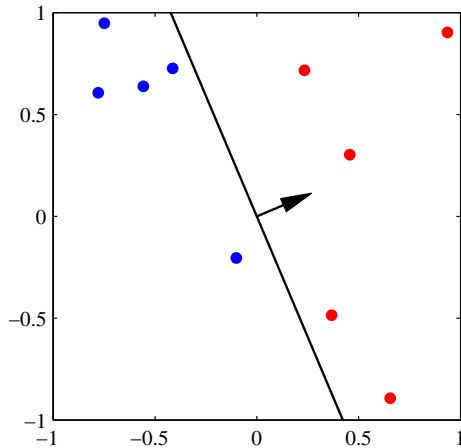


Linear classifiers

The whole idea behind linear classifiers is to directly represent the *decision* boundary (in binary classification problems). In decision trees and KNN, it is *implicit*.

We begin with the simplest kind of decision boundary: a hyperplane (a flat thing). We put points of each class on either side of the boundary. The mostly-vertical line in the figure below is a possible decision boundary between the blue (left) and red (right) points.



As a *representation* of a decision boundary, we store the vector that is orthogonal (aka “normal”) to the decision boundary. We usually call this \mathbf{w} (for “weights”... this will become clear soon). For instance, if the x-axis is the first coordinate and the y-axis is the second coordinate, the weight vector in the figure above (the arrow) might have value $\langle 0.4, 0.1 \rangle$, indicating that it is pointing right a lot and up a little.

The representation in terms of \mathbf{w} is overkill: if \mathbf{w} represents the hyperplane, then $\alpha\mathbf{w}$ also does, for any $\alpha > 0$ (it just makes the vector longer, but doesn’t change its direction). Thus, we’ll usually assume that $\|\mathbf{w}\| = 1$ (the 1 is for convenience; any positive constant would be fine).

This representation also assumes that the decision boundary passes through the origin. We will shift the hyperplane by adding an additional “bias” term b . In the above figure, if we wished to move the decision boundary in the direction of the arrow, we would use $b < 0$; to move it in the opposite direction, we use $b > 0$.

Given a representation of the decision boundary in terms of \mathbf{w} and b , we make our classification according to $\hat{y} = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$, where \mathbf{x} is an input and \hat{y} is the predicted class (note that we refer to our two classes as +1 and -1 instead of 0 and 1).

We can think about the hyperplane as being defined as the set $\{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + b = 0\}$... in other words, it’s the set of *all* points for which our classifier is totally uncertain.

In general, if there exists a hyperplane that perfectly separates our data onto two sides, we refer to the data as *linearly separable*. Except in very peculiar circumstances, if data is linearly separable, then there are infinitely many separating hyperplanes (just jiggle the one in the figure a little bit any direction).

We usually want to pick out the “best” hyperplane and the criteria that is often chosen is the *maximal margin hyperplane*, which is essentially the one with the most “wiggle room.” Formally, the *margin* of a hyperplane (denoted γ is the distance to the closest point): $\gamma = \min_n y_n (\mathbf{w}^\top \mathbf{x}_n + b)$.