CS 4540/CS 8803: Algorithmic Theory of Intelligence Fall 2023

Lecture 1: The Perceptron Algorithm

Instructor: Santosh Vempala *Lecture date: 8/21*

1 the Perceptron Algorithm

1.1 Settings

Consider labeled data $(x, l(x))$, where $x \in \mathbb{R}^n$, $l(x) \in \{1, -1\}$. Each data is labeled as

$$l(x) = \text{sign}(\langle w^*, x\rangle) = \text{sign}(\sum_{i=1}^{n} w_i^* x_i).$$

Here $w^* \in \mathbb{R}^n$ is the unknown normal vector to the separating hyperplane. We assume wlog that $\|w^*\| = 1, \|x\| \leq 1$. Moreover, we denote that the margin for w^* is $\gamma := \min_x \frac{|\langle w^*, x\rangle|}{\|w^*\|^2}$. This implies that for any x with label $l(x) = 1$, $\langle w^*, x\rangle \geq \gamma$. For any x with label $l(x) = -1$, $\langle w^*, x\rangle \leq -\gamma$.

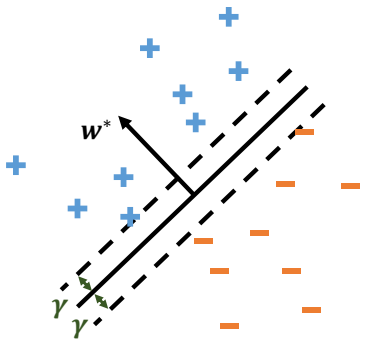

Figure 1: A sketch for the binary classification with normal vector w^* and margin γ.

1.2 Algorithm and Analysis

Algorithm 1 Perceptron Algorithm

Start with $w = 0$.
for each input x **do**
  Predict $\text{sign}(\langle w, x\rangle)$
  On a mistake, set $w \leftarrow w + l(x)x$
end for

Theorem 1.1. *The number of mistakes made by the Perceptron Algorithm on* ***any*** *input sequence is at most* $1/\gamma^2$.

Proof. Consider the potential function $\langle w, w^*\rangle / \|w\|$, which starts at 0. Whenever the algorithm makes a mistake, we update as $\tilde{w} := w + l(x)x$. For the numerator, we have

$$\langle \tilde{w}, w^*\rangle = \langle w + l(x)x, w^*\rangle = \langle w, w^*\rangle + l(x)\langle x, w^*\rangle \geq \langle w, w^*\rangle + \gamma$$

The last inequality comes from the definition of margin γ. For the denominator of the potential function, we have

$$\|\tilde{w}\|^2 = \langle \tilde{w}, \tilde{w} \rangle = \langle w + l(x)x, w + l(x)x \rangle = \langle w, w \rangle + \langle x, x \rangle + 2l(x)\langle w, x \rangle$$

Since x is incorrectly predicted by w, we know $l(x) \neq \text{sign}(\langle w, x \rangle)$. This indicates that $l(x)\langle w, x \rangle \leq 0$. Also we assume that $\|x\|$ is bounded by 1. So we have

$$\|\tilde{w}\|^2 \leq \langle w, w \rangle + \langle x, x \rangle \leq \|w\|^2 + 1$$

So after the algorithm makes T mistakes, the numerator is $\langle \tilde{w}, w^* \rangle \geq \gamma T$, while the denominator is $\|\tilde{w}\| \leq \sqrt{T}$. By the Cauchy-Schwarz inequality, $|\langle \tilde{w}, w^* \rangle| \leq \|\tilde{w}\|\|w^*\| = \|\tilde{w}\|$, which implies that $\langle \tilde{w}, w^* \rangle / \|\tilde{w}\| \leq \|w^*\| = 1$ at all time steps. This implies that

$$1 \geq \frac{\langle \tilde{w}, w^* \rangle}{\|\tilde{w}\|} \geq \frac{\gamma T}{\sqrt{T}} = \gamma\sqrt{T}$$

This derives that $T \leq 1/\gamma^2$.

□

The theorem shows that the number of mistakes made by the Perceptron Algorithm is bounded by $1/\gamma^2$. However, when the margin is tiny, for instance, $\gamma \sim 1/2^n$, the number of mistakes can be as large as $2^{O(n)}$, which is not polynomial with respect to the input dimension n.