Perceptron with non-separable data

Recall the setup for learning with the Perceptron algorithm. An example is $X_i \in \mathbb{R}^n$ and its label $y_i \in \{-1, 1\}$. A hypothesis is specified by $w \in \mathbb{R}^n$ and it represents a linear threshold element:

$$w^T \cdot X_i \geq 0 \Rightarrow \text{predict } +$$
$$w^T \cdot X_i < 0 \Rightarrow \text{predict } -$$

Using the fact that $y_i \in \{-1, 1\}$ we get the following facts on the performance of $w$:

$$y_i(w^T \cdot X_i) \geq 0 \text{ no mistake}$$
$$y_i(w^T \cdot X_i) < 0 \text{ mistake}$$

In the previous lecture we required that data to be "separable with some margin". More specifically, we required that $y_i(u^T \cdot X_i) \geq \gamma$ for some $u$. That is, $u$ is correct on all the examples and it gets the label right with some distance from the zero threshold. The theorem fails when the data is not separable. Here we show that similar bounds on the number of mistakes can be derived even for non-separable data.

Let $w$ be any hypothesis. We use $\xi_i$ to denote how far $w$ is from achieving a margin of $\gamma$ on $X_i$.

$$\xi_i = \max \left\{ 0, \gamma - y_i (U^T \cdot X_i) \right\}.$$ 

This is illustrated in Figure 1. Notice that if $w$ is wrong on $X_i$ then $\xi_i > \gamma$. If $w$ is correct and has margin more than $\gamma$ then $\xi_i = 0$. If $w$ is correct but its margin is small then $0 \leq \xi_i \leq \gamma$. Given

![Figure 1: non-linear separable data.](image)
these we can define aggregation quantities specifying how much the dataset fails to be separated with margin $\gamma$.

$$D_1 = \sum_{all \ i} \xi_i$$

$$D_2^2 = \sum_{all \ i} \xi_i^2$$

**Upper Bound for Perceptron in Terms of $D_1$**

**Theorem 1** Let $U$, with $\|U\|^2 = 1$ be the hypothesis minimizing $D_1$. The Perception algorithm makes no more than $\left(\frac{R + \sqrt{\gamma D_1}}{\gamma}\right)^2$ mistakes on any sequence of examples satisfying $\|X_i\|^2 \leq R^2$ for all $i$.

*Proof:* The proof follows the same outline as the one from the previous lecture. Recall that we initialize $w^0 = \mathbf{0}$ and update $w$ only when making mistakes. We have $w^t \leftarrow w^{t-1} + y_iX_i$ where $w^t$ is $w$ after the $t$’th mistake and update. We first derive an upper bound on the norm of $w^t$:

$$\|w^t\|^2 = \|w^{t-1} + y_iX_i\|^2 = \|w^{t-1}\|^2 + y_i^2 \|X_i\|^2 + 2y_i \langle (w^{t-1})^T \cdot X_i \rangle$$

item $< 0$, because $y_i \neq \hat{y}_i$

$$\leq \|w^{t-1}\|^2 + R^2 = tR^2$$

We next derive a lower bound on the margin that $U$ has on any example:

$$\xi_i = \max \{0, \gamma - y_i (U^T \cdot X_i)\} \geq \gamma - y_i (U^T \cdot X_i)$$

$$\Rightarrow y_i (U^T \cdot X_i) \geq \gamma - \xi_i$$

We next derive a lower bound on the inner product:

$$w^T \cdot U = (w^{t-1} + y_iX_i)^T \cdot U$$

$$= w^{t-1}^T \cdot U + y_i (X_i^T \cdot U)$$

$$\geq w^{t-1}^T \cdot U + \gamma - \xi_i$$

$$\geq \sum_{i \in \text{mistakes}} (\gamma - \xi_i)$$

$$\geq \sum_{i \in \text{mistakes}} \gamma - \sum_{all \ i} \xi_i$$

$$= t\gamma - D_1$$
Finally, combining the two bounds we get:

\[
(t\gamma - D_1)^2 \leq (w_i^T \cdot U)^2 \\
\leq \|w^t\|^2 \|U\|^2 \leq \|w^t\|^2 \\
\leq tR^2
\]

\[
\Rightarrow \quad \gamma^2 t^2 - (2\gamma D_1 + R^2) t + D_1^2 \leq 0
\]

By solving the equation \(\gamma^2 t^2 - (2\gamma D_1 + R^2) t + D_1^2 = 0\) for \(t = 0\), we get:

\[
t \leq \frac{(2\gamma D_1 + R^2) + \sqrt{(2\gamma D_1 + R^2)^2 - 4\gamma^2 D_1^2}}{2\gamma^2}
\]

\[
\leq \left( \frac{R + \sqrt{\gamma D_1}}{\gamma} \right)^2
\]

**Upper Bound for Perceptron in Terms of \(D_2\)**

**Theorem 2** Let \(U\), with \(\|U\|^2 = 1\) be the hypothesis minimizing \(D_2\). The Perception algorithm makes no more than \(\left( \frac{R + D_2}{\gamma} \right)^2\) mistakes on any sequence of examples satisfying \(\|X_i\|^2 \leq R^2\) for all \(i\).

**Proof:** The theorem is proved by reduction to the separable case. We map the original space to a new space as follows:

\[
\begin{align*}
X_i &\rightarrow \tilde{X}_i = (X_i, 0, 0, ..., \Delta, ..., 0, 0, 0), \Delta \text{ in the } i_{th} \text{ position of the new features.} \\
U &\rightarrow \tilde{U} = (U, 0, 0, ..., \frac{y_i\xi_i}{\Delta}, ..., 0, 0, 0)
\end{align*}
\]

We first observe that in this new space the data is separable:

\[
y_i\tilde{U}^T \cdot \tilde{X}_i = y_i U^T \cdot X_i + y_i^2 \xi_i \geq \gamma
\]

Moreover, because each new feature is non-zero exactly once, when perceptron is run in the new space its predictions and mistakes are identical to the ones in the original space. Therefore a mistake bound for the new space holds in the original space as well. It remains to normalize \(U\) and calculate the relevant quantities. We have:
\[ \|\hat{X}_i\|^2 = \|X_i\|^2 + \Delta^2 \leq R^2 + \Delta^2 = \hat{R}^2 \]
\[ \|\hat{U}\|^2 = \|U\|^2 + \frac{D^2}{\Delta^2} = 1 + \frac{D^2}{\Delta^2} \]

Normalize \( \hat{U} \rightarrow \hat{\hat{U}} = \frac{\hat{U}}{\sqrt{1 + \frac{D^2}{\Delta^2}}} \)
\[ \Rightarrow \|\hat{\hat{U}}\|^2 = 1 \]

We finally get that in the new space, the assumptions in the previous lecture are satisfied. We therefore get the mistake bound:

\[ t \leq \hat{R}^2 \frac{\hat{\gamma}^2}{\gamma^2} = \frac{(R^2 + \Delta^2) \left(1 + \frac{D^2}{\Delta^2}\right)}{\gamma^2} \]

We can still choose \( \Delta \) to optimize the bound. Taking derivatives (detail omitted) we are led to pick \( \Delta = \sqrt{RD} \). Plugging in this value we get:

\[ t \leq \frac{(R + D)^2}{\gamma^2} \]

**The Winnow Algorithm**

The Winnow Algorithm and some of its associated mistake bound were discussed. These are provided in separate notes.