

Fundamentals of Supervised Learning

SOCN report

Massih-Reza Amini

Different learning algorithms for binary classification have been proposed for the minimization of the following learning objective over the class of linear functions, $\mathcal{H} = \{h : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle\}$:

$$\hat{\mathcal{L}}_m(S, \mathbf{w}) = \frac{1}{m} \sum_{(\mathbf{x}, y) \in S} \ell(h(\mathbf{x}), y) + \frac{\lambda}{2} \|\mathbf{w}\|^2, \quad (1)$$

where $S = (\mathbf{x}_i, y_i)_{1 \leq i \leq m}$ is a training set of size m , $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ a vector representation of an observation, $y \in \{-1, +1\}$ its associated class label, and ℓ an instantaneous loss (called the hing loss) defined as :

$$\ell(h(\mathbf{x}), y) = \max(0, 1 - yh(\mathbf{x})). \quad (2)$$

In the following we will analysis the algorithm called PEGASOS (*Primal Estimated sub-Gradient SOLver for SVM*)¹ which procedure is summarized below.

Algorithm 1 Pegasos

- 1: **Input:** Training set $S = (\mathbf{x}_i, y_i)_{1 \leq i \leq m}$, constant $\lambda > 0$ and maximum number of iterations T
 - 2: **Initialize:** Set $\mathbf{w}^{(1)} \leftarrow 0$
 - 3: **for** $t = 1, 2, \dots, T$ **do**
 - 4: Set $S_t^+ = \{(\mathbf{x}, y) \in S; y \langle \mathbf{w}^{(t)}, \mathbf{x} \rangle < 1\}$
 - 5: Set $\eta_t = \frac{1}{\lambda t}$
 - 6: Update $\mathbf{w}^{(t+1)} \leftarrow (1 - \lambda \eta_t) \mathbf{w}^{(t)} + \frac{\eta_t}{m} \sum_{(\mathbf{x}, y) \in S_t^+} y \mathbf{x}$
 - 7: **end for**
 - 8: **Output:** $\mathbf{w}^{(T+1)}$
-

Begining from a null weight vector, the algorithm iteratively updates the weights over the subset of misclassified training examples S_t^+ by applying the following rule :

$$\forall t, \mathbf{w}^{(t+1)} \leftarrow (1 - \lambda \eta_t) \mathbf{w}^{(t)} + \frac{\eta_t}{m} \sum_{(\mathbf{x}, y) \in S_t^+} y \mathbf{x}, \quad (3)$$

¹S. Shalev-Shwartz, Y. Singer, N. Srebro and A. Cotter. Primal Estimated sub-Gradient SOLver for SVM (Pegasos) *Mathematical Programming* March 2011, Volume 127, Issue 1, pp 330

where, $\eta_t = \frac{1}{\lambda t}$ is the learning rate. In the following we will analysis the convergence property of the algorithm.

1. For an observation (\mathbf{x}, y) and a prediction function $h \in \mathcal{H}$, why the sign of the product $yh(\mathbf{x}) = y\langle \mathbf{w}, \mathbf{x} \rangle$ is an indicator of good/bad classification?
2. Which other learning algorithm updates the learning weights over misclassified training examples? In the case where $S_t^+ = (\mathbf{x}_t, y_t)$ is a singleton what is the update rule of this other learning algorithm and what is the difference with the one proposed in PEGASOS (3)?
3. Draw the binary classification loss $\ell_b : (h(\mathbf{x}), y) \mapsto \mathbb{1}_{yh(\mathbf{x}) < 0}$, and the hing loss (Eq. 2) with respect to the product $yh(\mathbf{x})$, i.e. the loss on the y -axis and $yh(\mathbf{x})$ on the x -axis.
4. For a given example (\mathbf{x}, y) , what does $\frac{|h(\mathbf{x})|}{\|\mathbf{w}\|}$ represent?
5. Why the learning objective (1) admits a single minimizer $\mathbf{w}^* \in \mathbb{R}^d$?
6. Explain why at the first iteration, S_1 is the whole training set; $S_1 = S$?
7. Show that the update (3) follows the gradient descent rule:

$$\forall t, \mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta_t \nabla_t$$

where $\nabla_t = \nabla_t \hat{\mathcal{L}}_m(S, \mathbf{w}^{(t)})$ denotes the gradient of the learning objective (Eq. 1) at $\mathbf{w}^{(t)}$.

8. For two consecutive weights $\mathbf{w}^{(t)}$ and $\mathbf{w}^{(t+1)}$, show that

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 = 2\eta_t \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla_t \rangle - \eta_t^2 \|\nabla_t\|^2$$

9. The objective learning function is λ -strongly convex (admitted), that is

$$\forall u \in \mathbb{R}^d, \langle \mathbf{w}^{(t)} - u, \nabla_t \rangle \geq \hat{\mathcal{L}}(\mathbf{w}^{(t)}) - \hat{\mathcal{L}}(u) + \frac{\lambda}{2} \|\mathbf{w}^{(t)} - u\|^2.$$

From this property and the previous question, deduce then

$$\sum_{t=1}^T (\hat{\mathcal{L}}(\mathbf{w}^{(t)}) - \hat{\mathcal{L}}(\mathbf{w}^*)) \leq \sum_{t=1}^T \left(\frac{\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2}{2\eta_t} - \frac{\lambda}{2} \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right) + \frac{1}{2} \sum_{t=1}^T \eta_t \|\nabla_t\|^2$$

10. Show that for two consecutive iterations t and $t + 1$, we have

$$\sum_{j=t}^{t+1} \left(\frac{\|\mathbf{w}^{(j)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(j+1)} - \mathbf{w}^*\|^2}{2\eta_j} - \frac{\lambda}{2} \|\mathbf{w}^{(j)} - \mathbf{w}^*\|^2 \right) = \frac{\lambda(t-1)}{2} \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \frac{\lambda(t+1)}{2} \|\mathbf{w}^{(t+2)} - \mathbf{w}^*\|^2$$

11. From the two previous questions deduce then

$$\begin{aligned} \sum_{t=1}^T \left(\hat{\mathcal{L}}(\mathbf{w}^{(t)}) - \hat{\mathcal{L}}(\mathbf{w}^*) \right) &\leq \frac{-\lambda T}{2} \|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 + \frac{1}{2} \sum_{t=1}^T \eta_t \|\nabla_t\|^2 \\ &\leq \frac{1}{2} \sum_{t=1}^T \eta_t \|\nabla_t\|^2 \end{aligned}$$

12. Suppose that the learning rate $\eta_t = \frac{1}{\lambda t}$, $\forall t$ and that the training data are contained in a ball of radius R ; if at each iteration, we normalize the weights $\mathbf{w}^{(t)}$ such that $\|\mathbf{w}^{(t)}\| \leq \frac{1}{\sqrt{\lambda}}$ show that

$$\|\nabla_t\| \leq \sqrt{\lambda} + R$$

and deduce that for $T \geq 3$

$$\frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{w}^{(t)}) \leq \frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{w}^*) + \frac{c(1 + \ln(T))}{2\lambda T},$$

where, $c = (\sqrt{\lambda} + R)^2$

13. As the learning objective is convex we have from the Jensen inequality that

$$\hat{\mathcal{L}} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)} \right) \leq \frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{w}^{(t)}).$$

Using the above inequality and question 10, prove that

$$\hat{\mathcal{L}}(\mathbf{w}^*) \leq \hat{\mathcal{L}}(\bar{\mathbf{w}}) \leq \hat{\mathcal{L}}(\mathbf{w}^*) + \frac{c(1 + \ln(T))}{2\lambda T},$$

where $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)}$, and finally

$$\lim_{t \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)} = \mathbf{w}^*.$$