CSCI 5525: Machine Learning

Homework 3 Extra

Jingxiang Li

November 15, 2015

Problem 1.a

The loss function is a non-smooth function of w.

Proof. Considering the loss function for each instance i,

$$L_i(w) = \max(0, -y_i w^T x_i)$$

We have

$$\nabla L_i(w) = \begin{cases} -y_i x_i & \text{if } \ y_i w^T x_i < 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that the gradient of $L(w)$ suddenly change at the hinge point, which suggests that L_i is not a smooth function.

Notice that the original loss function is simply the sum over $L_i, \ i = 1, \ldots, n$

$$\sum_{i=1}^{n} \max(0, -y_i w^T x_i) = \sum_{i=1}^{n} L_i(w)$$

since L_i is a non-smooth function of w, the original loss function is not smooth, either.

Q.E.D.

Problem 1.b

Given $\eta = 1$, the update equation of weight vector w for the perceptron algorithm is

$$w_{t+1} = w_t + \mathbb{1}(y_i w_t^T x_i < 0) y_i x_i$$

where $\mathbb{1}$ is the $\{0, 1\}$ indicator function, and $\mathbb{1}(y_i w_t^T x_i < 0)$ equals to 1 when the algorithm makes a mistake on instance $(y_i, \ x_i)$ given w_t as the weight vector. On convergence, since w_0 is 0, we will have

$$\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

where

$$\alpha_i = \sum_{t \in S} \mathbb{1}(y_i w_t^T x_i < 0), \ \ S = \{t \mid (y_i, \ x_i) \text{ is used for updating } w_t\}$$

i.e. α_i is the number of mistakes made by the perceptron algorithm on (y_i, x_i) before convergence.

Q.E.D.

Problem 1.c

Here we design a SGD algorithm which will converge even in the non-seperable setting as follows

1. Initialize step-size η_0
2. Set $w_0 = 0,\ \eta_t = \frac{\eta_0}{\sqrt{t}}$
3. For $t = 1, \ldots, T$
4. Randomly draw $i \in \{1, 2, \ldots, n\}$
5. Compute (sub)gradient $g_t = -\mathbb{1}(y_i w_t^T x_i < 0) y_i x_i$
6. Update: $w_{t+1} = w_t - \eta_t g_t$
7. Output $\bar{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$

To ensure the convergence of the algorithm, we use decaying step-size $\eta_t = \frac{\eta_0}{\sqrt{t}}$. In this way the step-size decreases as t increases, and the algorithm will guarantee convergence when η_t becomes negligible.

Since this is a SGD algorithm, and the loss function for this perceptron problem is in the form

$$\min_w f(w) = \frac{1}{n} \sum_{i=1}^{n} l((x_i, y_i), w) = \frac{1}{n} \sum_{i=1}^{n} \max(0, -y_i w^T x_i)$$

we have $\epsilon = \mathrm{E}[f(\bar{w}_T)] - f(w^*) \leq O(\frac{1}{\sqrt{T}})$, which implies that the iteration complexity is $T = O(\frac{1}{\epsilon^2})$. Since in each iteration the above SGD algorithm considers only 1 sample, i.e. for each iteration the complexity of the algorithm is $O(1)$, then the overall complexity of the SGD algorithm will be $O(\frac{1}{\epsilon^2})$. The expected rate of convergence of the algorithm is $O(\frac{1}{\epsilon^2})$