A theoretical understanding of self-paced learning

- Part 1-

Self-paced learning

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbf{E}(\mathbf{w},\mathbf{v},\lambda) = \sum_{i=1}^{n} (v_i L(y_i, f(\mathbf{x}_i,\mathbf{w})) + f(v_i,\lambda))$$

$$f^H(v,\lambda) = -\lambda v;\quad v^*(\ell,\lambda) = \begin{cases} 1, & \text{if } \ell < \lambda \\ 0, & \text{if } \ell \geq \lambda \end{cases}$$

$$f^L(v,\lambda) = \lambda(\tfrac{1}{2}v^2 - v);\quad v^*(\ell,\lambda) = \begin{cases} -\ell/\lambda + 1, & \text{if } \ell < \lambda \\ 0, & \text{if } \ell \geq \lambda \end{cases}$$

$$f^M(v,\lambda,\gamma) = \frac{\gamma^2}{v+\gamma/\lambda};\quad v^*(\ell,\lambda,\gamma) = \begin{cases} 1, & \text{if } \ell \leq \left(\frac{\lambda\gamma}{\lambda+\gamma}\right)^2 \\ 0, & \text{if } \ell \geq \lambda^2 \\ \gamma\left(\frac{1}{\sqrt{\ell}} - \frac{1}{\lambda}\right), & \text{otherwise.} \end{cases}$$

Alternative optimization strategy (AOS) commonly utilized to solve the SPL problem

It is still unclear where this SPL iteration converges and why SPL is robust in solving the learning problems especially with highly noisy data.

目录 »

CONTENT

01 通过AOS来优化SPL等价于通过majorization minimization(MM)算法来优化一个implicit SPL objective function

02 the loss function contained in this implicit SPL objective is closely related to the non-convex regularized penalty (NCRP)

03 解释了SPL为什么对噪声数据有效并给出了对SPL内在机制的理解

04 提出了一个新的SPL损失函数以使得其在弱标记数据上的表现更好

majorization minimization (MM) algorithm*

majorize的中文翻译_百度翻译

majorize

并优化

let $Q(w|w^k)$ denote a real-valued function of w whose form depends on w^k. The function $Q(w|w^k)$ is said to majorize a real-valued function F(w) at the point w^k provided

Majorization Step: Substitute $F(\mathbf{w})$ by a surrogate function $Q(\mathbf{w}|\mathbf{w}^k)$ such that:

$$F(\mathbf{w}) \le Q(\mathbf{w}|\mathbf{w}^k)$$

with equality holding at $\mathbf{w} = \mathbf{w}^k$.

Minimization Step: Obtain the next parameter estimate $\mathbf{w}^{k+1}$ by solving the following minimization problem:

$$\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} Q(\mathbf{w}|\mathbf{w}^k).$$

* Hunter D R, Lange K. **A tutorial on MM algorithms.** Am Stat[J]. American Statistician, 2004, 58(February):30-37.

SP-regularizer

Definition 3.1 (SP-regularizer). Suppose that v is a weight variable, ℓ is the loss, and λ is the age parameter. $f(v, \lambda)$ is called a self-paced regularizer, if

1. $f(v, \lambda)$ is convex with respect to $v \in [0, 1]$;
2. $v^*(\ell, \lambda)$ is monotonically decreasing with respect to ℓ, and it holds that $\lim_{\ell \to 0} v^*(\ell, \lambda) = 1$, $\lim_{\ell \to \infty} v^*(\ell, \lambda) = 0$;
3. $v^*(\ell, \lambda)$ is monotonically increasing with respect to λ, and it holds that $\lim_{\lambda \to \infty} v^*(\ell, \lambda) \le 1$, $\lim_{\lambda \to 0} v^*(\ell, \lambda) = 0$;

where

$$v^*(\ell, \lambda) = \arg\min_{v \in [0,1]} v\ell + f(v, \lambda). \quad (3)$$

$$f^H(v, \lambda) = -\lambda v; \quad v^*(\ell, \lambda) = \begin{cases} 1, & \text{if } \ell < \lambda \\ 0, & \text{if } \ell \ge \lambda \end{cases}$$

$$f^L(v, \lambda) = \lambda(\tfrac{1}{2}v^2 - v); \quad v^*(\ell, \lambda) = \begin{cases} -\ell/\lambda + 1, & \text{if } \ell < \lambda \\ 0, & \text{if } \ell \ge \lambda \end{cases}$$

$$f^M(v, \lambda, \gamma) = \frac{\gamma^2}{v + \gamma/\lambda}; \quad v^*(\ell, \lambda, \gamma) = \begin{cases} 1, & \text{if } \ell \le \left(\frac{\lambda\gamma}{\lambda+\gamma}\right)^2 \\ 0, & \text{if } \ell \ge \lambda^2 \\ \gamma\left(\frac{1}{\sqrt{\ell}} - \frac{1}{\lambda}\right), & \text{otherwise.} \end{cases}$$

构造Q函数(Implicit SPL obj.)

$$F_\lambda(\ell) = \int_0^\ell v^*(l,\lambda)dl.$$

泰勒公式

$$f(x) = \frac{f(a)}{0!} + \frac{f'(a)}{1!}(x-a)$$

令Q函数是F(l)在点w*处的一阶泰勒展开，由于V*对l来说是单调递减函数，所以他的积分函数F(l)是一个concave的函数。

$$Q_\lambda(\mathbf{w}|\mathbf{w}^*) = F_\lambda(\ell(\mathbf{w}^*)) + v^*(\ell(\mathbf{w}^*),\lambda)(\ell(\mathbf{w})-\ell(\mathbf{w}^*)).$$

If f is concave and differentiable（可微）, then it is bounded above by its first-order Taylor approximation*

$$F_\lambda(\ell(\mathbf{w})) \leq Q_\lambda(\mathbf{w}|\mathbf{w}^*)$$

只需要证明：AOS来优化SPL等价于MM算法来优化F(l)

* Varian H R. Microeconomic analysis /[M]// **Microeconomic analysis**. Norton, 1984:1 - 28.

MM

Majorization step: To obtain each $Q_{\lambda}^{(i)}(\mathbf{w}|\mathbf{w}^k)$, we only need to calculate $v^*(\ell_i(\mathbf{w}^k), \lambda)$ under the corresponding SP-regularizer $f(v_i, \lambda)$:

$$v^*(\ell_i(\mathbf{w}^k), \lambda) = \min_{v_i \in [0,1]} v_i \ell_i(\mathbf{w}^k) + f(v_i, \lambda).$$

This exactly complies with the AOS step in updating $\mathbf{v}$ in (1) under fixed $\mathbf{w}$.

Minimization step: We need to calculate the following:

$$\begin{aligned}\mathbf{w}^{k+1} &= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} F_{\lambda}(\ell_i(\mathbf{w}^k)) + v^*(\ell_i(\mathbf{w}^k), \lambda)(\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}^k)) \\ &= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} v^*(\ell_i(\mathbf{w}^k), \lambda)\ell_i(\mathbf{w}),\end{aligned}$$

which is exactly equivalent to the AOS step in updating $\mathbf{w}$ in (1) under fixed $\mathbf{v}$.

Conclusion

SPL is a MM algorithm of the implicit SPL objective $\sum_{i=1}^{n} F_\lambda(\ell_i(\boldsymbol{w}))$ with the implicit SPL loss $F_\lambda(\ell\,(\boldsymbol{w}))$

Various off-the-shelf theoretical results of MM can then be readily employed to explain the properties of such SPL solving strategies.

For example, based on the MM theory, the lower-bounded implicit SPL objective is monotonically decreasing during MM/AOS iteration, and the convergence of the SPL algorithm can then be guaranteed.

Thank you