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 $w$ 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 $S$.

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.
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 $w' \cdot x = 0$. 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 $S$. 

![Diagram of linear classification with margin](image)
Now, focus on the two copies of the plane in their final positions. If one copy has equation $w' \cdot x = \tau$, the other copy must have equation $w' \cdot x = -\tau$, where $\tau > 0$.

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

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

By dividing $\tau$ on both sides of each inequality, we have:

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

where

$$w = \frac{w'}{\tau}.$$
We will refer to the following plane as $\pi_1$

$$w \cdot x = 1$$

the following plane as $\pi_2$

$$w \cdot x = -1$$

The margin of the original separation plane is exactly half of the distance between $\pi_1$ and $\pi_2$: 

![Diagram showing the margin between $\pi_1$ and $\pi_2$]
**Lemma:** The distance between $\pi_1$ and $\pi_2$ is $\frac{2}{|w|}$.

Hence, the margin of the separation plane $w \cdot x = 0$ is $\frac{1}{|w|}$. 
Proof: Take an arbitrary point \( p_1 \) on \( \pi_1 \) and an arbitrary point \( p_2 \) on \( \pi_2 \). Hence, \( \mathbf{w} \cdot \mathbf{p}_1 = 1 \) and \( \mathbf{w} \cdot \mathbf{p}_2 = -1 \). It follows that \( \mathbf{w} \cdot (\mathbf{p}_1 - \mathbf{p}_2) = 2 \).

The distance between the two planes is precisely
\[
\frac{\mathbf{w}}{|\mathbf{w}|} \cdot (\mathbf{p}_1 - \mathbf{p}_2) = \frac{2}{|\mathbf{w}|}.
\]
In summary of the above, to solve the large margin separation problem, we want to find $w$ with the smallest $|w|$, subject to:

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

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

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 $\gamma_{opt}$ 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 $w \cdot x = 0$. The goal is to find a good $w$.

Our weapon is once again Perceptron. But we will correct $w$ 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 $\gamma_{\text{guess}} \leq \gamma_{\text{opt}}$. A point $p$ causes a violation in any of the following situations:

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

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

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

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

The algorithm finishes where no violation points are found.
Define $R = \max_{p \in S} \{|p|\}$, i.e., the maximum distance from the origin to the points in $S$.

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

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

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

The proof can be found in the appendix.
Margin Perceptron requires a parameter $\gamma_{\text{guess}} \leq \gamma_{\text{opt}}$. By the theorem on the previous slide, a larger $\gamma_{\text{guess}}$ promises a better quality guarantee.

Ideally, an ideal value for $\gamma_{\text{guess}}$ is $\gamma_{\text{opt}}$, but unfortunately, we do not know $\gamma_{\text{opt}}$. Next, we present a strategy to estimate $\gamma_{\text{opt}}$ 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 $\gamma_{guess}$.

- **[Self-Termination]**
  If the algorithm terminates with a plane $\pi$, return $\pi$ 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 $\gamma_i$ the value of $\gamma_{guess}$ 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

$$O\left(\sum_{i=1}^{h} \frac{R^2}{\gamma_i^2}\right) = O\left(\frac{R^2}{\gamma_h^2} + \frac{R^2}{4\gamma_h^2} + \frac{R^2}{4^2\gamma_h^2} + \ldots\right) = O(R^2/\gamma_h^2) = O(R^2/\gamma_{opt}^2).$$
Appendix: Proof of the Theorem on Slide 18.
Let $\pi^*$ be the optimal plane with margin $\gamma_{opt}$.

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

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

Recall that the perceptron algorithm adjusts $w$ in each iteration. Let $k$ be the total number of adjustments. Denote by $w_i$ ($i \geq 1$) the value of $w$ after the $i$-th adjustment; and define $w_0 = (0, ..., 0)$. 

Yufei Tao
Linear Classification: Maximizing the Margin
Claim 1: $|w_k| \geq w_k \cdot u \geq k\gamma_{opt}$.

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

$$w_{i+1} \cdot u \geq w_i \cdot u + \gamma_{opt}.$$  \hspace{1cm} (1)

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

It then follows from (1) that

$$|w_k| \geq w_k \cdot u$$
$$\geq w_{k-1} \cdot u + \gamma_{opt}$$
$$\geq w_{k-2} \cdot u + 2\gamma_{opt}$$
$$\cdots$$
$$\geq w_0 + k\gamma_{opt} = k\gamma_{opt}.$$
Claim 2: $|w_{i+1}| \leq |w_i| + R$.

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

$$|w_{i+1}| = |w_i + p| \leq |w_i| + |p| \leq |w_i| + R.$$
**Claim 3:** $|\mathbf{w}_{i+1}| \leq |\mathbf{w}_i| + \frac{R^2}{2|\mathbf{w}_i|} + \frac{\gamma_{opt}}{2}$.

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

$$
|\mathbf{w}_{i+1}|^2 = \mathbf{w}_{i+1} \cdot \mathbf{w}_{i+1} = (\mathbf{w}_i + \mathbf{p})^2 = \mathbf{w}_i \cdot \mathbf{w}_i + 2\mathbf{w}_i \cdot \mathbf{p} + \mathbf{p} \cdot \mathbf{p} = |\mathbf{w}_i|^2 + 2\mathbf{w}_i \cdot \mathbf{p} + |\mathbf{p}|^2.
$$

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

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

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

Proof: Directly follows from Claim 3. 

\[\square\]
Claim 5: $|\mathbf{w}_k| \leq \frac{2R^2}{\gamma_{\text{opt}}} + \frac{3k\gamma_{\text{opt}}}{4} + R$.

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

\[
|\mathbf{w}_k| \leq |\mathbf{w}_{k-1}| + (3/4)\gamma_{\text{opt}} \quad \text{(Claim 4)}
\]
\[
\leq |\mathbf{w}_{k-2}| + 2 \cdot (3/4)\gamma_{\text{opt}} \quad \text{(Claim 4)}
\]
...  
\[
\leq |\mathbf{w}_{j+1}| + (k - j - 1)(3/4)\gamma_{\text{opt}} \quad \text{(Claim 4)}
\]
\[
\leq |\mathbf{w}_{j+1}| + (3k/4)\gamma_{\text{opt}}
\]
\[
\leq |\mathbf{w}_j| + R + (3k/4)\gamma_{\text{opt}} \quad \text{(Claim 2)}
\]
\[
\leq \frac{2R^2}{\gamma_{\text{opt}}} + R + (3k/4)\gamma_{\text{opt}}.
\]

Combining Claims 1 and 5 gives:

\[ k \gamma_{opt} \leq \frac{2R^2}{\gamma_{opt}} + \frac{3k\gamma_{opt}}{4} + R \quad \Rightarrow \]

\[ k \leq \frac{8R^2}{\gamma_{opt}^2} + \frac{4R}{\gamma_{opt}} \quad \text{(by } R \geq \gamma_{opt}) \leq \frac{8R^2}{\gamma_{opt}^2} + \frac{4R^2}{\gamma_{opt}^2} \leq \frac{12R^2}{\gamma_{opt}^2}. \]

This completes the proof of the theorem.