the model selection procedure.

References

Y. Liu. Fisher consistency of multicategory support vector machines. In M. Meila and X. Shen, editors, *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 2 of *JMLR W&P*, pages 289–296, 2007.

	OVA	**MMR**	**WW**	**CS**	**LLW**	**AMO**	**ATS**	**ATM**	**RM**
Cover type	-1	-18	-11	2	7	-5	5	-5	-4
Letter	12	-10	-4	8	14	-12	11	-12	-8
News-20	1	0	1	2	6	10	6	10	5
Sector	1	1	1	2	8	14	8	14	7
Usps	-3	-11	1	-2	17	-4	13	-4	0
Abalone	-9	-14	3	-5	-5	-3	-3	-3	-5
Car	1	-17	10	0	19	-11	0	-11	-1
Glass	7	-3	2	5	8	9	6	9	8
Iris	2	-6	1	-2	2	-8	3	-8	4
Opt. digits	-1	-7	-2	-3	-13	-6	-12	-6	-6
Page blocks	11	-5	13	5	3	5	4	5	2
Sat	1	-23	2	5	5	0	-7	0	-11
Segment	-1	-10	12	8	2	14	7	14	6
Soybean	-2	-6	3	10	22	5	8	5	9
Vehicle	9	-4	8	7	7	11	11	11	-4
Red wine	-1	-8	-1	2	6	-11	-3	-11	3
White wine	-5	-5	0	-2	0	-8	1	-8	5

Table 5: Best hyperparameter values ($\log_2(C)$) for linear models found by the model selection procedure.