Linear separators

Daniel Hsu

September 3, 2023

1 Linear separators

A dataset $\mathcal{S}$ from $\mathbb{R}^d \times \{-1, 1\}$ is *linearly separable* if there exists $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y(w^\top x + b) > 0 \quad \text{for all } (x, y) \in \mathcal{S}.$$

We use the output space $\mathcal{Y} = \{-1, 1\}$ (instead of $\{0, 1\}$) for notational convenience. The linear classifier determined by this weight vector w and intercept parameter b is called a *linear separator* for the dataset $\mathcal{S}$.

2 Approximate MLE for logistic regression

How can we find a linear separator for a linearly separable dataset $\mathcal{S}$? One approach is to find an approximate maximizer of the log-likelihood from the logistic regression model. Any algorithm that can find (w, b) with log-likelihood arbitrarily close to the maximum log-likelihood will do the job.

The log-likelihood of (w, b) given $\mathcal{S}$ in the logistic regression model is

$$\ln L(w, b) = \sum_{(x,y) \in \mathcal{S}} \ln\left(\frac{1}{1 + \exp(-y(w^\top x + b))}\right).$$

Notice that, in each term from the summation, the argument to the logarithm is strictly between 0 and 1, and hence the value of the logarithm is negative. This means that $\ln L(w, b) < 0$, regardless of the choice of (w, b).

However, if $\mathcal{S}$ is linearly separable, then it is possible to achieve log-likelihood arbitrarily close to 0. Suppose (w, b) determines a linear separator for $\mathcal{S}$. Then, for any $c > 0$, (cw, cb) also determines a linear separator for $\mathcal{S}$, because

$$y(w^\intercal x + b) > 0 \quad \Leftrightarrow \quad y((cw)^\intercal x + cb) > 0.$$

Moreover, by choosing c sufficiently large, we can make

$$y((cw)^\intercal x + cb)$$

an arbitrarily large positive number, which in turn makes

$$\frac{1}{1 + \exp(-y((cw)^\intercal x + cb))}$$

arbitrarily close to 1. Therefore, each term in the log-likelihood of (cw, cb) can be made arbitrarily close to 0, and hence the log-likelihood of (cw, cb) itself can be made arbitrarily close to 0. This means that

$$\max_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \ln L(w, b) = 0,$$

i.e., the maximum log-likelihood is 0.[1]

It remains to show that any (w, b) with log-likelihood sufficiently close to the maximum log-likelihood (which is 0) must determine a linear separator for $\mathcal{S}$. Suppose $\ln L(w, b) > -\ln(2)$. Then

$$\ln\left(\frac{1}{2}\right) < \ln L(w, b) \leq \ln\left(\frac{1}{1 + \exp(-y(w^\intercal x + b))}\right) \quad \text{for all } (x, y) \in \mathcal{S}.$$

This implies that

$$\frac{1}{1 + \exp(-y(w^\intercal x + b))} > \frac{1}{2} \quad \text{for all } (x, y) \in \mathcal{S},$$

which is the same as (w, b) determining a linear separator for $\mathcal{S}$.

3 Perceptron

Another algorithm for finding a linear separator for a linearly separable dataset $\mathcal{S}$ is the *Perceptron* algorithm.

[1]Technically, it is the *supremum* of the log-likehood that is 0. But we will ignore such technicalities, since real analysis is not a prerequisite for this class.

- Start with $w = 0$ and $b = 0$
- While there exists $(x, y) \in \mathcal{S}$ such that $y(x^\top w + b) \le 0$:
  - Let $(x, y) \in \mathcal{S}$ be any such example
  - Update (w, b):

$$\begin{aligned} w &\leftarrow w + yx \\ b &\leftarrow b + y \end{aligned}$$

- Return (w, b)

It is clear from the description of the algorithm that if (w, b) is returned, then it must be a linear separator for $\mathcal{S}$. On the other hand, it is not clear if the algorithm will terminate; even if it does, it is not clear how many updates are needed. So the rest of this section is devoted to addressing these concerns.

We assume that $\mathcal{S}$ is linearly separable, so let $(w^\star, b^\star)$ be the weight vector and intercept parameter that satisfy

$$y(x^\top w^\star + b^\star) > 0 \quad \text{for all } (x, y) \in \mathcal{S}.$$

Moreover, it will be helpful to define two additional parameters:

$$\begin{aligned} \gamma &= \min_{(x,y)\in\mathcal{S}} y(x^\top w^\star + b^\star), \\ R &= \max_{(x,y)\in\mathcal{S}} \|x\|. \end{aligned}$$

Consider a single update in the execution of Perceptron: let (w, b) be the parameters before the update, and let $(\tilde{w}, \tilde{b})$ be the parameters after the update. Let (x, y) be the example chosen for the update. Then

$$\begin{aligned} \tilde{w}^\top w^\star + \tilde{b}b^\star &= (w + yx)^\top w^\star + (b + y)b^\star \\ &= w^\top w^\star + bb^\star + y(x^\top w^\star + b^\star) \\ &\ge w^\top w^\star + bb^\star + \gamma \end{aligned}$$

where the inequality uses the definition of γ. Moreover,

$$\begin{aligned} \|\tilde{w}\|^2 + \tilde{b}^2 &= \|w + yx\|^2 + (b + y)^2 \\ &= \|w\|^2 + b^2 + 2y(x^\top w + b) + \|x\|^2 + 1 \\ &\le \|w\|^2 + b^2 + \|x\|^2 + 1 \\ &\le \|w\|^2 + b^2 + R^2 + 1 \end{aligned}$$

where the inequalities use the choice of (x, y) for the update and the definition of R.

Before any updates, we have

$$w^\intercal w^\star + bb^\star = 0$$

and

$$\|w\|^2 + b^2 = 0.$$

So after T updates, we are left with (w, b) satisfying

$$w^\intercal w^\star + bb^\star \geq T\gamma$$

and

$$\|w\|^2 + b^2 \leq T(R^2 + 1).$$

Also, by the Cauchy-Schwarz inequality,

$$w^\intercal w^\star + bb^\star \leq \sqrt{\|w\|^2 + b^2}\sqrt{\|w^\star\|^2 + (b^\star)^2}$$

Combining these last three inequalities gives

$$T\gamma \leq \sqrt{T(R^2 + 1)}\sqrt{\|w^\star\|^2 + (b^\star)^2},$$

which simplifies to

$$T \leq \frac{(R^2 + 1)(\|w^\star\|^2 + (b^\star)^2)}{\gamma^2}.$$

Since $(w^\star, b^\star)$ determines a linear separator for $\mathcal{S}$, it must be that $\gamma > 0$, so the upper-bound on T is finite. This implies that Perceptron will terminate after at most

$$\frac{(R^2 + 1)(\|w^\star\|^2 + (b^\star)^2)}{\gamma^2}$$

updates.