CSci5525 Assginment 3 Extra

Rihan Chen

Nov 14, 2015

1 (a)

The loss function for perceptron is not smooth. It is smooth everywhere except the hinge point, where the subgradients exist for this point. Specifically, it is at the point $w^T x_i = 0$ the hinge loss is not smooth. The subgradient set consists of $[0,\ -y_i x_i]$

1 (b)

For perceptron algorithm, when we set $\eta = 1$,it seems like a subgradient descent method. if an error made the gradient will be $-y_i x_i$, otherwise zero:

$$w^{new} = w^{old} + 1(y_i w^T x_i < 0) y_i x_i$$

if we start with $w^T = 0$, then only times that we add $y_i x_i$ to it is that we made an error on sample (y_i, x_i). Therefore, finally w^T will be the sum of $\alpha_i y_i x_i$, α_i stands for the number of error we made at sample i.

1 (c)

The pesudocode for stochastic gradient method(SGD) is as follow:

For t = 1,...,T

  Randomly draw $i \in \{1, \ldots, m\}$

  Compute subgradient $g_t = -1(y_i w^T x_i < 0) y_i x_i$

  $w^{t+1} = w^t - \eta_t g_t$

Output w_T

In order for the algorithm to converge on non-seperable data set, decaying learning rate is implemented in this case, where $\eta_t = \frac{\eta_0}{\sqrt{t}}$. The η_0 is initial value, the step size is decreasing with the increasing of number of iteration till it's almost zero. So the algorithm must converage finally.

For the rate of convergence, $E[f(\bar{w}_T)] - f(w^*) \leq O(\frac{1}{\sqrt{T}})$, therefore the iteration complexity is $O(\frac{1}{\varepsilon^2})$. For SGD on non-smooth function, the number of iteration should be $O(\frac{1}{\varepsilon^2})$. For each iteration, as only one sample was chosen for updating the weights, the complexity for each each iteration, hence, is $O(1)$. Therefore, the total runtime is $O(\frac{1}{\varepsilon^2})$