Caio Brighenti
COSC 480 - Learning From Data
Fall 2019
Problem Set 3

1. Define $M := \min_i y^{(i)}(w^* \cdot x^{(i)})$. Show that $M > 0$.

   Based on our standard perceptron notation, i represents the ith observation, w^* the final set of weights produced by the PLA algorithm that perfectly seperates the data, and $x^{(i)}$ is the data associated with the ith observation. Thus, the expression $y^{(i)}(w^* \cdot x^{(i)})$ represents the predicted outcome of this observation using the PLA algorithm multiplied by the actual value of that observation.

   Intuitively, this quantity represents the distance of each observation from the decision boundary, as the decision boundary of a perceptron is technically 0 (the bias term is subsumed into $w^* \cdot x^{(i)}$). Since the final weights perfectly predict for every observation, $w^* \cdot x^{(i)})$ and y *must* have the same sign, which makes the entire quantity always be positive. For this quantity to be negative, the two terms must necessarily have different signs, which would violate the assumption that w^* accurately classifies all predictions.

   By taking the minimum of these values, we find the observation in the data that is closest to the decision boundary. In other words, we find the data point that is the closest to being incorrectly classified. This quantity is called the *margin*, and it is desirable to have higher values for this.

2. Show that for any $t \geq 1, w(t) \cdot w^* \geq w(t-1) \cdot w^* + M$.

   After each iteration of the PLA algorithm, the weights are updated by choosing a misclassified observation i and adding $y^i(w \cdot x^i)$ to the weights. Thus we can define $w(t)$ as $w(t) = w(t-1) + y^i(w \cdot x^i)$, where i is the misclassified observation chosen. By the definition of M, it also must be that $y^i(w \cdot x^i) \geq M$, as M is the minimum margin within the observations. Therefore we have:

   $$\begin{aligned} w(t) &\geq w(t-1) + M && && (1)\\ w(t) * w^* &\geq w(t-1) * w^* + M && \text{multiply by } w^* && (2) \end{aligned}$$

   Thus $w(t) * w^* \geq w(t-1) * w^* + M$ follows from the definition of M and the PLA algorithm.

3. Use induction to show that $w(t) \cdot w^* \geq tM$.

   We start by providing a base case and inductive hypothesis.

   Base case: Our base case is that $t = 1$, meaning we are on the first iteration of the PLA algorithm. We thus have $w(1) \cdot w^* \geq M$. This must be true as M is the minimum distance to the PLA decision boundary.

   Inductive hypothesis: Assume the claim holds for $t-1$ and $t-2$. We thus show the claim must also hold for t.

   $$\begin{aligned} w(t-1)w^* &\geq w(t-2)w^* + M && \text{shown in 2} && (3)\\ w(t-1)w^* + M &\geq w(t-2)w^* + 2M && \text{add } M \text{ to both sides} && (4)\\ w(t)w^* \geq w(t)w^* + M &\geq w(t-2)w^* + 2M && \text{shown in 2} && (5)\\ w(t)w^* &\geq tM && \text{simplify} && (6) \end{aligned}$$

   We thus have that the claim holds for all $t \geq 1$.