Online Learning with Knowledge–Based Support Vector Machines

Gautam Kunapuli
University of Wisconsin–Madison
kunapg@biostat.wisc.edu

Kristin Bennett
Rensselaer Polytechnic Institute
bennek@rpi.edu

Richard Maclin
University of Minnesota–Duluth
rmaclin@d.umn.edu

Jude Shavlik
University of Wisconsin–Madison
shavlik@cs.wisc.edu

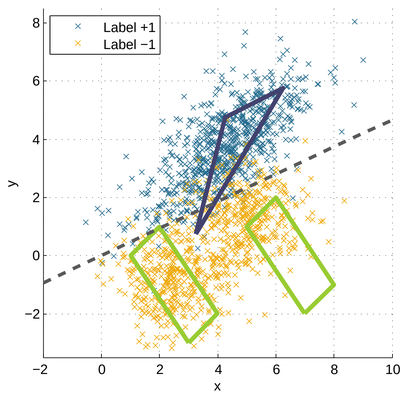
This research was supported by DAPRA grant FA8650-06-C-7606.

Abstract. We propose a novel approach for incorporating prior knowledge into the online binary support vector classification problem. An existing advice-taking approach, when prior knowledge is in the form of *polyhedral knowledge sets* in input space of data, is via *knowledge-based support vector machines* (KBSVMs). We adopt the formalism of passive-aggressive algorithms to derive an online version of KBSVMs when the advice is fixed for every learning round. The goal is to successively update the decision function taking into account *prior knowledge* in the form of soft polyhedral advice so as to make increasingly accurate predictions on subsequent rounds. The advice helps speed up and bias learning so that better generalization can be obtained with less data.

Adding Knowledge to SVMs

Polyhedral knowledge sets in the input space of data, can be added to SVMs via *knowledge-based support vector machines* (KBSVMs) [FMS03]. The knowledge sets typically characterize an area of input space as belonging to one of the two classes and *the advice is labeled*: $z = \pm 1$.

Knowledge is specified using $D\mathbf{x} \leq \mathbf{d} \Rightarrow z(\mathbf{w}'\mathbf{x} - \gamma) \geq 1$. This means every point $\mathbf{x} \in D\mathbf{x} \leq \mathbf{d}$ lies above $\mathbf{w}'\mathbf{x} - \gamma = 1$ (if labeled $z = 1$) or below $\mathbf{w}'\mathbf{x} - \gamma = -1$ (if labeled $z = -1$).



Using *theorems of the alternative*, the *logical implication* can be reformulated as constraints, where prior knowledge is characterized by the *knowledge variables*, $\mathbf{u}$:

$$D'\mathbf{u} + z\,\mathbf{w} = 0, \quad -d'\mathbf{u} - z\,\gamma \geq 1, \quad \mathbf{u} \geq 0.$$

Soft advice can be allowed by relaxing the constraints:

$$D'\mathbf{u} + z\,\mathbf{w} + \boldsymbol{\eta} = 0, \quad -d'\mathbf{u} - z\,\gamma + \zeta \geq 1, \quad \mathbf{u} \geq 0.$$

Online Knowledge-Based SVMs

There are m labeled knowledge sets $(D_i, \mathbf{d}^i, z_i)_{i=1}^m$, and at round t, the algorithm receives labeled data, $(\mathbf{x}^t, y_t)$. We use the *passive-aggressive approach* [CDK$^+$06], in which updates are only computed if the loss of the data point $(\mathbf{x}^t, y_t)$ with respect to the current hypothesis $\mathbf{w}^t$ is non-zero.

At round t, given $\mathbf{w}^t$ and $\mathbf{u}^{i,t}$, the updates, $(\mathbf{w}^{t+1}, \mathbf{u}^{i,t+1})$, can be computed as the optimal solution to

$$\begin{aligned} \underset{\mathbf{w},\mathbf{u}^i,\xi,\boldsymbol{\eta}^i,\zeta_i}{\arg\min}\ & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|_2^2 + \frac{1}{2}\sum_{i=1}^m \|\mathbf{u}^i - \mathbf{u}^{i,t}\|_2^2 + \frac{\lambda}{2}\,\xi^2 + \frac{\mu}{2}\sum_{i=1}^m \|\boldsymbol{\eta}^i\|_2^2 + \zeta_i^2 \\ \text{sub. to}\ & y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\ & \left.\begin{aligned} D_i'\mathbf{u}^i + z_i\mathbf{w} + \boldsymbol{\eta}^i &= 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i &\geq 0 \\ \mathbf{u}^i &\geq 0 \end{aligned}\right\}\ i = 1, \ldots, m. \end{aligned}$$

This formulation, unlike most passive-aggressive approaches, has *no closed form solution*.

Learning with Fixed Advice

If advice does not change during learning ($\mathbf{u}^i$ is fixed), we have the reduced problem of updating $\mathbf{w}^{t+1}$ using $\mathbf{w}^t$ if we

$$\begin{aligned} \underset{\mathbf{w},\xi,\boldsymbol{\eta}^i}{\text{minimize}}\ & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|_2^2 + \frac{\lambda}{2}\,\xi^2 + \frac{\mu}{2}\sum_{i=1}^m \|\boldsymbol{\eta}^i\|_2^2 \\ \text{sub. to}\ & y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, && \textit{(data)} \\ & D_i'\mathbf{u}^i + z_i\mathbf{w} + \boldsymbol{\eta}^i = 0,\ i = 1, \ldots, m. && \textit{(advice)} \end{aligned}$$

The objective minimizes a *proximal term* (requiring that $\mathbf{w}^{t+1}$ be close to $\mathbf{w}^t$), the squared-loss with respect to data (ξ^2), and squared-loss with respect to advice ($\|\boldsymbol{\eta}^i\|^2$). Denoting the multipliers for data and advice constraints α and $\boldsymbol{\beta}^i$, we have

$$\text{the closed-form update:}\quad \mathbf{w}^{t+1} = \mathbf{w}^t + \alpha y_t\mathbf{x}^t + \sum_{i=1}^m z_i\boldsymbol{\beta}^i.$$

Let $\mathbf{r}^i = -z_i D_i'\mathbf{u}^i$, which represents information about each advice set as a *point in the hypothesis space*. The centroid of these points, the *average advice*, is $\mathbf{r}^t = \frac{1}{m}\sum_{i=1}^m \mathbf{r}^i$. Eliminating $\boldsymbol{\beta}^i$, the update can be expressed as a *convex combination* of estimates according to the data and advice,

$$\mathbf{w}^{t+1} = \nu(\mathbf{w}^t + \alpha_t y_t\mathbf{x}^t) + (1-\nu)\mathbf{r}^t,\ \ \nu = 1/(1+m\mu)$$

where $\nu \in [0, 1]$ controls the learning rate according to the advice. If $\nu = 1$, the update is same as derived in [CDK$^+$06].

Learning Algorithm

Data: At each round t, a new labeled data point $(\mathbf{x}^t, y_t)$,
Labeled advice sets $D_i\mathbf{x} \leq \mathbf{d}^i \implies z_i\mathbf{w}'\mathbf{x} \geq 1$, $i = 1, \ldots, m$,
Aggressiveness parameters $\lambda,\ \mu > 0$ and $\nu = 1/(1+m\mu)$

Input: $\mathbf{u}^{i,1}$ learned from advice only by sampling, $\mathbf{w}^1 = 0$

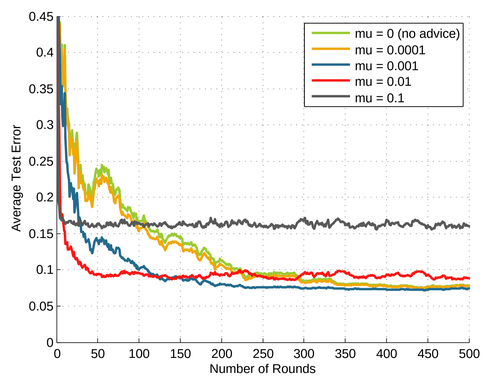
foreach $(\mathbf{x}^t, y_t)$ **do**
- predict label $\hat{y}_t = \text{sign}(\mathbf{w}^{t\prime}\mathbf{x}^t)$
- receive correct label y_t
- suffer loss $\ell_t = \max\left(1 - \nu\, y_t\mathbf{w}^{t\prime}\mathbf{x}^t - (1-\nu)\, y_t\mathbf{r}^{t\prime}\mathbf{x}^t,\ 0\right)$
- update hypothesis using $\mathbf{w}^t$ and $\mathbf{r}^t$

$$\alpha_t = \ell_t/(\tfrac{1}{\lambda} + \nu\|\mathbf{x}^t\|_2^2),$$
$$\mathbf{w}^{t+1} = \nu\,(\,\mathbf{w}^t + \alpha_t\, y_t\mathbf{x}^t\,) + (1-\nu)\,\mathbf{r}^t$$

Here, ℓ_t is the hinge loss applied to a convex combination of distance of $\mathbf{x}^t$ from the *current hypothesis* and distance from the *advice-estimate of the hypothesis*.

Numerical Results

Training was performed on 20 randomized runs of the synthetic data set in the figure to the left, with $\lambda = 10^{-3}$ and different values of μ. The results below show the average test error on a hold-out set.



For small μ, the results are identical to $\mu = 0$ (no advice), with slow convergence. For large μ, there is an initial rapid convergence, but to a less optimal hypothesis that is dominated by advice rather than data. An optimal choice of μ ($= 10^{-3}$) balances both these objectives.

References

[CDK$^+$06] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Rsrch.*, 7:551–585, 2006.

[FMS03] G. Fung, O. Mangasarian, and J. Shavlik. Knowledge-based support vector classifiers. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, volume 15, pages 521–528, 2003.