Can Entropic Regularization Be Replaced by Squared Euclidean Distance Plus Additional Linear Constraints

Manfred K. Warmuth

Univ. of Calif. at Santa Cruz

Abstract. There are two main families of on-line algorithms depending on whether a relative entropy or a squared Euclidean distance is used as a regularizer. The difference between the two families can be dramatic. The question is whether one can always achieve comparable performance by replacing the relative entropy regularization by the squared Euclidean distance plus additional linear constraints. We formulate a simple open problem along these lines for the case of learning disjunctions.

Assume the target concept is a k literal disjunction over n variables. The instances are bit vectors $\boldsymbol{x} \in \{0,1\}^n$ and the disjunction $V_{i_1} \vee V_{i_2} \vee \ldots V_{i_k}$ is true on instance $\boldsymbol{x}$ iff at least one bit in the positions $i_1, i_2, \ldots, i_k$ is one. We can represent the above disjunction as a weight vector $\boldsymbol{w}$: all relevant weights w_{i_j} are set to some threshold $\theta > 0$ and the remaining $n-k$ irrelevant weights are zero. Now the disjunction is a linear threshold function: the disjunction is true on $\boldsymbol{x}$ iff $\boldsymbol{w} \cdot \boldsymbol{x} \geq \theta$.

The following type of on-line algorithm makes at most $O(k \log n)$ mistakes on sequences of examples $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, when the labels y_t are consistent[1] with a k-literal monotone disjunction: The algorithm predicts true on instance $\boldsymbol{x}_t$ iff $\boldsymbol{w}_t \cdot \boldsymbol{x}_t \geq \theta$. The weight vector $\boldsymbol{w}_t$ for predicting at trial t is determined by minimizing the *relative entropy* to the initial weight vector $\boldsymbol{w}_1$ subject to some linear constraints implied by the examples. Here the relative entropy is defined as $\Delta(\boldsymbol{w}, \boldsymbol{w}_1) = \sum_i w_i \ln \frac{w_i}{w_{1,i}} + w_{1,i} - w_i$. More precisely, $\boldsymbol{w}_t := \min_{\boldsymbol{w}} \Delta(\boldsymbol{w}, \boldsymbol{w}_1)$ subject to the following *example constraints* (where $\theta, \alpha > 0$ are fixed):

- $\boldsymbol{w} \cdot \boldsymbol{x}_q = 0$, for all $1 \leq q < t$ and $y_t = false$,
- $\boldsymbol{w} \cdot \boldsymbol{x}_q \geq \alpha\theta$, for all $1 \leq q < t$ and $y_t = true$.

This algorithm is a variant of the Winnow algorithm [Lit88] which, for $\boldsymbol{w}_1 = (1, \ldots, 1)$, $\alpha = e$ and $\theta = \frac{n}{e}$, makes at most $e + ke \ln n$ mistakes on any sequence of examples that is consistent with a k out of n literal disjunction.[2]

The crucial fact is that the mistake bound of Winnow and its variants grows logarithmically in the number of variables, whereas the mistake bound of the Perceptron algorithm is $\Omega(kn)$ [KWA97]. The question is, what is responsible for this dramatic difference?

[1] For the sake of simplicity we only consider the noise-free case.

[2] An elegant proof of this bound was first given in [LW04] for the case when the additional constraint $\sum_i \boldsymbol{w}_i = 1$ is enforced: for $\boldsymbol{w}_1 = (\frac{1}{n}, \ldots, \frac{1}{n})$, $\alpha = e$ and $\theta = \frac{1}{ek}$, this algorithm makes at most $ek \ln n$ mistakes.

G. Lugosi and H.U. Simon (Eds.): COLT 2006, LNAI 4005, pp. 653–654, 2006.

The Perceptron type algorithm is motivated by minimizing a different divergence subject to threshold constraints defined by the positive and negative examples: the squared Euclidean distance $||\boldsymbol{w}||_2^2$. The algorithms in this family maintain a weight vector that is a linear combination of the past instances and all algorithms in this family require at least $\Omega(n+k)$ mistakes when learning k out of n literal disjunctions [KWA97].

The question is whether[3] minimizing $||w||_2^2$ subject to the example constraints **plus** some additional linear constraints can achieve the same feat as the relative entropy minimization and lead to the improved mistake bound of $O(k \log n)$. In our experiments on artificial data, two additional constraints do the trick: if $\sum_i w_i = 1$ and the n non-negativity constraints $w_i \geq 0$ are enforced in addition to the example constraints, then choosing $\theta = \frac{1}{2k}$ and $\alpha = 2$ seems to achieve the $O(k \log n)$ mistake bound. Dropping the $\sum_i w_i = 1$ constraint only slightly increases the number of mistakes. On the other hand, with only the example constraints, the mistake bound grows linearly with n. Adding the $\sum_i w_i = 1$ constraint helps only slightly and adding the $\sum_i |w_i| = 1$ constraint gives moderate improvements.

The advantage of the family that uses the squared Euclidean distance is that the algorithms can be kernelized. However, both the non-negativity constraints as well as the one-norm constraint destroy this property. See [KW97], Section 9.6, and [KRS01, SM05] for additional discussion in the context of regression.

Acknowledgements. Dima Kuzmin for providing experimental evidence.

References

[KRS01] Roni Khardon, Dan Roth, and Rocco Servedio. Efficiency versus convergence of Boolean kernels for on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 423–430. MIT Press, Cambridge, MA, 2001.

[KW97] J. Kivinen and M. K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, January 1997.

[KWA97] J. Kivinen, M. K. Warmuth, and P. Auer. The perceptron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. *Artificial Intelligence*, 97:325–343, December 1997.

[Lit88] N. Littlestone. Learning when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[LW04] P. M. Long and Xinyu Wu. Mistake bounds for maximum entropy discrimination. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, December 2004.

[SM05] Vishwanathan S.V.N. and Warmuth M.K. Leaving the span. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT 05)*, Bertinoro, Italy, June 2005. Springer. A longer journal version is in preperation.

[3] A slightly more general case is minimizing $||\boldsymbol{w} - \boldsymbol{w}_1||_2^2$ for some uniform start vector $\boldsymbol{w}_1$.