Title: Linear separability

URL: https://en.wikipedia.org/wiki/Linear_separability

PageID: 523173

Categories: Category:Convex analysis, Category:Geometry, Category:Machine learning

Source: Wikipedia (CC BY-SA 4.0).

In Euclidean geometry , linear separability is a property of two sets of points . This is most easily visualized in two dimensions (the Euclidean plane) by thinking of one set of points as being colored blue and the other set of points as being colored red. These two sets are linearly separable if there exists at least one line in the plane with all of the blue points on one side of the line and all the red points on the other side. This idea immediately generalizes to higher-dimensional Euclidean spaces if the line is replaced by a hyperplane .

The problem of determining if a pair of sets is linearly separable and finding a separating hyperplane if they are, arises in several areas. In statistics and machine learning , classifying certain types of data is a problem for which good algorithms exist that are based on this concept.

Mathematical definition

Let X ⊂ R d {\displaystyle X\subset \mathbb {R} ^{d}} be a set of m {\displaystyle m} points and Y ⊂ R d {\displaystyle Y\subset \mathbb {R} ^{d}} be a set of n {\displaystyle n} points in a d {\displaystyle d} -dimensional Euclidean space . X {\displaystyle X} and Y {\displaystyle Y} are linearly separable if they can be "separated" by a d {\displaystyle d} -dimensional hyperplane such that every point in X {\displaystyle X} lies on one side of the hyperplane and every point in Y {\displaystyle Y} lies on the other side.

The separating hyperplane is composed of points { z ∈ R d : w ■ z + k = 0 } {\displaystyle \left\{z\in \mathbb {R} ^{d}:w^{\top }z+k=0\right\}} , where w ∈ R d {\displaystyle w\in \mathbb {R} ^{d}} is the normal vector to the hyperplane and k ∈ R {\displaystyle k\in \mathbb {R} } is a scalar offset. X {\displaystyle X} and Y {\displaystyle Y} are linearly separable if there exists a normal vector w {\displaystyle w} and a scalar offset k {\displaystyle k} such that either every point x ∈ X {\displaystyle x\in X} satisfies w ■ x + k > 0 {\displaystyle w^{\top }x+k>0} and every point y ∈ Y {\displaystyle y\in Y} satisfies w ■ y + k < 0 {\displaystyle w^{\top }y+k<0} , or every point x ∈ X {\displaystyle x\in X} satisfies w ■ x + k < 0 {\displaystyle w^{\top }x+k<0} and every point y ∈ Y {\displaystyle y\in Y} satisfies w ■ y + k > 0 {\displaystyle w^{\top }y+k>0} .

Equivalently, two sets are linearly separable precisely when their respective convex hulls are disjoint (colloquially, do not overlap). [1]

Examples

Three non- collinear points in two classes ('+' and '-') are always linearly separable in two dimensions. This is illustrated by the three examples in the following figure (the all '+' case is not shown, but is similar to the all '-' case):

However, not all sets of four points, no three collinear, are linearly separable in two dimensions. The following example would need two straight lines and thus is not linearly separable:

Notice that three points which are collinear and of the form "+ ⋅⋅⋅ — ⋅⋅⋅ +" are also not linearly separable.

Number of linear separations

Let T (N , K) {\displaystyle T(N,K)} be the number of ways to linearly separate N points (in general position) in K dimensions, then [2] T (N , K) = { 2 N K ≥ N 2 ∑ k = 0 K − 1 (N − 1 k) K < N {\displaystyle T(N,K)=\left\{{\begin{array}{cc}2^{N}&K;\geq N\\2\sum _{k=0}^{K-1}\left({\begin{array}{c}N-1\\k\end{array}}\right)&K; 2 K {\displaystyle N>2K} . In words, one perceptron unit can almost certainly memorize a random assignment of binary labels on N points when N ≤ 2 K {\displaystyle N\leq 2K} , but almost certainly not when N > 2 K {\displaystyle N>2K} .

Linear separability of Boolean functions in n variables

A Boolean function in n variables can be thought of as an assignment of 0 or 1 to each vertex of a Boolean hypercube in n dimensions. This gives a natural division of the vertices into two sets. The Boolean function is said to be linearly separable provided these two sets of points are linearly separable. The number of distinct Boolean functions is 2 2 n {\displaystyle 2^{2^{n}}} where n is the number of variables passed into the function. [3]

Such functions are also called linear threshold logic, or perceptrons . The classical theory is summarized in, [4] as Knuth claims. [5]

The value is only known exactly up to n = 9 {\displaystyle n=9} case, but the order of magnitude is known quite exactly: it has upper bound 2 n 2 – n log 2 ■ n + O (n) {\displaystyle 2^{n^{2}-n\log _{2}n+O(n)}} and lower bound 2 n 2 – n log 2 ■ n – O (n) {\displaystyle 2^{n^{2}-n\log _{2}n-O(n)}} . [6]

It is co-NP-complete to decide whether a Boolean function given in disjunctive or conjunctive normal form is linearly separable. [6]

Threshold logic

A linear threshold logic gate is a Boolean function defined by n {\displaystyle n} weights w 1 , … , w n {\displaystyle w_{1},\dots ,w_{n}} and a threshold θ {\displaystyle \theta } . It takes n {\displaystyle n} binary inputs x 1 , … , x n {\displaystyle x_{1},\dots ,x_{n}} , and outputs 1 if ∑ i w i x i > θ {\displaystyle \sum _{i}w_{i}x_{i}>\theta } , and otherwise outputs 0.

For any fixed n {\displaystyle n} , because there are only finitely many Boolean functions that can be computed by a threshold logic unit, it is possible to set all w 1 , … , w n , θ {\displaystyle w_{1},\dots ,w_{n},\theta } to be integers. Let W (n) {\displaystyle W(n)} be the smallest number W {\displaystyle W} such that every possible real threshold function of n {\displaystyle n} variables can be realized using integer weights of absolute value ≤ W {\displaystyle \leq W} . It is known that [8] 1 2 n log ■ n – 2 n + o (n) ≤ log 2 ■ W (n) ≤ 1 2 n log ■ n – n + o (n) {\displaystyle {\frac {1}{2}}n\log n-2n+o(n)\leq \log _{2}W(n)\leq {\frac {1}{2}}n\log n-n+o(n)} See [9] : Section 11.10 for a literature review.

Support vector machines

Classifying data is a common task in machine learning .

Suppose some data points, each belonging to one of two sets, are given and we wish to create a model that will decide which set a new data point will be in. In the case of support vector machines , a data point is viewed as a p -dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a (p – 1)-dimensional hyperplane . This is called a linear classifier . There are many hyperplanes that might classify (separate) the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two sets. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier .

More formally, given some training data D {\displaystyle {\mathcal {D}}} , a set of n points of the form

where the y i is either 1 or −1, indicating the set to which the point x i {\displaystyle \mathbf {x} _{i}} belongs. Each x i {\displaystyle \mathbf {x} _{i}} is a p -dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having y i = 1 {\displaystyle y_{i}=1} from those having y i = – 1 {\displaystyle y_{i}=-1} . Any hyperplane can be written as the set of points x {\displaystyle \mathbf {x} } satisfying

where · {\displaystyle \cdot } denotes the dot product and w {\displaystyle {\mathbf {w} }} the (not necessarily normalized) normal vector to the hyperplane. The parameter b ■ w ■ {\displaystyle {\tfrac {b}{\|\mathbf {w} \|}}} determines the offset of the hyperplane from the origin along the normal vector w {\displaystyle {\mathbf {w} }} .

If the training data are linearly separable, we can select two hyperplanes in such a way that they separate the data and there are no points between them, and then try to maximize their distance.

See also

Clustering (statistics)

Hyperplane separation theorem

Kirchberger's theorem

Perceptron

Vapnik–Chervonenkis dimension

References