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.

Linear Classification

Let $A_1, ..., A_d$ be d **attributes**, each with a domain $\mathbb{R}$, i.e., $dom(A_i) = \mathbb{R}^d$ for each $i \in [1, d]$.

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

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

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

- $\boldsymbol{x}$ is a d-dimensional vector. Since every dimension has a real domain, we can regard $\boldsymbol{x}$ as a d-dimensional point.
- We use $\boldsymbol{x}[i]$ to represent the i-th coordinate of point $\boldsymbol{x}$.

Linear Classification

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

- $h(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \cdot \boldsymbol{w} \geq 0$ (note: "$\cdot$" represents dot product);
- $h(\boldsymbol{x}) = -1$ 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:

$$err_{\mathcal{D}}(h) = \boldsymbol{Pr}_{\boldsymbol{x} \sim \mathcal{D}}[h(\boldsymbol{x}) \neq h^*(\boldsymbol{x})].$$

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

Linear Classification

Alice provides a **training set** S which contains objects $(\boldsymbol{x}, y)$ obtained as follows:

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

The goal of linear classification is to learn a classifier h from S whose error on $\mathcal{D}$ is as low as possible.

Linear Classification

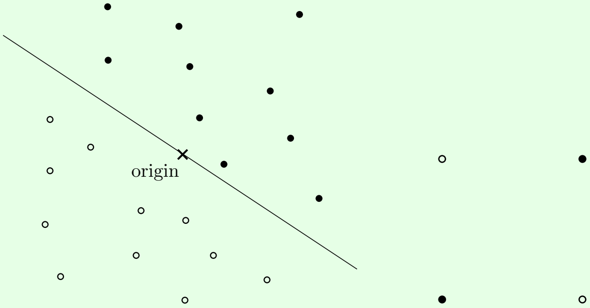
S is **linearly separable** if there is a d-dimensional vector $\boldsymbol{w}$ such that for each $\boldsymbol{p} \in S$:

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

The plane $\boldsymbol{w} \cdot \boldsymbol{x} = 0$ is a **separation plane** of S.

We will discuss only the scenario where S is linearly separable.

Example:



Linearly separable　　Linearly non-separable

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 $\boldsymbol{w} = (0, 0, ..., 0)$ and, then, runs in **iterations**.

In each iteration, it looks for a **violation point** $\boldsymbol{p} \in S$:

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

If $\boldsymbol{p}$ exists, the algorithm adjusts $\boldsymbol{w}$ as follows:

- If $\boldsymbol{p}$ has label 1, then $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{p}$.
- If $\boldsymbol{p}$ has label -1, then $\boldsymbol{w} \leftarrow \boldsymbol{w} - \boldsymbol{p}$.

The algorithm finishes when there are no more violation points.

Example: Suppose that S has points: $\boldsymbol{p}_1 = (1, 0)$, $\boldsymbol{p}_2 = (0, -1)$, $\boldsymbol{p}_3 = (0, 1)$, and $\boldsymbol{p}_4 = (-1, 0)$. Points $\boldsymbol{p}_1$ and $\boldsymbol{p}_3$ have label 1, and the other have label -1.

The algorithm starts with $\boldsymbol{w} = (0, 0, ..., 0)$.

- Iteration 1: $\boldsymbol{p}_1$ is a violation point because it has label 1 but $\boldsymbol{p}_1 \cdot \boldsymbol{w} = 0$. Hence, we update $\boldsymbol{w}$ to $\boldsymbol{w} + \boldsymbol{p}_1 = (1, 0)$.
- Iteration 2: $\boldsymbol{p}_w$ is a violation point because it has label -1 but $\boldsymbol{p}_2 \cdot \boldsymbol{w} = 0$. Hence, we update $\boldsymbol{w}$ to $\boldsymbol{w} - \boldsymbol{p}_2 = (1, 0) - (0, -1) = (1, 1)$.
- Iteration 3: No more violation points. The algorithm finishes with $\boldsymbol{w} = (1, 1)$.

We now analyze the number of iterations performed by Perceptron.

Given a vector $\boldsymbol{v} = (v_1, ..., v_d)$, we define its **length** as

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

For any vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$, it holds that $\boldsymbol{v}_1 \cdot \boldsymbol{v}_2 \leq |\boldsymbol{v}_1||\boldsymbol{v}_2|$.

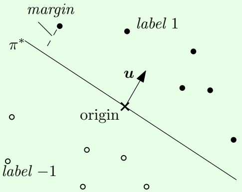
Define:

$$R = \max_{\boldsymbol{p} \in S}\{|\boldsymbol{p}|\}.$$

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 π, define its **margin** as the smallest distance from the points of S to π.

Example:



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

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

We have:

$$\gamma = \min_{\boldsymbol{p} \in S} |\boldsymbol{u}^* \cdot \boldsymbol{p}|.$$

Theorem: Perceptron terminates after at most $(R/\gamma)^2$ adjustments of $\boldsymbol{w}$.

Proof: Let $\boldsymbol{w}_i$ ($i \geq 1$) be the value of $\boldsymbol{w}$ after the i-th adjustment. As a special case, define $\boldsymbol{w}_0 = (0, ..., 0)$. Denote by k the total number of violations.

We first show that $\boldsymbol{w}_{i+1} \cdot \boldsymbol{u}^* \geq \boldsymbol{w}_i \cdot \boldsymbol{u}^* + \gamma$ for any $i \geq 0$. Consider the violation point $\boldsymbol{p}$ used to change $\boldsymbol{w}$ from $\boldsymbol{w}_i$ to $\boldsymbol{w}_{i+1}$:

- Case 1: $\boldsymbol{p}$ has label 1. Thus, $\boldsymbol{p} \cdot \boldsymbol{w}_i < 0$ and $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i + \boldsymbol{p}$. Hence, $\boldsymbol{w}_{i+1} \cdot \boldsymbol{u}^* = \boldsymbol{w}_i \cdot \boldsymbol{u}^* + \boldsymbol{p} \cdot \boldsymbol{u}^*$. From the definition of γ, we know that $\boldsymbol{p} \cdot \boldsymbol{u}^* \geq \gamma$. This gives $\boldsymbol{w}_{i+1} \cdot \boldsymbol{u}^* \geq \boldsymbol{w}_i \cdot \boldsymbol{u}^* + \gamma$.
- Case 2: $\boldsymbol{p}$ has label -1. The proof is similar and left to you.

Therefore:

$$\begin{aligned} \boldsymbol{w}_k \cdot \boldsymbol{u}^* &\geq \boldsymbol{w}_{k-1} \cdot \boldsymbol{u}^* + \gamma \\ &\geq \boldsymbol{w}_{k-2} \cdot \boldsymbol{u}^* + 2\gamma \\ &\ldots \\ &\geq \boldsymbol{w}_0 \cdot \boldsymbol{u}^* + k\gamma \\ &= k\gamma. \end{aligned} \tag{1}$$

Next, we show that $|\boldsymbol{w}_{i+1}|^2 \le |\boldsymbol{w}_i|^2 + R^2$ for any $i \ge 0$. Consider the violation point $\boldsymbol{p}$ used to change $\boldsymbol{w}$ from $\boldsymbol{w}_i$ to $\boldsymbol{w}_{i+1}$:

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

$$\begin{aligned} |\boldsymbol{w}_{i+1}|^2 = \boldsymbol{w}_{i+1} \cdot \boldsymbol{w}_{i+1} &= (\boldsymbol{w}_i + \boldsymbol{p}) \cdot (\boldsymbol{w}_i + \boldsymbol{p}) \\ &= \boldsymbol{w}_i \cdot \boldsymbol{w}_i + 2\boldsymbol{w}_i \cdot \boldsymbol{p} + |\boldsymbol{p}|^2 \\ \text{(by def. of } R\text{)} &\le |\boldsymbol{w}_i|^2 + 2\boldsymbol{w}_i \cdot \boldsymbol{p} + R^2 \\ &\le |\boldsymbol{w}_i|^2 + R^2 \end{aligned}$$

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

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

Therefore:

$$|\boldsymbol{w}_k|^2 \le |\boldsymbol{w}_{k-1}|^2 + R^2 \le |\boldsymbol{w}_{k-2}|^2 + 2R^2 ... \le |\boldsymbol{w}_0|^2 + kR^2 = kR^2. \quad (2)$$

From (1), we know:

$$|\boldsymbol{w}_k| = |\boldsymbol{w}_k||\boldsymbol{u}^*| \geq \boldsymbol{w}_k \cdot \boldsymbol{u}^* \geq k\gamma.$$

Therefore, $|\boldsymbol{w}_k|^2 \geq k^2\gamma^2$. Comparing this to (2) gives:

$$\begin{aligned} kR^2 &\geq k^2\gamma^2 \Rightarrow \\ k &\leq \frac{R^2}{\gamma^2} \end{aligned}$$

□

We have learned how to obtain a linear classifier h with 0 empirical error on S. Does h have a small generalization error $err_{\mathcal{D}}(h)$? 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 $err_{\mathcal{D}}(h)$.