THE GEOMETRY OF PERCEPTRONS

MATT RAYMOND

Lemma 0.1. *Define D to be the set of input-output pairs. We call the elements of D data-points. Fix $(\mathbf{x}_i, y_i) \in D$, $\mathbf{w} \in \mathbb{R}^n$ and $\varphi : \mathbb{R} \to \{-1, 1\}$ the binary step activation function.*

(a) *The data-point $(\mathbf{x}_i, y_i)$ is misclassified if and only if $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) \leq 0$.*

(b) *The inequality $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) > 0$ holds if and only if $(\mathbf{x}_i, y_i)$ was classified correctly.*

Proof. If $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) \leq 0$ then $y_i > 0$ and $\varphi(\mathbf{w}^T\mathbf{x}_i) < 0$ or $y_i < 0$ and $\varphi(\mathbf{w}^T\mathbf{x}_i) > 0$. It follows that either $y_i > \varphi(\mathbf{w}^T\mathbf{x}_i)$ or $y_i < \varphi(\mathbf{w}^T\mathbf{x}_i)$. In both cases, $y_i \neq \varphi(\mathbf{w}^T\mathbf{x}_i)$. Hence, $(\mathbf{x}_i, y_i)$ is misclassified. By a similar argument, it is easy to show the converse. Then (a) holds. Since (b) is the contrapositive of (a), (b) holds. This completes the proof. □

Definition 0.2. Let V be a k-dimensional vector space over $\mathbb{R}$. A subspace H is called a hyperplane if it has codimension 1.

Theorem 0.3. *For each $\mathbf{x}_i$, define a orthogonal hyperplane $H(\mathbf{x}_i)$ to $\mathbf{x}_i$. That is, for each $\mathbf{w} \in H(\mathbf{x}_i)$, $\mathbf{w}^T\mathbf{x}_i = 0$. Define $W_\uparrow(\mathbf{x}_i)$ to be the set of $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w}^T\mathbf{x}_i < 0$, and $W_\downarrow(\mathbf{x}_i)$ the set of $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w}^T\mathbf{x}_i > 0$.*

(a) *If $y_i > 0$, then any $\mathbf{w}$ with $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) > 0$ is an element of $W_\uparrow$.*

(b) *If $y_i < 0$, then any $\mathbf{w}$ with $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) > 0$ is an element of $W_\downarrow$.*

(c) *The complement of $W_\uparrow(\mathbf{x}_i) \cup W_\downarrow(\mathbf{x}_i)$ is $H(\mathbf{x}_i)$.*

Proof. If $y_i > 0$ and $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) > 0$ then $\varphi(\mathbf{w}^T\mathbf{x}_i) > 0$. It follows that $\mathbf{w}^T\mathbf{x}_i > 0$, so $\mathbf{w} \in W_\uparrow(\mathbf{x}_i)$. Hence (a) holds. The proof of (b) is so similar it is omitted. If $\mathbf{w}$ is in the complement of $W_\uparrow(\mathbf{x}_i) \cup W_\downarrow(\mathbf{x}_i)$, then $0 \leq \mathbf{w}^T\mathbf{x}_i$ and $\mathbf{w}^T\mathbf{x}_i \leq 0$. It follows that $\mathbf{w}^T\mathbf{x}_i = 0$ so $\mathbf{w} \in H(\mathbf{x}_i)$. The other inclusion is similar, so (c) holds. This completes the proof. □

Corollary 0.4. *Fix $\mathbf{x}_i \in \mathbb{R}^n$.*

(a) *The union $H(\mathbf{x}_i) \cup W_\uparrow(\mathbf{x}_i) \cup W_\downarrow(\mathbf{x}_i) = \mathbb{R}^n$.*

(b) *The intersection $H(\mathbf{x}_i) \cap W_\uparrow(\mathbf{x}_i) \cap W_\downarrow(\mathbf{x}_i) = \varnothing$.*

Definition 0.5. Fix $\mathbf{w} \in \mathbb{R}^n$. Define the geometric margin of $H(\mathbf{w})$ as

$$\gamma_{H(\mathbf{w})} = \min_{(\mathbf{x}_i, y_i) \in D} \|\mathbf{w}^T\mathbf{x}_i\|$$

Theorem 0.6. *If there exists some $\mathbf{w}^*$ with $y_i(\varphi(\mathbf{x}_i^T\mathbf{w}^*)) > 0$ for every choice of $(\mathbf{x}_i, y_i)$, then the perceptron learning algorithm converges in a finite number of steps.*

Proof. Fix $R > 0$, set $\|\mathbf{w}^*\| = R$ and constrain $\|\mathbf{x}_i\| \leq R$. Choose $\mathbf{w}$ with $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) \leq 0$. Then after k updates $\mathbf{w}^T\mathbf{w}^* \geq k\gamma_{H(\mathbf{w}^*)}$ and $\mathbf{w}^T\mathbf{w} \leq k$. It suffices to show that k is bounded above. By elementary algebra, $k \leq R/\gamma^2_{H(\mathbf{w}^*)}$. The result follows. □

Remark 0.7. This result of course depends on the existence of $\mathbf{w}^*$. In 1969, Minsky and Papert showed - among other things, that the perceptron could not classify datasets which are not linearly separable. Perhaps the most famous example of this is the XOR problem.

TRINITY GRAMMAR SCHOOL

Date: March 29, 2020.