Machine Learning

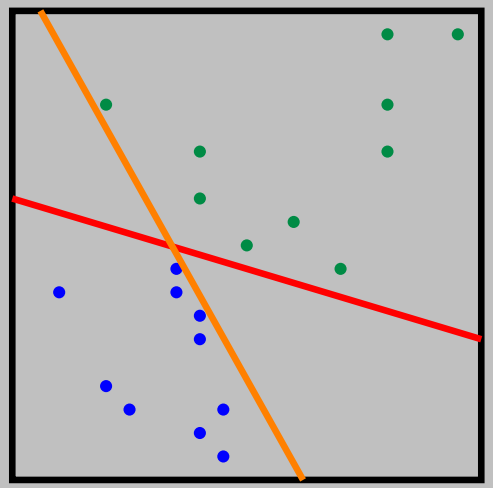
Support Vector Classifier

Irving Gómez Méndez

August-December, 2021

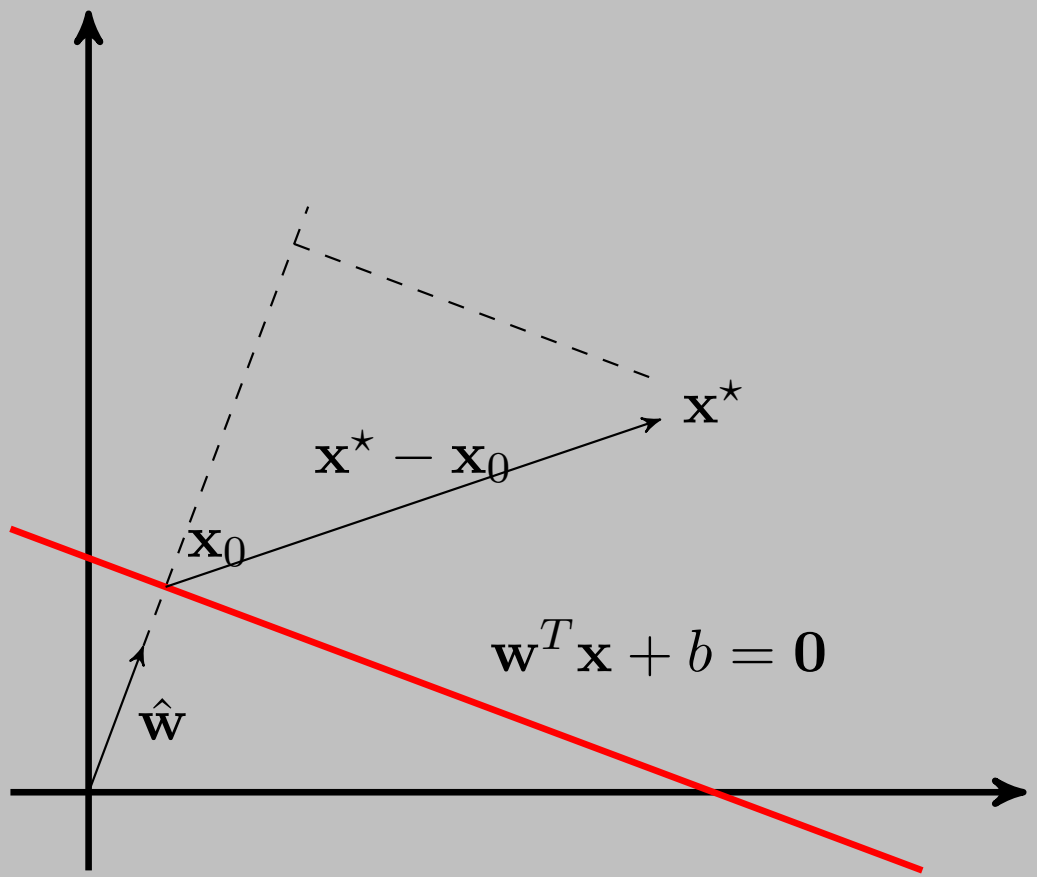
Separating Hyperplanes

These procedures construct linear decision boundaries that explicitly try to separate the data into different classes as well as possible. The next figure shows 20 data points in two classes. These data can be separated by a linear boundary. Included in the figure are two of the infinitely many possible *separating hyperplanes*. The idea is to find the *optimal separating hyperplane*.



Signed Distance

Let be $D(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, then the hyperplane is given by the equation $D(\mathbf{x}) = 0$. The figure depicts an *affine set* L defined by this equation; since we are in $\mathbb{R}^2$ this is a line.



- $L = \{\mathbf{x} \text{ such that } \mathbf{w}^T\mathbf{x} + b = \mathbf{0}\}$.
- For any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ lying in L, $\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$, hence $\hat{\mathbf{w}} = \mathbf{w}/\left\|\mathbf{w}\right\|$ is the vector normal to the surface of L.
- For any point $\mathbf{x}_0$ on L, $\mathbf{w}^T\mathbf{x}_0 = -b$.
- The signed distance of any point $\mathbf{x}^\star$ to L is given by.

$$\begin{aligned}\langle \hat{\mathbf{w}}, (\mathbf{x}^\star - \mathbf{x})\rangle &= \mathbf{w}^{\star T}(\mathbf{x}^\star - \mathbf{x}_0) \\ &= \frac{1}{\|\mathbf{w}\|}(\mathbf{w}^T\mathbf{x}^\star - b) \\ &\propto D(\mathbf{x}^\star)\end{aligned}$$

Hence, $D(\mathbf{x})$ is proportional to the signed distance from $\mathbf{x}$ to the hyperplane defined by $D(\mathbf{x}) = 0$.

Given n training observations $\mathbf{x}_i$ in a p-dimensional space, which belong to one of two classes, labeled as $y_i = 1$ or $y_i = -1$, for classes 1 or -1, respectively, the data is said to be linearly separable if it is possible to determine a decision function of the form:

$$D(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

and for $i = 1, \ldots, n$ we have

$$\mathbf{w}^T\mathbf{x}_i + b \begin{cases} \geq 0 & \text{for} \quad y_i = 1 \\ \leq 0 & \text{for} \quad y_i = -1 \end{cases}$$

If the training data are linearly separable, it is possible to assume that no observation satisfies $\mathbf{w}^T\mathbf{x} + b = 0$. Then, to control the separability, instead we consider

$$\mathbf{w}^T\mathbf{x}_i + b \begin{cases} \geq M & \text{for} \quad y_i = 1 \\ \leq -M & \text{for} \quad y_i = -1 \end{cases}$$

The *optimal separating hyperplane* separates the two classes and maximizes the distance to the closest point from either class. Hence, we consider the optimization problem

$$\underset{\mathbf{w},b,\|\mathbf{w}\|=1}{\text{maximize}} \frac{1}{2}M^2$$
$$\text{subject to } y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq M, \quad i = 1, \ldots, n$$

The set of conditions ensure that all the points are at least signed distance M from the decision boundary, and we seek the largest such M and associated parameters $(\mathbf{w}, b)$. We can get rid of the $\|\mathbf{w}\| = 1$ constraint by replacing the conditions with

$$\frac{1}{\|\mathbf{w}\|} y_i(\mathbf{x}_i^T \mathbf{w} + b) \geq M,$$

or equivalently

$$y_i(\mathbf{x}_i^T \mathbf{w} + b) \geq \|\mathbf{w}\| M.$$

Since for any $\mathbf{w}$ and b satisfying these inequalities, any positively scaled multiple satisfies them too, we can arbitrarily set $\|\mathbf{w}\| = 1/M$.

Thus the optimization problem is equivalent to

$$\underset{\mathbf{w},b}{\text{minimize}} \frac{1}{2} \|\mathbf{w}\|^2$$
$$\text{subject to } y_i(\mathbf{x}_i^T \mathbf{w} + b) \geq 1, \quad i = 1, \ldots, n$$

The constraints define an empty margin around the linear decision boundary of thickness $1/\|\mathbf{w}\|$. Hence, we choose $\mathbf{w}$ and b to maximize its thickness.

We can rewrite the optimization into the more usual form

$$\underset{\mathbf{w},b}{\text{minimize}}\, \frac{1}{2}\mathbf{w}^T\mathbf{w} \equiv f(\mathbf{w},b)$$

$$\text{subject to } g_i(\mathbf{w},b) \equiv 1 - y_i(\mathbf{x}_i^T\mathbf{w} + b) \leq 0, \quad i = 1,\ldots,n$$

The Lagrange function is given by

$$\begin{aligned} L(\mathbf{w},b,\boldsymbol{\alpha}) &= f(\mathbf{w},b) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w},b) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \alpha_i [1 - y_i(\mathbf{x}_i^T\mathbf{w} + b)] \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i b \end{aligned}$$

Let $\mathbf{w}^\star$, $b^\star$ and $\boldsymbol{\alpha}^\star$ be the optimal parameters, this parameters must satisfy

$$\left.\frac{\partial L}{\partial \mathbf{w}}\right|_{(\mathbf{w}^\star, b^\star, \boldsymbol{\alpha}^\star)} = 0 \Rightarrow \mathbf{w}^\star = \sum_{i=1}^{n} \alpha_i^\star y_i \mathbf{x}_i$$

$$\left.\frac{\partial L}{\partial b}\right|_{(\mathbf{w}^\star, b^\star, \boldsymbol{\alpha}^\star)} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i^\star y_i = 0$$

To find $\boldsymbol{\alpha}$, we use the so-called Wolfe dual, given by

$$\underset{\boldsymbol{\alpha}}{\text{maximize}}\, f(\mathbf{w}^\star(\alpha), b^\star(\alpha) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w}^\star(\alpha), b^\star(\alpha)))$$

$$\text{subject to } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, \quad i = 1, \ldots, n$$

That is,

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, \quad i = 1, \ldots, n$$

To explicitly calculate $b^\star$, we use the remaining Karush-Kuhn-Tucker condition

$$\alpha_i^\star g_i(\mathbf{w}^\star, b^\star) = 0 \quad i = 1, \ldots, n$$

that is,

$$\alpha_i^\star [y_i(\mathbf{x}_i^T \mathbf{w}^\star + b^\star) - 1] = 0 \quad i = 1, \ldots, n$$

- If $\alpha_i^\star > 0$, then $y_i(\mathbf{x}_i^T\mathbf{w}^\star + b^\star) = 1$, or in other words, $\mathbf{x}_i$ is on the boundary of the slab. These points are known as *support points*.
- If $y_i(\mathbf{x}_i^T\mathbf{w}^\star + b^\star) > 1$, $\mathbf{x}_i$ is not on the boundary of the slab, and $\alpha_i^\star = 0$

To calculate $b^\star$ we can pick any support point $(\mathbf{x}_j, y_j)$ and compute

$$b^\star = \frac{1}{y_j} - \mathbf{x}_j^T\mathbf{w}^\star$$

The optimal separating hyperplane produces a function $D^\star(\mathbf{x})$ for classifying new observations:

$$\widehat{G}(\mathbf{x}) = \operatorname{sign} D^\star(\mathbf{x})$$