Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contributors:

This set of notes are based on contributions from many of graduate students, post-doctoral fellows and other collaborators. Here is a partial list:

Juncai He, Qingguo Hong, Li Jiang.....

Contents

1

Logistic Regression

In this chapter, we will mainly consider one statistic model, namely logistic regression, to define classifier for linearly separable sets. In the next chapter, we will study another statistical model, namely support vector machine (SVM). These two linear models form the foundation of deep learning, since the final fully connected output layer of a deep neural network is often give by one of these two linear classifiers. The general intuition is that linear classification models will work well when the different classes are approximately linearly separable. This assumption is made explicit for the SVM model, but the situation is a bit more subtle with the logistic regression.

1.1 Definition of linearly separable sets

In this section, we consider a special class of separable sets, namely linearly separable sets. Let us formally introduce the following definition.

1.1.1 Binary classification

For $k = 2$, there is a very simple geometric interpretation of two linearly separable sets.

Definition 1. *The two sets $A_1, A_2 \subset \mathbb{R}^d$ are linearly separable if there exists a hyperplane*

$$H_0 = \{x : wx + b = 0\}, \tag{1.1}$$

such that $wx + b > 0$ if $x \in A_1$ and $wx + b < 0$ if $x \in A_2$.

Lemma 1. *The two sets $A_1, A_2 \subset \mathbb{R}^d$ are linearly separable if there exists a hyperplane linearly separable if there exists*

$$W = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \in \mathbb{R}^{2\times d}, b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \in \mathbb{R}^{2\times d}, \tag{1.2}$$

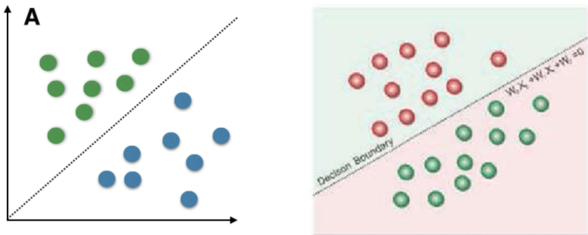

Fig. 1.1. Two linearly separable sets

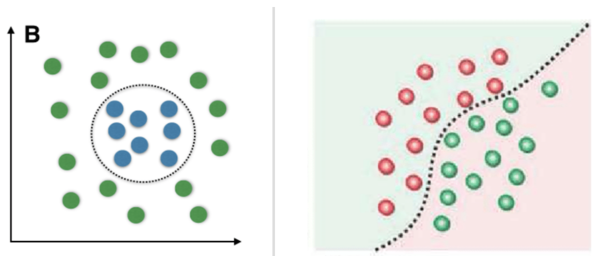

Fig. 1.2. Two Non-linearly separable sets

such that, for each $1 \leq i \leq 2$ *and* $j \neq i$

$$w_1 x + b_1 > w_2 x + b_2, \ \forall x \in A_1, \tag{1.3}$$

and

$$w_1 x + b_1 < w_2 x + b_2, \ \forall x \in A_2. \tag{1.4}$$

Proof. Here, we can just take $w = w_1 - w_2$ and $b = b_1 - b_2$, then we can check that the hyperplane $wx + b$ satisfies the definition as presented before. □

1.1.2 Multi-class classification

To begin with the definition, let us assume that the data space is divided into k classes represented by k disjoint sets $A_1, A_2, \cdots, A_k \subset \mathbb{R}^d$, which means

$$A = A_1 \cup A_2 \cup \cdots \cup A_k, \ A_i \cap A_j = \emptyset, \forall i \neq j. \tag{1.5}$$

Definition 2 (Linearly Separable). *A collection of subsets* $A_1, ..., A_k \subset \mathbb{R}^d$ *are linearly separable if there exist*

$$W = \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix} \in \mathbb{R}^{k\times d}, b = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} \in \mathbb{R}^{k\times d}, \tag{1.6}$$

such that, for each $1 \leq i \leq k$ *and* $j \neq i$

$$(Wx+b)_i > (Wx+b)_j, \ \forall x \in A_i, \tag{1.7}$$

or

$$w_i x + b_i > w_j x + b_j, \ \forall x \in A_i. \tag{1.8}$$

1.1.3 Geometric interpretation for multi-label cases ($k > 2$)

The geometric interpretation for linearly separable sets is less obvious when $k > 2$.

Lemma 2. *Assume that* $A_1, ..., A_k$ *are linearly separable and* $W \in \mathbb{R}^{k\times d}$ *and* $b \in \mathbb{R}^k$ *satisfy* (1.6). *Define*

$$\Gamma_i(W,b) = \{x \in \mathbb{R}^d : (Wx+b)_i > (Wx+b)_j, \ \forall j \neq i\} \tag{1.9}$$

Then for each i,

$$A_i \subset \Gamma_i(W,b) \tag{1.10}$$

We note that each $\Gamma_i(W,b)$ is a polygon whose boundary consists of hyperplanes

$$H_{ij} = \{(w_i - w_j) \cdot x + (b_i - b_j) = 0\}, \quad \forall j \neq i. \tag{1.11}$$

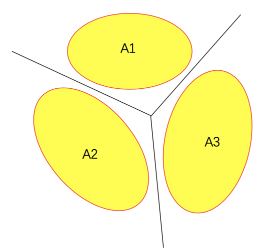

Fig. 1.3. Linearly separable sets in 2-d space (k = 3)

1.1.4 Two more definitions of linearly separable sets

We next introduce two more definitions of linearly separable sets that have more clear geometric interpretation.

Definition 3 (All-vs-One Linearly Separable). *A collection of subsets* $A_1, ..., A_k \subset \mathbb{R}^d$ *is all-vs-one linearly separable if for each* $i = 1, ..., k$, A_i *and* $\cup_{j\neq i} A_j$ *are linearly separable.*

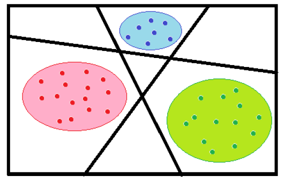

Fig. 1.4. All-vs-One linearly separable sets (k = 3)

Definition 4 (Pairwise Linearly Separable). *A collection of subsets* $A_1, ..., A_k \subset \mathbb{R}^d$ *is pairwise linearly separable if for each pair of indices* $1 \leq i < j \leq k$*,* A_i *and* A_j *are linearly separable.*

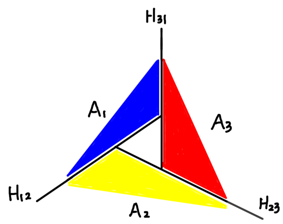

Fig. 1.5. Pairwise linearly separable sets in 2-d space (k = 3)

1.1.5 Comparison of different definitions of linearly separable sets

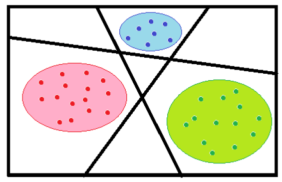

Fig. 1.6. All-vs-One linearly separable sets (k = 3)

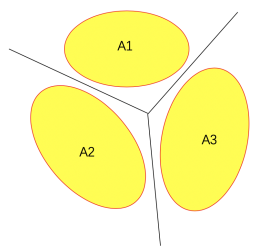

Fig. 1.7. Linearly separable sets in 2-d space (k = 3)

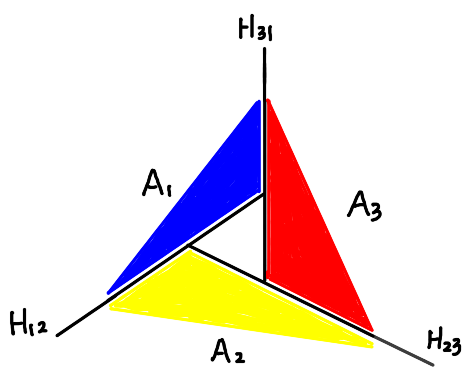

Fig. 1.8. Pairwise separable but not linearly separable sets

We begin by comparing our notion of linearly separable to the two other previously introduced geometric definitions of all-vs-one linearly separable and pairwise lineaerly separable. Obviously, in the case of two classes, they are all equivalent, however, with more than two classes this is no longer the case. We do have the following implications, though.

Lemma 3. *If $A_1, ..., A_k \subset \mathbb{R}^d$ are all-vs-one linearly separable, then they are linearly separable as well.*

Proof. Assume that $A_1, ..., A_k$ are all-vs-one linearly separable. For each i, let w_i, b_i be such that $w_i x + b_i$ separates A_i from $\cup_{j\neq i} A_j$, i.e. $w_i x + b_i > 0$ for $x \in A_i$ and $w_i x + b_i < 0$ for $x \in \cup_{j\neq i} A_j$.

Set $W = (w_1^T, w_2^T, \cdots, w_k^T)^T$, $b = (b_1, b_2, \cdots, b_k)^T$ and observe that if $x \in A_i$, then $(Wx + b)_i > 0$ while $(Wx + b)_j < 0$ for all $j \neq i$. □

Lemma 4. *If $A_1, ..., A_k \subset \mathbb{R}^n$ are linearly separable, then they are pairwise linearly separable as well.*

Proof. If $A_1, ..., A_k \subset \mathbb{R}^d$ are linearly separable, suppose that $W = (w_1^T, w_2^T, \cdots, w_k^T)^T$, $b = (b_1, b_2, \cdots, b_k)^T$. So we have

$$\begin{cases} w_i x + b_i > w_j x + b_j & x \in A_i \\ w_i x + b_i < w_j x + b_j & x \in A_j \end{cases} \tag{1.12}$$

Take $w_{i,j} = w_i - w_j, b_{i,j} = b_i - b_j$, then we have

$$w_{i,j} x + b_{i,j} \begin{cases} > 0 & x \in A_i \\ < 0 & x \in A_j \end{cases} \tag{1.13}$$

So $A_1, ..., A_k$ are pairwise linearly separable. □

However, the converses of both of these statements are false, as the following examples show.

Example 1 (Linearly separable but not all-vs-one linearly separable). Consider the sets $A_1, A_2, A_3 \subset \mathbb{R}$ given by $A_1 = [-4, -2]$, $A_2 = [-1, 1]$, and $A_3 = [2, 4]$. These sets are clearly not one-vs-all linearly separable because A_2 cannot be separated from both A_1 and A_3 by a single plane (in $\mathbb{R}$ this is just cutting the real line at a given number, and A_2 is in the middle).

However, these sets are linearly separated by $W = [-2, 0, 2]^T$ and $b = [-3, 0, -3]^T$, for example.

Example 2 (Pairwise linearly separable but not linearly separable). Consider the sets $A_1, A_2, A_3 \subset \mathbb{R}^2$ shown in figure 1.9. Note that A_i and A_j are separated by hyperplane $H_{i,j}$ (drawn in the figure) and so these sets are pairwise linearly separable. We will show that they are not linearly separable.

Assume to the contrary that $W \in \mathbb{R}^{3\times 2}$ and $b \in \mathbb{R}^2$ separates A_1, A_2, and A_3. Then $(w_i - w_j)x + (b_i - b_j)$ must be a plane which separates A_i and A_j. Now consider the point z in figure 1.9. We see from the figure that given any plane separating A_1 from A_2, z must be on the same side as A_2, given any plane separating A_2 from A_3, z must be on the same side as A_3, and given any plane separating A_3 from A_1, z must be on the same side as A_1.

This means that $(w_2 - w_1)z + (b_2 - b_1) > 0$, $(w_3 - w_2)z + (b_3 - b_2) > 0$, and $(w_1 - w_3)z + (b_1 - b_3) > 0$. Adding these together, we obtain $0 > 0$, a contradiction.

The essence behind this example is that although the sets A_1, A_2, and A_3 are pairwise linearly separable, no possible pairwise separation allows us to consistently classify arbitrary new points. However, a linear separation would give us a consistent scheme for classifying new points.

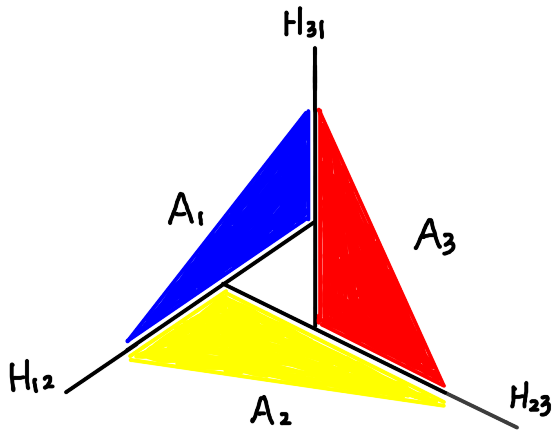

Fig. 1.9. Pairwise separable but not linearly separable sets

So the notion of linear separability is sandwiched in between the more intuitive notions of all-vs-one and pairwise separability. It turns out that linear separability is the notion which is most useful for the k-class classification problem and so we focus on this notion of separability from now on.