CSCI 5525: Machine Learning (Fall'15)

Homework 3, Extra Credit problem
Due 11/15/15

1. **(20 points)** The perceptron algorithm considers the hinge loss function over the training data $\{(y_1, \mathbf{x}_i), \ldots, (y_n, \mathbf{x}_n)\}$, with the hinge being at 0 (instead of 1 as for SVMs). The loss function considered by perceptrons can be written as:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \max(0, -y_i \mathbf{w}^T \mathbf{x}_i) \ .$$

   (a) (5 points) Is the loss function a smooth or non-smooth function of $\mathbf{w}$? Clearly explain your answer.

   (b) (8 points) Recall that the basic perceptron algorithm considers updates of the form

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta y_i \mathbf{x}_i$$

   if $\mathbf{w}_t$ makes a mistake on $(y_i, \mathbf{x}_i)$. Assuming the learning problem to be separable and $\eta = 1$, show that the final parameter after convergence is of the form $\hat{\mathbf{w}} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ where α_i is the number of mistakes made by the perceptron algorithm on $(y_i, \mathbf{x}_i)$ before convergence.

   (c) (7 points) When the problem is non-separable, the basic perceptron algorithm is not guaranteed to converge. Based on your knowledge of stochastic gradient descent (SGD), design a SGD algorithm which will converge even in the non-separable setting.[1] Clearly describe the algorithm using pseudo-code, and state the expected rate of convergence of the algorithm.

[1] This will require suitable choice of the learning parameter η_t as discussed in class.