Behaviors of large-margin homogeneous linear classifiers on n points in $\mathbb{R}^d$

Daniel Hsu

December 7, 2023

Let $f_w \colon \mathbb{R}^d \to \{0,1\}$ denote the homogeneous linear classifier on $\mathbb{R}^d$ with weight vector $w \in \mathbb{R}^d$, given by

$$f_w(x) = \mathbb{1}\{\langle w, x\rangle > 0\}, \quad x \in \mathbb{R}^d.$$

Let $x^{(1)}, \dots, x^{(n)}$ be n feature vectors in $\mathbb{R}^d$, each with $\|x^{(i)}\| \leq 1$. For $\gamma > 0$, let $\mathcal{F}_\gamma$ denote the set of homogeneous linear classifiers on $\mathbb{R}^d$ with margin at least γ on the (unlabeled) dataset $x^{(1)}, \dots, x^{(n)}$, i.e.,

$$\mathcal{F}_\gamma = \{f_w : w \in \mathbb{R}^d \text{ with } \|w\| = 1 \text{ and } |\langle w, x^{(i)}\rangle| \geq \gamma \text{ for all } i = 1, \dots, n\}$$

The question we ask in this note is:

> What is the number of behaviors
>
> $$S(\mathcal{F}_\gamma; (x^{(i)})_{i=1}^n) = |\{(f_w(x^{(1)}), \dots, f_w(x^{(n)})) : f_w \in \mathcal{F}_\gamma\}|$$
>
> of $\mathcal{F}_\gamma$ on $x^{(1)}, \dots, x^{(n)}$?

For the purpose of upper-bounding $S(\mathcal{F}_\gamma; (x^{(i)})_{i=1}^n)$, let us define a function $T \colon \mathcal{F}_\gamma \to \mathbb{R}^d$ as follows. Given $f_w \in \mathcal{F}_\gamma$, define $T(f_w)$ to be the weight vector $\hat{w} \in \mathbb{R}^d$ returned by the Perceptron algorithm when run on the dataset $((x^{(i)}, f_w(x^{(i)})))_{i=1}^n$. In these executions of Perceptron, since the labels of the dataset are provided by some $f_w \in \mathcal{F}_\gamma$, Perceptron is guaranteed to halt within $1/\gamma^2$ updates and return a weight vector. Moreover, the linear classifier $f_{\hat{w}}$ associated with the weight vector $\hat{w} = T(f_w)$ is guaranteed to agree with f_w on all n feature vectors $x^{(1)}, \dots, x^{(n)}$. Therefore,

$$\begin{aligned} S(\mathcal{F}_\gamma; (x^{(i)})_{i=1}^n) &= |\{(f_{\hat{w}}(x^{(1)}), \dots, f_{\hat{w}}(x^{(n)})) : \hat{w} = T(f_w) \text{ for some } f_w \in \mathcal{F}_\gamma\}| \\ &\leq |\mathrm{range}(T)|. \end{aligned}$$

What is the cardinality of the range of T? A weight vector $\hat{w} \in \mathbb{R}^d$ returned by Perceptron on the dataset $((x^{(i)}, f_w(x^{(i)})))_{i=1}^n$ is obtained by starting with the zero vector and then adding at most $1/\gamma^2$ feature vectors from $\{\pm x^{(1)}, \ldots, \pm x^{(n)}\}$. So $\hat{w}$ must be of the form

$$\hat{w} = \sum_{t=1}^{k} \alpha_t \, x^{(i_t)}$$

for some $k \leq 1/\gamma^2$, some $\alpha_1, \ldots, \alpha_k \in \{-1, 1\}$, and some $i_1, \ldots, i_k \in \{1, \ldots, n\}$.[1] The number of vectors of this form is at most

$$\begin{aligned} \sum_{k=0}^{\lfloor 1/\gamma^2 \rfloor} (2n)^k &= (2n)^{\lfloor 1/\gamma^2 \rfloor} \sum_{k=0}^{\lfloor 1/\gamma^2 \rfloor} (2n)^{k - \lfloor 1/\gamma^2 \rfloor} \\ &= (2n)^{\lfloor 1/\gamma^2 \rfloor} \sum_{\ell=0}^{\lfloor 1/\gamma^2 \rfloor} (2n)^{-\ell} \\ &\leq \frac{(2n)^{\lfloor 1/\gamma^2 \rfloor}}{1 - (2n)^{-1}}. \end{aligned}$$

Therefore

$$S(\mathcal{F}_\gamma; (x^{(i)})_{i=1}^n) \leq \frac{(2n)^{\lfloor 1/\gamma^2 \rfloor}}{1 - (2n)^{-1}} = O(n^{1/\gamma^2}).$$

Notice that this bound is independent of the dimension d of the feature vectors. In fact, this argument works even when $d = \infty$.

[1] Not *all* vectors of this form might be returned by Perceptron, so we might be overcounting the range of T. This is okay because we are only trying to obtain an upper-bound on $|\mathrm{range}(T)|$.