SVM 的随机梯度下降算法

对于线性核函数而言，我们方法的运行时间是˜O (d/(λϱ)),d 取决于每个样例中的非零特征数，因为训练时间并不直接取决于训练集的大小，所以我们的方法可以适用于较大规模的数据。（For a linear kernel, the total run-time of our method is, where d is a bound on the number of non-zero features in each example. Since the run-time does *not* depend directly on the size of the training set, the resulting algorithm is especially suited for learning from large datasets.）

对于 SVM，我们实际上是要找到适合的算法来求解下面的这个规划问题：

$$\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{(\mathbf{x},y)\in S} \ell(\mathbf{w};(\mathbf{x},y))$$

其中

$$\ell(\mathbf{w};(\mathbf{x},y)) = \max\{0, 1 - y\langle\mathbf{w},\mathbf{x}\rangle\} .$$

为了解决这个问题， 我们设计了一个新的算法，称为 Pegasos 算法，算法的流程如下：

INPUT: S, λ, T, k
INITIALIZE: Choose $\mathbf{w}_1$ s.t. $\|\mathbf{w}_1\| \leq 1/\sqrt{\lambda}$
FOR $t = 1, 2, \ldots, T$
  Choose $A_t \subseteq S$, where $|A_t| = k$
  Set $A_t^+ = \{(\mathbf{x}, y) \in A_t : y\langle\mathbf{w}_t, \mathbf{x}\rangle < 1\}$
  Set $\eta_t = \frac{1}{\lambda t}$
  Set $\mathbf{w}_{t+\frac{1}{2}} = (1 - \eta_t\lambda)\mathbf{w}_t + \frac{\eta_t}{k}\sum_{(\mathbf{x},y)\in A_t^+} y\,\mathbf{x}$
  Set $\mathbf{w}_{t+1} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{t+\frac{1}{2}}\|}\right\} \mathbf{w}_{t+\frac{1}{2}}$
OUTPUT: $\mathbf{w}_{T+1}$

该文章同时分析了该算法的收敛性：该文章证明了几个引理，通过数学推导严格证明了该算法所需要的时间：

Lemma 1. *Let $f_1, \ldots, f_T$ be a sequence of λ-strongly convex functions w.r.t. the function $\frac{1}{2}\|\cdot\|^2$. Let B be a closed convex set and define $\Pi_B(\mathbf{w}) = \arg\min_{\mathbf{w}'\in B}\|\mathbf{w} - \mathbf{w}'\|$. Let $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$ be a sequence of vectors such that $\mathbf{w}_1 \in B$ and for $t \geq 1$, $\mathbf{w}_{t+1} = \Pi_B(\mathbf{w}_t - \eta_t\nabla_t)$, where ∇_t is a subgradient of f_t at $\mathbf{w}_t$ and $\eta_t = 1/(\lambda t)$. Assume that for all t, $\|\nabla_t\| \leq G$. Then, for all $\mathbf{u} \in B$ we have*

$$\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{w}_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{u}) + \frac{G^2(1+\ln(T))}{2\lambda T} .$$

Theorem 1. *Assume that for all $(\mathbf{x}, y) \in S$ the norm of $\mathbf{x}$ is at most R. Let $\mathbf{w}^\star$ be as defined in Eq. (5) and let $c = (\sqrt{\lambda} + R)^2$. Then, for $T \geq 3$,*

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t; A_t) \leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}^\star; A_t) + \frac{c\ln(T)}{\lambda T} .$$

Corollary 1. *Assume the conditions stated in Thm. 1 and that $A_t = S$ for all t. Let $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$. Then,*

$$f(\bar{\mathbf{w}}) \leq f(\mathbf{w}^\star) + \frac{c\ln(T)}{\lambda T} .$$

Theorem 2. *Assume that the conditions stated in Thm. 1 hold and for all t, A_t is chosen i.i.d. from S. Let r be an integer picked uniformly at random from $[T]$. Then,*

$$\mathbb{E}_{A_1,\ldots,A_T}\mathbb{E}_r[f(\mathbf{w}_r)] \leq f(\mathbf{w}^\star) + \frac{c\ln(T)}{\lambda T} .$$

Theorem 3. *Assume that the conditions stated in Thm. 2 hold. Let $\delta \in (0,1)$. Then, with probability of at least $1-\delta$ over the choices of $(A_1,\ldots,A_T)$ and the index r we have that*

$$f(\mathbf{w}_r) \leq f(\mathbf{w}^\star) + \frac{c\ln(T)}{\delta\lambda T} .$$

通过上面三个定理，我们可严格推出以下的结论：

1.从定理三我们可以看出我们如果要有1-δ的概率有ε的准确率，我们需要的迭代次数为 $\tilde{O}(1/\lambda\delta\varepsilon)$，在之前的研究中如果想要达到同等效果，需要的迭代次数则是

$$\tilde{O}(\ln(1/\delta)/\lambda\varepsilon^2)$$

可看出我们的算法明显比之前的算法来得好。

2.我们有不小于1-δ的概率，可以得出在某一向量满足如下的不等式

$$f(\hat{\mathbf{w}}_i) - f(\mathbf{w}^\star) \leq \frac{c\,e\,\ln(T)}{\lambda T/s} \leq \frac{c\,e\,\ln(T)\left\lceil \ln\left(\frac{1}{\delta}\right)\right\rceil}{\lambda T} .$$

即函数值十分靠近真实值。