HW 1

Ben Lowman

February 18th, 2017

1. (a) i. $\delta = 0 \Rightarrow$ **Linear Separability:**

If a solution to the linear constraint

$$y_i(\vec{w}^T\vec{x}_i + \theta) = 1 - \delta$$

is found with $\delta = 0$, it must be the case that for $y_i = 1$

$$\vec{w}^T\vec{x}_i + \theta \geq 1 \text{ and } \vec{w}^T\vec{x}_i + \theta \geq 0$$

and for $y_i = -1$

$$\vec{w}^T\vec{x}_i + \theta \leq -1 \text{ and } \vec{w}^T\vec{x}_i + \theta < 0$$

ii. **Linear Separability** $\Rightarrow \delta = 0$**:**

Linear separability implies that there exists a hyperplane $\vec{v}^T x + \rho$ such that

$$\min_{\substack{(\vec{x},y)\in D \\ y=1}} (\vec{v}^T\vec{x} + \rho) \geq 0 > \max_{\substack{(\vec{x},y)\in D \\ y=-1}} (\vec{v}^T\vec{x} + \rho)$$

Let x_+ be the closest positive sample to this hyperplane, and let x_- be the closest negative sample to this hyperplane. It follows that

$$p^+ = \vec{v}^T\vec{x_+} + \rho$$
$$p^- = \vec{v}^T\vec{x_-} + \rho$$

The hyperplane $\vec{v}^T x + \rho$ can be shifted such that it lies exactly between x_+ and x_-. Intuitively, this shift is the negative average of p^+ and p^-, $\frac{p^- - p^+}{2}$. Given this new hyperplane $\vec{v}^T x + \rho + \frac{p^- - p^+}{2}$ seperates x_+ and x_- with equal distance:

$$y_i(\vec{v}^T\vec{x_i} + \rho + \frac{p^- - p^+}{2}) \geq \frac{p^+ - p^-}{2}$$

This equation can be re-written to be of the form

$$y_i(\vec{w}^T\vec{x_i} + \theta) \geq 1 - \delta$$

where $\delta = 0$.

(b) A trivial solution to this problem is to set all of the free variables to zero, $\vec{w} = \theta = \delta = 0$. This solution is not really useful, as it doesn't output any information. In addition to avoiding this useless solution, formulating the linear program constraints as

$$y_i(\vec{w}^T\vec{x}_i + \theta) = 1 - \delta$$

$$\delta \geq 0$$

intuitively ensures that while minimizing δ, the classification of training samples must be preserved as much as possible. A negative value yielded by inference on $(\vec{w}, \theta)$ will be made positive by the y_i factor (also negative) if the classification is correct. Similarly, a positive inference will remain positive if the classification is correct.

(c) Since the data is linearly separable, any optimum solution will have $\delta = 0$. Enumerating the linear program constraint for each element of D (after taking the dot product of w and x vectors):

$$1(w_1 + w_2..., +w_n + \theta) \geq 1 - 0$$

$$-1(-w_1 - w_2..., -w_n + \theta) \geq 1 - 0$$

Both equations can be expressed more succinctly:

$$\sum_{i=1}^{N} w_i \geq 1 \pm \theta$$

Reducing to one equation:

$$\sum_{i=1}^{N} w_i \geq 1 + |\theta|$$

Any (w, θ) that satisfies this equation is an optimal solution.

2. (a)

$$\frac{\partial g_i(w)}{\partial w_k} = \begin{cases} \frac{w_k}{N} & \widetilde{y} = k \text{ and } y_i = k \\ \frac{w_k}{N} - Cx_i & \widetilde{y} \neq k \text{ and } y_i = k \\ \frac{w_k}{N} + Cx_i & \widetilde{y} = k \text{ and } y_i \neq k \\ \frac{w_k}{N} & \widetilde{y} \neq k \text{ and } y_i \neq k \end{cases}$$

(b) Answers are listed in the order questions appear in the given gradient descent algorithm (algorithm not copied here):

$$w_k \leftarrow w_k - \eta \frac{w_k}{N}$$

$$w_{y_i} \leftarrow w_{y_i} + \eta C x_i$$

$$w_{\widetilde{y}} \leftarrow w_{\widetilde{y}} - \eta C x_i$$

(c) Accuracy of 92.68% was achieved when $C = 10^{-6}$. All powers of 10 were tested on the range $[10^{-8}, 10^{2}]$. 10-fold cross validation was used.

(d) Accuracy of 92.71% was achieved when $C = 10^{-2}$. All powers of 10 were tested on the range $[10^{-4}, 10^{2}]$. 10-fold cross validation was used, and each model trained for 100 epochs.

For both parts (b) and (c), training performance greatly limited my ability to attain a better C value. In part (b), the binary classifiers for each class are trained in parallel. In part (c), each candidate C value is tested in parallel. Despite these optimizations, each model took on the order of hours to train, mostly due to 10-fold cross validation. The similar accuracy of both models leads me to believe that my C values are close to optimal.