GEOMETRICAL ANALYSIS OF THE PERCEPTRON

MATT RAYMOND

Lemma 0.1. *Define D to be the set of input-output pairs. We call the elements of D datapoints. 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 datapoint $(\mathbf{x}_i, y_i)$ 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 $H(\mathbf{x}_i) \cap W_\uparrow(\mathbf{x}_i) \cap W_\downarrow(\mathbf{x}_i) = \varnothing$.*

Theorem 0.5. *(Rosenblatt) 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.*

Definition 0.6. Let $W_\uparrow^*(D)$ be the intersection of sets $W_\uparrow(\mathbf{x}_i)$ such that every $\mathbf{w} \in W_\uparrow(\mathbf{x}_i)$ has $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i) > 0$. Similarly, let $W_\downarrow^*(D)$ be the intersection of sets $W_\downarrow(\mathbf{x}_i)$ such that every $\mathbf{w} \in W_\downarrow(\mathbf{x}_i)$ has $y_i(\varphi(\mathbf{w}^T\mathbf{x}_i)) > 0$.

Definition 0.7. Suppose D is a dataset. The set D is not linearly separable if there does not exist an $\mathbf{w}^* \in \mathbb{R}^n$ such that for each $(\mathbf{x}_i, y_i) \in D$, $y_i(\varphi(\mathbf{x}_i^T\mathbf{w}^*)) > 0$.

Lemma 0.8. *Let D be a dataset. Then the following are equivalent.*

(a) *The set D is not linearly separable.*

(b) *The set $W_\uparrow^*(D)$ fails to intersect $W_\downarrow^*(D)$.*

(c) *The learning algorithm does not converge in D.*

Proof. Suppose D is not linearly separable. Since $W_\uparrow^*(D) \cap W_\downarrow^*(D)$ is a subset of $\mathbb{R}^n$, it is clear that (a) implies (b). Suppose $W_\uparrow^*(D)$ and $W_\downarrow^*(D)$ do not intersect but D is linearly separable. Then there is some $\mathbf{w}^* \in \mathbb{R}^d$ such that for each $(\mathbf{x}_i, y_i) \in D$, $y_i(\varphi(\mathbf{x}_i^T\mathbf{w}^*)) > 0$. It follows that $W_\uparrow^*(D)$ and $W_\downarrow^*(D)$ intersect, which is a contradiction. Hence (b) implies (a). Suppose the learning algorithm converges in D. Then there is some $\mathbf{w}^*$ with $y_i(\varphi(\mathbf{x}_i^T\mathbf{w}^*)) > 0$ for every choice of $(\mathbf{x}_i, y_i)$. Then $\mathbf{w}^* \in W_\uparrow^*(D) \cap W_\downarrow^*(D)$, so (b) implies (c). The other direction is clear. This completes the proof. □

Example 0.9. Let $D = \{((-1,-1),-1),((1,-1),1),((1,1),-1),((-1,1),1)\}$. This is the XOR problem. It is clear that

$$W_\uparrow^*(D) \cap W_\downarrow^*(D) = W_\uparrow((1,-1)) \cap W_\uparrow((-1,1)) \cap W_\downarrow((-1,-1)) \cap W_\downarrow((1,1)) = \varnothing. \tag{0.10}$$

From the previous lemma, it follows that the learning algorithm does not converge in D.

Date: March 30, 2020.