Linear Classification: Maximizing the Margin

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

Recall:

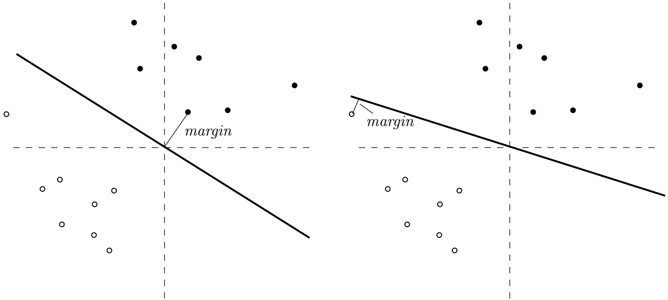
S is **linearly separable** if there is a d-dimensional vector $\boldsymbol{w}$ such that for each $\boldsymbol{p} \in S$:

- $\boldsymbol{w} \cdot \boldsymbol{p} > 0$ 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.

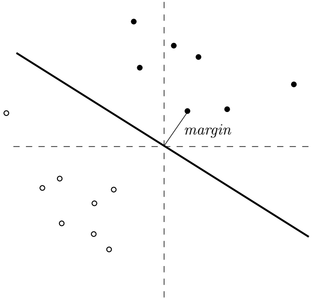
There can be many separation planes. As discussed previously, we should find the plane with the **largest margin**. In this lecture, we will discuss how to achieve the purpose.

Review: Margins



We prefer the left plane.

Let S be a linearly separable set of points in $\mathbb{R}^d$. In the **large margin separation problem**, we want to find a separation plane with the maximum margin.



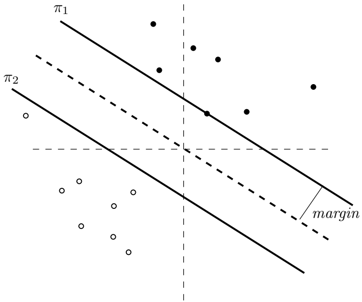
An algorithm solving this problem is called a **support vector machine**.

Next, we will discuss two methods. The first one finds the **optimal** solution but is computationally expensive. The second method is (much) faster but gives an **approximate** solution close to optimality.

Finding the Optimal Plane

We will model the problem as a **quadratic programing** problem.

Consider an arbitrary separation plane $\boldsymbol{w'} \cdot \boldsymbol{x} = 0$. Imagine two copies of the plane, one moving up and the other down, at the same speed. They stop as soon as a plane hits a point in S.



Now, focus on the two copies of the plane in their final positions. If one copy has equation $\boldsymbol{w'} \cdot x = \tau$, the other copy must have equation $\boldsymbol{w'} \cdot x = -\tau$, where $\tau > 0$.

For each point $p \in S$, we must have:

- if p has label 1, then $\boldsymbol{w'} \cdot \boldsymbol{p} \geq \tau$;
- if p has label -1, then $\boldsymbol{w'} \cdot \boldsymbol{p} \leq -\tau$.

By dividing τ on both sides of each inequality, we have:

- if p has label 1, then $\boldsymbol{w} \cdot \boldsymbol{p} \geq 1$;
- if p has label -1, then $\boldsymbol{w} \cdot \boldsymbol{p} \leq -1$

where

$$\boldsymbol{w} = \frac{\boldsymbol{w'}}{\tau}.$$

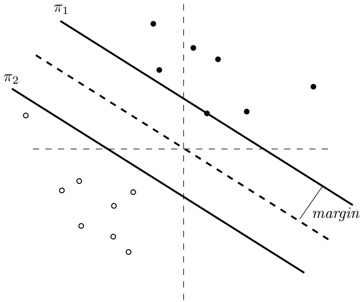
We will refer to the following plane as π_1

$$\boldsymbol{w} \cdot \boldsymbol{x} = 1$$

the following plane as π_2

$$\boldsymbol{w} \cdot \boldsymbol{x} = -1$$

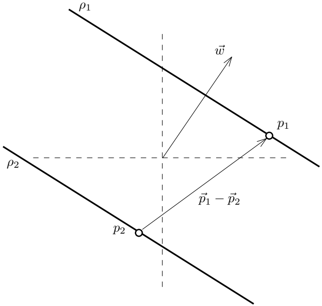
The margin of the original separation plane is exactly half of the distance between π_1 and π_2:



Lemma: The distance between π_1 and π_2 is $\frac{2}{|w|}$.

Hence, the margin of the separation plane $\boldsymbol{w} \cdot \boldsymbol{x} = 0$ is $\frac{1}{|w|}$.

Proof: Take an arbitrary point p_1 on π_1 and an arbitrary point p_2 on π_2. Hence, $\boldsymbol{w} \cdot \boldsymbol{p}_1 = 1$ and $\boldsymbol{w} \cdot \boldsymbol{p}_2 = -1$. It follows that $\boldsymbol{w} \cdot (\boldsymbol{p}_1 - \boldsymbol{p}_2) = 2$.



The distance between the two planes is precisely
$\frac{\boldsymbol{w}}{|\boldsymbol{w}|} \cdot (\boldsymbol{p}_1 - \boldsymbol{p}_2) = \frac{2}{|\boldsymbol{w}|}$. □

In summary of the above, to solve the large margin separation problem, we want to find $\boldsymbol{w}$ with the smallest $|\boldsymbol{w}|$, subject to:

- For each point $p \in S$ of label 1:

$$\boldsymbol{w} \cdot \boldsymbol{p} \geq 1$$

- For each point $p \in S$ of label -1:

$$\boldsymbol{w} \cdot \boldsymbol{p} \leq -1$$

This is an instance of **quadratic programming**.

In theory, the quadratic programming instance can be solved using convex-optimization techniques whose efficiency is rather difficult to analyze. We will not discuss this direction further.

Finding an Approximate Plane

Define γ_{opt} as the maximum margin of all separation planes.
A separation plane is **c-approximate** if its margin is at least $c \cdot \gamma_{opt}$.

We will give an algorithm to find a $(1/4)$-approximate separation plane.

Recall that a separation plane is given by $\boldsymbol{w} \cdot \boldsymbol{x} = 0$. The goal is to find a good $\boldsymbol{w}$.

Our weapon is once again Perceptron. But we will correct $\boldsymbol{w}$ **not only** when a point falls on the wrong side of the plane, **but also** when the point is too close to the plane.

For now, let us assume we are given an arbitrary value $\gamma_{guess} \leq \gamma_{opt}$. A point p causes a **violation** in any of the following situations:

- Its distance to the plane $\boldsymbol{w} \cdot \boldsymbol{x} = 0$ is less than $\gamma_{guess}/2$, regardless of its label.
- p has label 1 but $\boldsymbol{w} \cdot \boldsymbol{p} < 0$.
- p has label -1 but $\boldsymbol{w} \cdot \boldsymbol{p} > 0$.

Margin Perceptron

The algorithm starts with $\boldsymbol{w} = \mathbf{0}$ and runs in **iterations**.

In each iteration, it tries to find a **violation point** $p \in S$. If found, the algorithm adjusts $\boldsymbol{w}$ as follows:

- if p has label 1, $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{p}$.
- otherwise, $\boldsymbol{w} \leftarrow \boldsymbol{w} - \boldsymbol{p}$.

The algorithm finishes where no violation points are found.

Define $R = \max_{p \in S}\{|\boldsymbol{p}|\}$, i.e., the maximum distance from the origin to the points in S.

Theorem: If $\gamma_{guess} \leq \gamma_{opt}$, margin Perceptron terminates in at most

$$12R^2/\gamma_{opt}^2$$

iterations and returns a separation plane with margin at least $\gamma_{guess}/2$.

The proof can be found in the appendix.

Margin Perceptron requires a parameter $\gamma_{guess} \leq \gamma_{opt}$. By the theorem on the previous slide, a larger γ_{guess} promises a better quality guarantee.

Ideally, an ideal value for γ_{guess} is γ_{opt}, but unfortunately, we do not know γ_{opt}. Next, we present a strategy to estimate γ_{opt} up to a factor of $1/2$.

An Incremental Algorithm

1. $R \leftarrow$ the maximum distance from the origin to the points in S
2. $\gamma_{guess} \leftarrow R$
3. Run margin Perceptron with parameter γ_{guess}.
   - **[Self-Termination]**
     If the algorithm terminates with a plane π, return π as the final answer.
   - **[Forced-Termination]**
     If the algorithm has not terminated after $\frac{12R^2}{\gamma_{guess}^2}$ iterations:
     - Stop the algorithm manually.
     - Set $\gamma_{guess} \leftarrow \gamma_{guess}/2$.
     - Repeat Line 3.

Theorem: Our incremental algorithm returns a separation plane with margin at least $\gamma_{opt}/4$. Furthermore, it performs $O(R^2/\gamma_{opt}^2)$ iterations in total (including all the repeats at Line 3).

Proof: Suppose that we repeat Line 3 in total h times. For each $i \in [1, h]$, denote by γ_i the value of γ_{guess} at the i-th time we execute Line 3.

By the fact that the $(i-1)$-th repeat required a forced termination, we know that $\gamma_{h-1} > \gamma_{opt}$. Hence, $\gamma_h = \gamma_{h-1}/2 > \gamma_{opt}/2$. It thus follows that the plane we return must have a margin at least $\gamma_h/2 > \gamma_{opt}/4$.

The total number of iterations performed is

$$\begin{aligned} O\left(\sum_{i=1}^{h} \frac{R^2}{\gamma_i^2}\right) &= O\left(\frac{R^2}{\gamma_h^2} + \frac{R^2}{4\gamma_h^2} + \frac{R^2}{4^2\gamma_h^2} + \ldots\right) \\ &= O(R^2/\gamma_h^2) = O(R^2/\gamma_{opt}^2). \end{aligned}$$

$\square$

Appendix: Proof of the Theorem on Slide 18.

Let π^* be the the optimal plane with margin γ_{opt}.

Define $\boldsymbol{u}$ as the unit normal vector of π^* pointing to the positive side of π^*; in other words, we have:

- $|\boldsymbol{u}| = 1$.
- For every point $p \in S$ of label 1, $\boldsymbol{p} \cdot \boldsymbol{u} > 0$.
- For every point $p \in S$ label -1, $\boldsymbol{p} \cdot \boldsymbol{u} < 0$.
- $\gamma_{opt} = \min_{p \in S}\{|\boldsymbol{p} \cdot \boldsymbol{u}|\}$.

Recall that the perceptron algorithm adjusts $\boldsymbol{w}$ in each iteration. Let k be the total number of adjustments. Denote by $\boldsymbol{w}_i$ ($i \geq 1$) the value of $\boldsymbol{w}$ after the i-th adjustment; and define $\boldsymbol{w}_0 = (0, ..., 0)$.

Claim 1: $|\boldsymbol{w}_k| \geq \boldsymbol{w}_k \cdot \boldsymbol{u} \geq k\gamma_{opt}$.

Proof: We will first prove: for any $i \geq 0$, it holds that.

$$\boldsymbol{w}_{i+1} \cdot \boldsymbol{u} \geq \boldsymbol{w}_i \cdot \boldsymbol{u} + \gamma_{opt}. \tag{1}$$

Due to symmetry, we will prove the above only for the case where $\boldsymbol{w}_{i+1}$ is adjusted from $\boldsymbol{w}_i$ due to a violation point $\boldsymbol{p}$ of label 1. In this case, $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i + \boldsymbol{p}$; and hence, $\boldsymbol{w}_{i+1} \cdot \boldsymbol{u} = \boldsymbol{w}_i \cdot \boldsymbol{u} + \boldsymbol{p} \cdot \boldsymbol{u}$. From the definition of γ_{opt}, we know that $\boldsymbol{p} \cdot \boldsymbol{u} \geq \gamma_{opt}$, which gives (1).

It then follows from (1) that

$$\begin{aligned}
|\boldsymbol{w}_k| &\geq \boldsymbol{w}_k \cdot \boldsymbol{u} \\
&\geq \boldsymbol{w}_{k-1} \cdot \boldsymbol{u} + \gamma_{opt} \\
&\geq \boldsymbol{w}_{k-2} \cdot \boldsymbol{u} + 2\gamma_{opt} \\
&\ldots \\
&\geq \boldsymbol{w}_0 + k\gamma_{opt} = k\gamma_{opt}.
\end{aligned}$$

$\square$

Claim 2: $|\boldsymbol{w}_{i+1}| \le |\boldsymbol{w}_i| + R$.

Proof: We will prove only the case where $\boldsymbol{w}_{i+1}$ is adjusted from $\boldsymbol{w}_i$ using a violation point $\boldsymbol{p}$ of label 1. In this case:

$$|\boldsymbol{w}_{i+1}| = |\boldsymbol{w}_i + \boldsymbol{p}| \le |\boldsymbol{w}_i| + |\boldsymbol{p}| \le |\boldsymbol{w}_i| + R.$$

$\square$

Claim 3: $|\boldsymbol{w}_{i+1}| \leq |\boldsymbol{w}_i| + \frac{R^2}{2|\boldsymbol{w}_i|} + \frac{\gamma_{opt}}{2}$.

Proof: We will prove only the case where $\boldsymbol{w}_{i+1}$ is adjusted from $\boldsymbol{w}_i$ using a violation point $\boldsymbol{p}$ of label 1. In other words, $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i + \boldsymbol{p}$. Hence:

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

Since p is a violation point, it must hold that $\frac{\boldsymbol{w}_i}{|\boldsymbol{w}_i|} \cdot \boldsymbol{p} < \gamma_{guess}/2 \leq \gamma_{opt}/2$. Furthermore, obviously, $|\boldsymbol{p}|^2 \leq R^2$. We thus have:

$$|\boldsymbol{w}_{i+1}|^2 \leq |\boldsymbol{w}_i|^2 + \gamma_{opt}|\boldsymbol{w}_i| + R^2 \leq \left(|\boldsymbol{w}_i| + \frac{R^2}{2|\boldsymbol{w}_i|} + \frac{\gamma_{opt}}{2}\right)^2.$$

The claim then follows. $\square$

Claim 4: When $|\boldsymbol{w}_i| \geq \frac{2R^2}{\gamma_{opt}}$, $|\boldsymbol{w}_{i+1}| \leq |\boldsymbol{w}_i| + (3/4)\gamma_{opt}$.

Proof: Directly follows from Claim 3. $\square$

Claim 5: $|\boldsymbol{w}_k| \le \frac{2R^2}{\gamma_{opt}} + \frac{3k\gamma_{opt}}{4} + R$.

Proof: Let j be the largest i satisfying $|\boldsymbol{w}_i| < \frac{2R^2}{\gamma_{opt}}$. If $j = k$, then $|\boldsymbol{w}_k| < \frac{2R^2}{\gamma_{opt}}$, and we are done. Next, we focus on the case $j < k$; note that this means $|\boldsymbol{w}_{j+1}|, |\boldsymbol{w}_{j+2}|, ..., |\boldsymbol{w}_k|$ are all at least $2R^2/\gamma_{opt}$.

$$\begin{aligned}
|\boldsymbol{w}_k| &\le |\boldsymbol{w}_{k-1}| + (3/4)\gamma_{opt} && \text{(Claim 4)} \\
&\le |\boldsymbol{w}_{k-2}| + 2 \cdot (3/4)\gamma_{opt} && \text{(Claim 4)} \\
&\ldots \\
&\le |\boldsymbol{w}_{j+1}| + (k-j-1)(3/4)\gamma_{opt} && \text{(Claim 4)} \\
&\le |\boldsymbol{w}_{j+1}| + (3k/4)\gamma_{opt} \\
&\le |\boldsymbol{w}_j| + R + (3k/4)\gamma_{opt} && \text{(Claim 2)} \\
&\le \frac{2R^2}{\gamma_{opt}} + R + (3k/4)\gamma_{opt}.
\end{aligned}$$

□

Combining Claims 1 and 5 gives:

$$\begin{aligned}
k\gamma_{opt} &\leq \frac{2R^2}{\gamma_{opt}} + \frac{3k\gamma_{opt}}{4} + R \quad \Rightarrow \\
k &\leq \frac{8R^2}{\gamma_{opt}^2} + \frac{4R}{\gamma_{opt}} \\
(\text{by } R \geq \gamma_{opt}) &\leq \frac{8R^2}{\gamma_{opt}^2} + \frac{4R^2}{\gamma_{opt}^2} \\
&\leq \frac{12R^2}{\gamma_{opt}^2}.
\end{aligned}$$

This completes the proof of the theorem.