Linear Classification: Perceptron

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong
Today, we start a series of lectures devoted to **linear classification**, which harbors a deep theory and is one of the most important topics in machine learning.
Let \( A_1, \ldots, A_d \) be \( d \) attributes, each with a domain \( \mathbb{R} \), i.e., \( \text{dom}(A_i) = \mathbb{R}^d \) for each \( i \in [1, d] \).

**Instance space:** \( \mathcal{X} = \text{dom}(A_1) \times \text{dom}(A_2) \times \ldots \times \text{dom}(A_d) = \mathbb{R}^d \).

**Label space:** \( \mathcal{Y} = \{-1, 1\} \) (where \(-1\) and \(1\) are class labels).

**Instance-label pair** (a.k.a. object): a pair \((x, y)\) in \( \mathcal{X} \times \mathcal{Y} \).

- \( x \) is a \( d \)-dimensional vector. Since every dimension has a real domain, we can regard \( x \) as a \( d \)-dimensional point.

- We use \( x[i] \) to represent the \( i \)-th coordinate of point \( x \).
**Linear Classification**

**Linear classifier:** A function $h : \mathcal{X} \rightarrow \mathcal{Y}$ where $h$ is defined by a $d$-dimensional **weight vector** $w$ such that

- $h(x) = 1$ if $x \cdot w \geq 0$ (note: “$\cdot$” represents dot product);
- $h(x) = -1$ otherwise.

Suppose that Alice chooses a linear classifier $h^*$ and a distribution $\mathcal{D}$ over $\mathcal{X}$ (note: $\mathcal{D}$ is defined in the instance space, not the instance-label space).

For any linear classifier $h$, its **error on** $\mathcal{D}$ is defined as:

$$\text{err}_\mathcal{D}(h) = \Pr_{x \sim \mathcal{D}}[h(x) \neq h^*(x)].$$

Note that the error of $h^*$ on $\mathcal{D}$ is 0.
Linear Classification

Alice provides a training set $S$ which contains objects $(x, y)$ obtained as follows:

- First, draw $x$ independently from $\mathcal{X}$.
- Then, set $y = h^*(x)$.

The goal of linear classification is to learn a classifier $h$ from $S$ whose error on $\mathcal{D}$ is as low as possible.
**Claim:** $S$ is linearly separable, i.e., there is a $d$-dimensional vector $w$ such that for each $p \in S$:

- $w \cdot p > 0$ if $p$ has label $1$;
- $w \cdot p < 0$ if $p$ has label $-1$.

The plane $w \cdot x = 0$ is a separation plane of $S$.

**Proof (sketch):** Denote by $w^*$ the weight vector of the classifier $h^*$ chosen by Alice. By the way $S$ is generated, we know that $w^* \cdot p \geq 0$ if $p$ has label $1$, and $w^* \cdot p < 0$ otherwise. We can obtain the desired $w$ by adjusting $w^*$ infinitesimally.
Example:

```
Linearly separable       Linearly non-separable
```

- Origin
In this lecture, we will study the following problem:

**Problem (Finding a Separation Plane):** Given a linearly separable set $S$, find a separation plane.

The separation plane gives a linear classifier $h$ with $err_S(h) = 0$, i.e., empirical error 0.

We will solve the problem with a surprisingly simple algorithm called perceptron.
Perceptron

The algorithm starts with \( \mathbf{w} = (0, 0, \ldots, 0) \) and, then, runs in iterations. In each iteration, it looks for a violation point \( p \in S \):

- If \( p \) has label 1, \( p \) is a violation point if \( \mathbf{w} \cdot p \leq 0 \);
- If \( p \) has label \(-1\), \( p \) is a violation point if \( \mathbf{w} \cdot p \geq 0 \);

If \( p \) exists, the algorithm adjusts \( \mathbf{w} \) as follows:

- If \( p \) has label 1, then \( \mathbf{w} \leftarrow \mathbf{w} + p \).
- If \( p \) has label \(-1\), then \( \mathbf{w} \leftarrow \mathbf{w} - p \).

The algorithm finishes when there are no more violation points.
Example: Suppose that $S$ has points: $p_1 = (1, 0)$, $p_2 = (0, -1)$, $p_3 = (0, 1)$, and $p_4 = (-1, 0)$. Points $p_1$ and $p_3$ have label 1, and the other have label $-1$.

The algorithm starts with $w = (0, 0, \ldots, 0)$.

- Iteration 1: $p_1$ is a violation point because it has label 1 but $p_1 \cdot w = 0$. Hence, we update $w$ to $w + p_1 = (1, 0)$.

- Iteration 2: $p_2$ is a violation point because it has label $-1$ but $p_2 \cdot w = 0$. Hence, we update $w$ to $w - p_2 = (1, 0) - (0, -1) = (1, 1)$.

- Iteration 3: No more violation points. The algorithm finishes with $w = (1, 1)$. 

We now analyze the number of iterations performed by Perceptron.

Given a vector \( \mathbf{v} = (v_1, \ldots, v_d) \), we define its length as

\[
|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{\sum_{i=1}^{d} v[i]^2}.
\]

For any vectors \( \mathbf{v}_1, \mathbf{v}_2 \), it holds that

\[
\mathbf{v}_1 \cdot \mathbf{v}_2 \leq |\mathbf{v}_1||\mathbf{v}_2|.
\]

Define:

\[
R = \max_{\mathbf{p} \in S} \{|\mathbf{p}|\}.
\]

In other words, all the points of \( S \) fall in a ball that centers at the origin and has radius \( R \).
Given a separation plane $\pi$, define its **margin** as the smallest distance from the points of $S$ to $\pi$.

**Example:**

Denote by $\gamma$ the **largest** margin of all the separation planes. Let $\pi^*$ be the origin-passing plane with margin $\gamma$; the plane has a **unit normal vector** $u^*$ such that

- for every $p \in S$ with label 1, $u^* \cdot p > 0$;
- for every $p \in S$ with label $-1$, $u^* \cdot p < 0$.

We have:

$$\gamma = \min_{p \in S} |u^* \cdot p|.$$
Theorem: Perceptron terminates after at most \((R/\gamma)^2\) adjustments of \(w\).

Proof: Let \(w_i\) \((i \geq 1)\) be the value of \(w\) after the \(i\)-th adjustment. As a special case, define \(w_0 = (0, ..., 0)\). Denote by \(k\) the total number of violations.
We first show that $\mathbf{w}_{i+1} \cdot \mathbf{u}^* \geq \mathbf{w}_i \cdot \mathbf{u}^* + \gamma$ for any $i \geq 0$. Consider the violation point $\mathbf{p}$ used to change $\mathbf{w}$ from $\mathbf{w}_i$ to $\mathbf{w}_{i+1}$:

- **Case 1:** $\mathbf{p}$ has label 1. Thus, $\mathbf{p} \cdot \mathbf{w}_i < 0$ and $\mathbf{w}_{i+1} = \mathbf{w}_i + \mathbf{p}$. Hence, $\mathbf{w}_{i+1} \cdot \mathbf{u}^* = \mathbf{w}_i \cdot \mathbf{u}^* + \mathbf{p} \cdot \mathbf{u}^*$. From the definition of $\gamma$, we know that $\mathbf{p} \cdot \mathbf{u}^* \geq \gamma$. This gives $\mathbf{w}_{i+1} \cdot \mathbf{u}^* \geq \mathbf{w}_i \cdot \mathbf{u}^* + \gamma$.

- **Case 2:** $\mathbf{p}$ has label $-1$. The proof is similar and left to you.

Therefore:

\[
\mathbf{w}_k \cdot \mathbf{u}^* \geq \mathbf{w}_{k-1} \cdot \mathbf{u}^* + \gamma \\
\geq \mathbf{w}_{k-2} \cdot \mathbf{u}^* + 2\gamma \\
\vdots \\
\geq \mathbf{w}_0 \cdot \mathbf{u}^* + k\gamma \\
= k\gamma.
\] (1)
Next, we show that $|\mathbf{w}_{i+1}|^2 \leq |\mathbf{w}_i|^2 + R^2$ for any $i \geq 0$. Consider the violation point $\mathbf{p}$ used to change $\mathbf{w}$ from $\mathbf{w}_i$ to $\mathbf{w}_{i+1}$:

- **Case 1:** $\mathbf{p}$ has label 1. Thus, $\mathbf{p} \cdot \mathbf{w}_i < 0$ and $\mathbf{w}_{i+1} = \mathbf{w}_i + \mathbf{p}$. Hence:

$$
|\mathbf{w}_{i+1}|^2 = \mathbf{w}_{i+1} \cdot \mathbf{w}_{i+1} = (\mathbf{w}_i + \mathbf{p}) \cdot (\mathbf{w}_i + \mathbf{p})
= \mathbf{w}_i \cdot \mathbf{w}_i + 2\mathbf{w}_i \cdot \mathbf{p} + |\mathbf{p}|^2
\quad \text{(by def. of } R) \leq |\mathbf{w}_i|^2 + 2\mathbf{w}_i \cdot \mathbf{p} + R^2
\leq |\mathbf{w}_i|^2 + R^2
$$

where the last step used the fact that $\mathbf{p} \cdot \mathbf{w}_i < 0$.

- **Case 2:** $\mathbf{p}$ has label $-1$. The proof is similar and left to you.

Therefore:

$$
|\mathbf{w}_k|^2 \leq |\mathbf{w}_{k-1}|^2 + R^2 \leq |\mathbf{w}_{k-2}|^2 + 2R^2 \ldots \leq |\mathbf{w}_0|^2 + kR^2 = kR^2. \quad (2)
$$
From (1), we know:

$$|w_k| = |w_k||u^*| \geq w_k \cdot u^* \geq k\gamma.$$ 

Therefore, $|w_k|^2 \geq k^2\gamma^2$. Comparing this to (2) gives:

$$kR^2 \geq k^2\gamma^2 \Rightarrow k \leq \frac{R^2}{\gamma^2}$$
We have learned how to obtain a linear classifier $h$ with 0 empirical error on $S$. Does $h$ have a small generalization error $err_D(h)$? The answer is yes, but this does not follow from the generalization theorem we currently have (think: why not?). In the next lecture, we will discuss a more powerful generalization theorem that will allow us to bound $err_D(h)$. 