An Exact Solver for the Weston-Watkins SVM Subproblem

Yutong Wang

University of Michigan

June 20, 2021

Joint work with my advisor

Clay Scott
EECS & Statistics

Review: Binary linear classifier

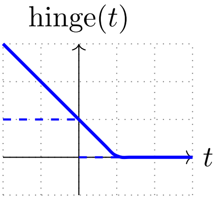
- Data: $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, $i = 1, \ldots, n$
- Vector: $w \in \mathbb{R}^d$
- i-th Margin: $y_i w' x_i$
- Classifier: $x \mapsto \mathrm{sign}(w'x)$
- Goal: find w such that $\uparrow$ is accurate

Review: Binary Support vector machines (SVM)

- Data: $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, $i = 1, \ldots, n$
- Hyperparameter: $C \in \mathbb{R}_{>0}$

Boser, Guyon, and Vapnik (1992)

$$\tilde{w} = \arg\min_{w \in \mathbb{R}^d} \ \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \mathrm{hinge}(y_i w' x_i)$$



Multiclass linear classifier

- Data: $(x_i, y_i) \in \mathbb{R}^d \times [k]$, where $[k] = \{1, \ldots, k\}$
- Matrix: $\mathbf{w} = \begin{bmatrix} w_1 & \cdots & w_k \end{bmatrix} \in \mathbb{R}^{d \times k}$
- (i, j)-th margin: $w'_{y_i} x_i - w'_j x_i$
- Classifier: $x \mapsto \arg\max_{j \in [k]} \; w'_j x$
- Goal: find $\mathbf{w}$ such that $\uparrow$ is accurate

Multiclass SVM

- Data: $(x_i, y_i) \in \mathbb{R}^d \times [k]$, $i = 1, \dots, n$
- Hyperparameter: $C \in \mathbb{R}_{>0}$

Crammer and Singer (2001) (abbrev. CS)

$$\tilde{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^{d \times k}} \frac{1}{2} \|\mathbf{w}\|_F^2 + C \sum_{i=1}^{n} \max_{j \in [k] : j \neq y_i} \mathrm{hinge}(w_{y_i}' x_i - w_j' x_i)$$

Multiclass SVM

- Data: $(x_i, y_i) \in \mathbb{R}^d \times [k]$, $i = 1, \ldots, n$
- Hyperparameter: $C \in \mathbb{R}_{>0}$

Weston and Watkins (1999) (abbrev. WW)

$$\tilde{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^{d \times k}} \frac{1}{2} \|\mathbf{w}\|_F^2 + C \sum_{i=1}^{n} \sum_{j \in [k] : j \neq y_i} \mathrm{hinge}(w_{y_i}' x_i - w_j' x_i)$$

Solving the SVM optimization via the dual

Dual optimization

- Dual variables: $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \cdots & \alpha_n\end{bmatrix} \in \mathbb{R}^{k\times n}$
- Solve: $\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathcal{F}} \ f(\boldsymbol{\alpha})$
- Primal solution: $\tilde{\mathbf{w}} = -\sum_{i\in[n]} x_i\tilde{\alpha}_i'$

Optimizer (block coordinate descent, Keerthi et al. (2008)):

1. Initialize $\boldsymbol{\alpha}$
2. **Solve i-th subproblem:** optimize α_i, fixing all other α_s $s \neq i$
3. Repeat the above step, cycling through $i \in [n]$ multiple times

Which multiclass SVM variant to use?

Doğan, Glasmachers, and Igel (2016)

- benchmarked 9 variants of multiclass SVMs
- recommendation: "WW SVM should be used as the default since it gives robust performance at moderate training times"

Solving the subproblem

For the CS SVM

1. $\exists$ exact subproblem solver in $O(k \log k)$ time
2. References: Crammer and Singer (2001), Duchi et al. (2008), Blondel, Fujino, and Ueda (2014), and Condat (2016)

For the WW SVM,

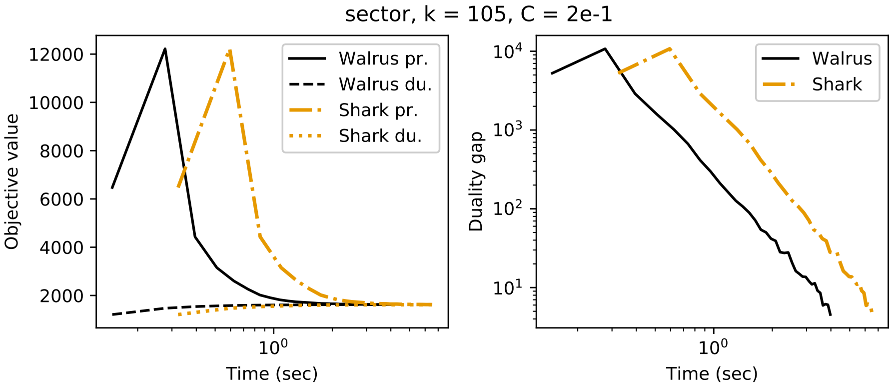
1. No exact subproblem solver known
2. Doğan, Glasmachers, and Igel (2016) used Shark (state-of-the-art iterative solver)

Our contributions

For the WW SVM,

1. Walrus: first exact subproblem solver in $O(k \log k)$ time
2. implementation in LIBLINEAR available on github
3. significant speed up over Shark

Implementation and benchmarking



Thank you!

Questions? Feel free to reach me at:

`yutongw@umich.edu`

Blondel, Mathieu, Akinori Fujino, and Naonori Ueda (2014). "Large-scale multiclass support vector machine training via Euclidean projection onto the simplex". In: *2014 22nd International Conference on Pattern Recognition*. IEEE, pp. 1289–1294.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Condat, Laurent (2016). "Fast projection onto the simplex and the l_1 ball". In: *Mathematical Programming* 158.1-2, pp. 575–585.

Crammer, Koby and Yoram Singer (2001). "On the algorithmic implementation of multiclass kernel-based vector machines". In: *Journal of Machine Learning Research* 2.Dec, pp. 265–292.

Doğan, Ürün, Tobias Glasmachers, and Christian Igel (2016). "A unified view on multi-class support vector classification". In: *The Journal of Machine Learning Research* 17.1, pp. 1550–1831.

Duchi, John et al. (2008). “Efficient projections onto the l_1-ball for learning in high dimensions”. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279.

Keerthi, S Sathiya et al. (2008). “A sequential dual method for large scale multi-class linear SVMs”. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 408–416.

Weston, Jason and Chris Watkins (1999). “Support Vector Machines for Multi-Class Pattern Recognition”. In: *Proc. 7th European Symposium on Artificial Neural Networks, 1999*.