Fixed-Size Pegasos for Large Scale Pinball Loss SVM

Vilen Jumutc Xiaolin Huang Johan A.K. Suykens

Katholieke Universiteit Leuven, ESAT-SCD, Belgium

ROKS Workshop, July 8 - 10, 2013

Outline

Stochastic programming

- By stochastic programming [Nemirovski, 2009] we assume the following unconstrained optimization problem

$$\min_{x \in X}\{f(x) = \mathbb{E}[F(x,\xi)]\}. \tag{1}$$

Here $X \in \mathbb{R}^n$ is a nonempty bounded closed convex set, ξ is a random vector whose probability distribution P is supported on set $\Xi \in \mathbb{R}^d$ and $F : X \times \Xi \to \mathbb{R}$.

Pegasos

- Pegasos [Shalev-Shwartz et al., 2007] has become a widely acknowledged algorithm for learning linear SVMs. It utilizes strongly convex optimization objective and hinge loss which replaces linear constraints.
- As a result we benefit from the faster convergence rates and can directly apply stochastic approaches via instantaneous optimization objective

$$f(w; \mathcal{A}_t) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{|\mathcal{A}_t|}\sum_{(x,y)\in\mathcal{A}_t} \mathbb{L}(w; (x, y)), \tag{2}$$

where $\mathcal{A}_t$ is our current data at evaluation step t and $\mathbb{L}(w; (x, y)) = \max\{0, 1 - y\langle w, x\rangle\}$ stands for the hinge loss.

Pegasos cont'd

- Pegasos in a stochastic programming setting is an iterative subgradient descent algorithm where at every step t we are working with a subsample $\mathcal{A}_t$ and the subgradient of the instantaneous optimization objective is defined as

$$\nabla_t = \lambda w_t - \frac{1}{|\mathcal{A}_t|} \sum_{(x,y)\in\mathcal{A}_t^+} yx, \tag{3}$$

where $\mathcal{A}_t^+$ denotes the subset of $\mathcal{A}_t$ where $\mathbb{L}(w;(x,y)) > 0$. Our bounded closed convex set is $\mathcal{B} = \{w : \|w\| \leq 1/\sqrt{\lambda}\}$ and in theory expectation over ξ is taken *w.r.t.* our iterates.

Pinball Loss

- Pinball Loss [Huang et al., 2012] $\mathbb{L}_\tau$ for SVM classifier is

$$\mathbb{L}_\tau(w;(x,y)) = \begin{cases} 1 - y\langle w,x\rangle & y\langle w,x\rangle \leq 1, \\ \tau(y\langle w,x\rangle - 1), & y\langle w,x\rangle > 1, \end{cases} \tag{4}$$

  where the reasonable range of τ is $[0,1]$. The pinball loss $\mathbb{L}_\tau$ has been successfully applied for quantile regression, see e.g. [Koenker, 2005].
- Hinge loss is a special case of pinball loss with $\tau = 0$.

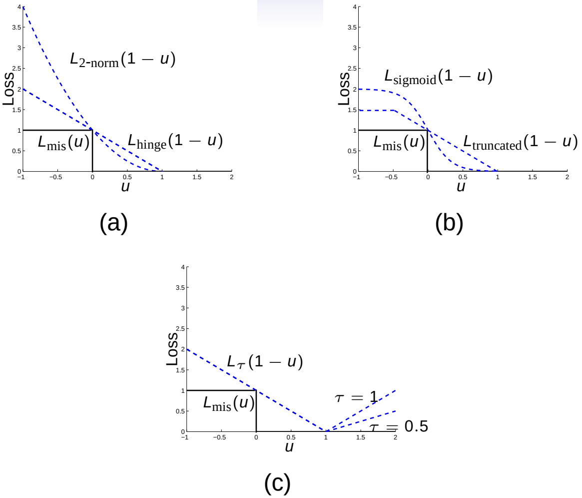

Figure : Loss $L_{\text{mis}}(u)$ is shown by solid lines and some loss functions are displayed by dashed lines: (a) the hinge loss and the 2-norm loss; (b) the normalized sigmoid loss and the truncated hinge loss; (c) the pinball loss with $\tau = 0.5$ and $\tau = 1$.

Pinball Loss vs. Hinge Loss

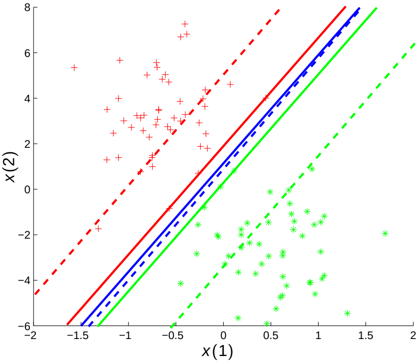

Figure : Points in two classes are marked by red crosses and green stars. The "hyperplanes" are shown by green, blue, and red lines, corresponding to $\langle w, x \rangle = 1, 0,$ and -1, respectively. The solution of the hinge loss SVM is marked by the solid lines.

Pinball Loss vs. Hinge Loss (cont'd)

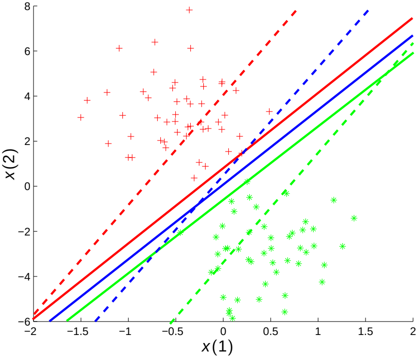

Figure : The results of the hinge loss SVM (the solid lines) differ significantly. In contrast, the results of the pinball loss SVM (the dashed lines) are more stable to re-sampling, which is suitable for stochastic subgradient methods.

Algorithm 1: Pegasos with pinball loss

Data: $\mathcal{S}, \lambda, \tau, T, k, \epsilon$

Select w_1 randomly s.t. $\|w^{(1)}\| \leq 1/\sqrt{\lambda}$;

for $t = 1 \rightarrow T$ **do**

- Set $\eta_t = \frac{1}{\lambda t}$
- Select $\mathcal{A}_t \subseteq \mathcal{S}$, where $|\mathcal{A}_t| = k$;
- $\rho = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{A}_t} (y - \langle w_t, x\rangle)$;
- $\mathcal{A}_t^+ = \{(x,y) \in \mathcal{A}_t : y(\langle w_t, x\rangle + \rho) < 1\}$;
- $\mathcal{A}_t^- = \{(x,y) \in \mathcal{A}_t : y(\langle w_t, x\rangle + \rho) > 1\}$;
- $w_{t+\frac{1}{2}} = w_t - \eta_t(\lambda w_t - \frac{1}{k}\left[\sum_{(x,y)\in\mathcal{A}_t^+} yx - \sum_{(x,y)\in\mathcal{A}_t^-} \tau yx\right])$;
- $w_{t+1} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\|w_{t+\frac{1}{2}}\|}\right\} w_{t+\frac{1}{2}}$;
- **if** $\|w_{t+1} - w_t\| \leq \epsilon$ **then**
  - **return** $(w_{t+1}, \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} (y - \langle w_t, x\rangle))$;
- **end**

end

return $(w_{T+1}, \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} (y - \langle w_{T+1}, x\rangle))$;

Convergence bounds

Based on the Lemma 1 in [Shalev-Shwartz et al., 2007], we can bound the average instantaneous objective of Algorithm 1 in Theorem 1 [Jumutc et al., 2013].

Theorem

Assume $\|x\| \leq R$ *for all* $(x, y) \in \mathcal{S}$. *Let*

$$w^* = \arg\min_{\mathrm{w}} \left[\frac{\lambda}{2}\|\mathrm{w}\|^2 + \frac{1}{|\mathcal{A}_{\mathrm{t}}|} \sum_{(\mathrm{x},\mathrm{y}) \in \mathcal{A}_{\mathrm{t}}} \mathbb{L}_{\tau}(\mathrm{w}; (\mathrm{x}, \mathrm{y})) \right]$$

and let $c = (\sqrt{\lambda} + (\tau + 1)R)$. *Then, for* $T \geq 3$ *we have*

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t; \mathcal{A}_t) \leq \frac{1}{T}\sum_{t=1}^{T} f(w^*; \mathcal{A}_t) + \frac{c^2 \ln(T)}{\lambda T}.$$

Fixed-Size approach

- Algorithm 1 operates only in the primal space. To handle this restriction we go for the Fixed-Size approach [Suykens et al., 2002].
- Entropy based criterion is used to select m prototype vectors and construct $m \times m$ RBF kernel matrix K.
- Nyström approximation [Williams and Seeger, 2001] gives an expression for the entries of the approximated feature map $\hat{\Phi}(x) : \mathbb{R}^d \rightarrow \mathbb{R}^m$ with $\hat{\Phi}(x) = (\hat{\Phi}_1(x), \ldots, \hat{\Phi}_m(x))^T$ and

$$\hat{\Phi}_i(x) = \frac{1}{\sqrt{\lambda_{i,m}}} \sum_{t=1}^{m} u_{ti,m} k(x_t, x), \tag{5}$$

where $\lambda_{i,m}$ and $u_{i,m}$ denote the i-th eigenvalue and the i-th eigenvector of K and $k(x, y)$ denotes the RBF function.

Algorithm 2: Complete procedure

Data : training data $\mathcal{S}$ with $|\mathcal{S}| = n$, labeling Y, parameters $\lambda, \tau, T, k, \epsilon, m$

Return: mapping $\hat{\Phi}(x), \forall x \in \mathcal{S}$, SVM model given by w and ρ

begin

$\quad \mathcal{S}_r \leftarrow \texttt{FindActiveSet}(\mathcal{S}, m)$;

$\quad \hat{\Phi}(x) \leftarrow \texttt{ComputeNystromApprox}(\mathcal{S}_r)$;

$\quad X \leftarrow [\hat{\Phi}(x_1)^T, \ldots, \hat{\Phi}(x_n)^T]$;

$\quad [w, \rho] \leftarrow \texttt{PegasosPBL}(X, Y, \lambda, \tau, T, k, \epsilon)$;

end

General notes on the procedure

- In Algorithm 2 "PegasosPBL" function stands for the shortcut of Algorithm 1.
- "ComputeNystromApprox" function denotes the Fixed-Size part.
- "FindActiveSet" function denotes entropy based selection of prototype vectors.

Toy datasets and evaluation

Table : Test errors for Pegasos$_{pbl}$ with the dataset of size 10000

Dataset (% of distortion)	Hinge Loss	Pinball Loss $\tau = 0.1$	Pinball Loss $\tau = 0.5$	Pinball Loss $\tau = 1$
Toy Data (5%)	0.08262	**0.06908**	0.06926	0.07446
Toy Data (15%)	0.18753	**0.15843**	0.16141	0.16538
Toy Data (35%)	0.36094	0.31829	0.32335	**0.31571**

UCI datasets and evaluation

Table : Test errors for Pegasos$_{pbl}$ with $k = 1$(fully stochastic)

Dataset	Size	Hinge Loss	Pinball Loss $\tau = 0.1$	Pinball Loss $\tau = 0.5$	Pinball Loss $\tau = 1$
Pima	768	0.28896	0.29422	**0.28870**	0.29198
Spambase	4601	0.21444	0.21229	**0.20816**	0.21903
Transfusion	748	0.23406	0.23465	**0.23396**	0.23465
White Wine	4898	0.29607	0.29526	0.29694	**0.28898**
Magic	19020	0.22667	**0.22385**	0.22481	0.22750
Shuttle	58000	0.04505	0.04145	**0.03499**	0.03736
Skin	245057	0.02705	0.02498	**0.02172**	0.02401

Convergence of Pegasos algorithms

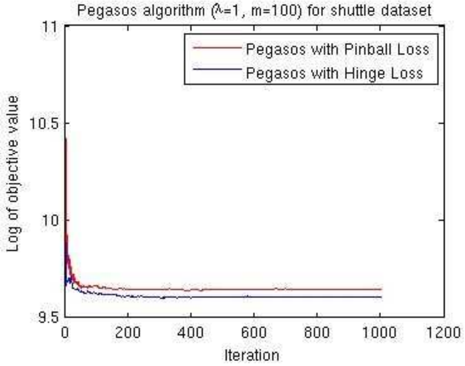

Figure : Convergence of Pegasos algorithm for Shuttle dataset in a long term (1000 iterations) for hinge loss (blue) and pinball loss (red) respectively. In the experimental setup $\lambda = 1$ and $k = 100$.

Pegasos's "pros"

- Pegasos algorithm in general is suitable for large-scale linear and fixed-size SVM learning.
- Pegasos algorithm in a fully stochastic setting is suitable for online learning.
- Incorporating other loss functions (e.g. pinball loss) might be beneficial in terms of the generalization error and convergence.

References

X. Huang, L. Shi, and J. A. K. Suykens.
Support vector machine classifier with pinball loss.
Technical Report KUL-12-162, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven, 2012.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro.
Robust stochastic approximation approach to stochastic programming.
SIAM J. on Optimization, 19(4):1574–1609, January 2009.

S. Shalev-Shwartz, Y. Singer and N. Srebro.
Pegasos: Primal Estimated sub-GrAdient SOlver for SVM.
In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 807–814, New York, NY, USA, 2007.

R. Koenker.
Quantile Regression.
Econometric Society Monographs. Cambridge University Press, 2005.

C. Williams and M. Seeger.
Using the Nyström method to speed up kernel machines.
In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle.
Least Squares Support Vector Machines.
World Scientific, Singapore, 2002.

V. Jumutc, X. Huang, and J. A. K. Suykens.
Fixed-size pegasos for hinge and pinball loss SVM.
Technical Report KUL-13-31, KULeuven, Kasteelpark Arenberg 10, Leuven, 2013. Accepted in IJCNN 2013.

Thank you for your attention!