Support vector machines (SVMs)

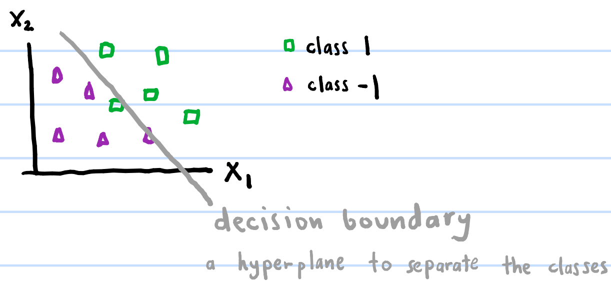
... for binary classification.

training data: $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$

- feature vector $\vec{x}_i \in \mathbb{R}^d$ ← # features
- labels $y_i \in \{-1, 1\}$ ← for math convenience

SVM in words: find the hyperplane that "best" separates the classes in feature space

e.g. $d=2$

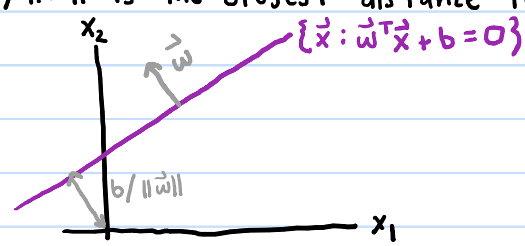


preliminaries on planes

a plane is described by the set of all points such that $\vec{w}^T\vec{x} + b = 0$.

$$\{\vec{x} : f(\vec{x}) = \vec{w}^T\vec{x} + b = 0\}$$

- $\vec{w}$ is a vector normal to the plane (*)
- $b/\|\vec{w}\|$ is the closest distance to the origin (**)



to see (*), consider two different points $\vec{x}_1, \vec{x}_2$ on the plane.

$$\Rightarrow \begin{array}{l} \vec{w}^T\vec{x}_1 + b = 0 \\ \vec{w}^T\vec{x}_2 + b = 0 \end{array} \Rightarrow \vec{w}^T(\vec{x}_1 - \vec{x}_2) = 0 \quad \text{i.e. } \vec{w} \text{ is orthogonal to } \vec{x}_1 - \vec{x}_2$$

the vector $\vec{x}_1 - \vec{x}_2$ lies in the plane.

since this is true $\forall\, \vec{x}_1, \vec{x}_2$ on the plane, $\vec{w}$ is orthogonal to the plane ∎

to see (**), consider the optimization problem of finding the closest point on the plane to the origin:

$$\min_{\vec{x}} \|\vec{x} - \vec{0}\|^2 \text{ subject to } \vec{w}^T\vec{x} + b = 0$$

$$\Updownarrow$$

$$\min_{\vec{x}} \|\vec{x}\|^2 + \lambda(\vec{w}^T\vec{x} + b)$$

↖ Lagrange multiplier

$$\vec{\nabla}_{\vec{x}}\left[\|\vec{x}\|^2 + \lambda(\vec{w}^T\vec{x} + b)\right] = \vec{0}$$

$$2\vec{x} + \lambda\vec{w} = \vec{0}$$

$$\vec{x} = \frac{\lambda}{2}\vec{w}$$

enforce constraint: $\vec{w}^T\left(\frac{\lambda}{2}\vec{w}\right) + b = 0$

$$\Rightarrow \frac{\lambda}{2} = b/\|\vec{w}\|^2$$

$$\Rightarrow \vec{x} = \frac{b}{\|\vec{w}\|^2}\vec{w} \Rightarrow \|\vec{x}\| = \frac{b}{\|\vec{w}\|}$$ ∎

- the closest (signed) distance of any point $\vec{y}$ to the plane $\{\vec{x}: f(\vec{x}) = \vec{w}^T\vec{x}+b=0\}$ is:

$$\frac{1}{\|\vec{w}\|}\left(\vec{w}^T\vec{y}+b\right)$$

to see this, consider the optimization problem:

a pt. on the plane

$$\min_{\vec{x}} \|\vec{y}-\vec{x}\|^2 \quad \text{subject to} \quad \vec{w}^T\vec{x}+b=0$$

Lagrange formulation:

$$\min_{\vec{x}} \|\vec{y}-\vec{x}\|^2 + \lambda\left(\vec{w}^T\vec{x}+b\right)$$

$$\vec{\nabla}_{\vec{x}}\left(\qquad \cdots \qquad\right) = \vec{0}$$

$$2(\vec{y}-\vec{x}) + \lambda\vec{w} = \vec{0}$$

$$\Rightarrow \vec{x} = \frac{\lambda}{2}\vec{w} + \vec{y}$$

enforce constraint:

$$\vec{w}^T\left(\frac{\lambda}{2}\vec{w}+\vec{y}\right)+b=0$$

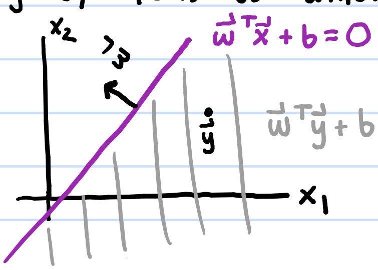
$$\frac{\lambda}{2}\|\vec{w}\|^2 + \vec{w}^T\vec{y} + b = 0$$

$$\frac{\lambda}{2} = \left(\vec{w}^T\vec{y}+b\right)/\|\vec{w}\|^2$$

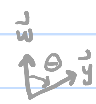
$$\Rightarrow \vec{x}-\vec{y} = \frac{\lambda}{2}\vec{w} = \left(\vec{w}^T\vec{y}+b\right)\frac{\vec{w}}{\|\vec{w}\|^2}$$

$$\|\vec{x}-\vec{y}\| = \|\vec{w}^T\vec{y}+b\| / \|\vec{w}\| \qquad \blacksquare$$

- $\text{sign}\left(\vec{w}^T\vec{y}+b\right)$ tells us which side of the plane $\vec{y}$ is on

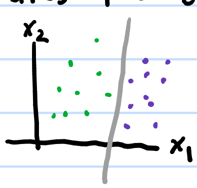


think: $\vec{w}^T\vec{y} = \|\vec{w}\|\|\vec{y}\|\cos\theta$

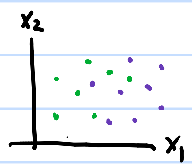


CASE (1): separable classes

the classes are separable if there exists a hyperplane in feature space that perfectly separates the classes.

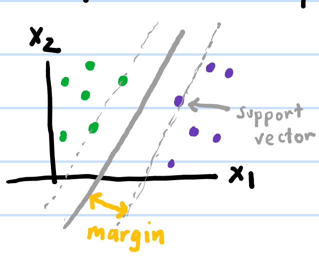


separable

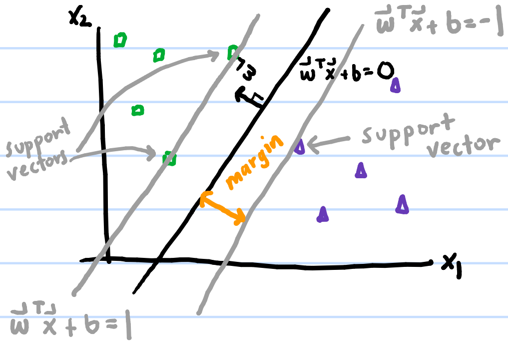


not separable

an SVM finds | the separating hyperplane that gives the largest margin:



note: there are infinite separating planes.
SVM gives the plane with the largest margin



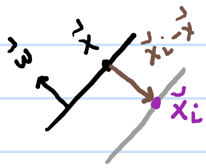
margin = $\frac{1}{\|\vec{w}\|}$

to see this, let $\vec{x}$ be the closest point on $\{\vec{x} : \vec{w}^T\vec{x}+b=0\}$ to a support vector $\vec{x}_i$.

$\Rightarrow \vec{w}^T\vec{x}_i + b = 1$ (or -1)

$\Rightarrow \vec{w}^T(\vec{x}_i-\vec{x}) = 1 = \|\vec{w}\|\,\|\vec{x}_i-\vec{x}\|$ b/c $\vec{w}$, $\vec{x}_i-\vec{x}$ point in same direction.

$\Rightarrow \|\vec{x}_i-\vec{x}\| = \text{margin} = 1/\|\vec{w}\|$



classification rule for SVM:

$$y = \text{sign}(\vec{w}^T\vec{x}+b)$$

i.e. which side of the plane is $\vec{x}$ on?

if all y_i classified correctly:

$$y_i\,\text{sign}(\vec{w}^T\vec{x}_i+b) = 1 \quad \forall i$$

this is where it's nice that $y_i \in \{-1,1\}$.

i.e. y_i and $\vec{w}^T\vec{x}_i+b$ have the same sign.

the (convex) optimization problem for SVM for separable classes:

$$\max_{\vec{w},b} \frac{1}{\|\vec{w}\|}$$

choose $\vec{w}, b$ to maximize the margin

subject to $y_i(\vec{w}^T\vec{x}_i+1) \geq 1$

contraint: no data mis-classified

&, even stronger, no data inside margins

the optimal $\vec{w}, b$ end up being determined completely by the data points closest to the separating hyperplane, the support vectors.

CASE (2): non-separable classes

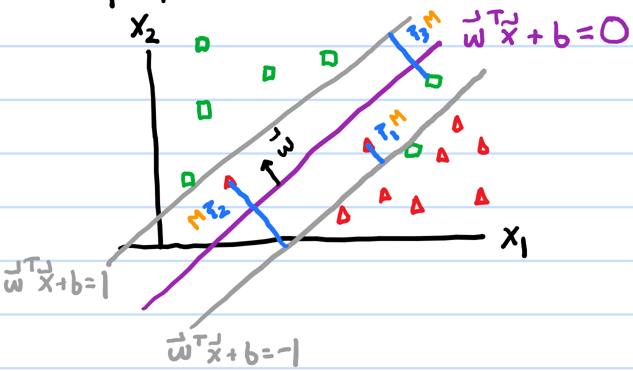
a solution to the above optimization problem does not exist.

we must accept that some data will be mis-classified, yet choose the plane to minimize mis-classifications.

define "slack variables" $\xi_i \geq 0$ associated with each data point.

the meaning of these variables?

ξ_i is proportional to how far data pt. i is from the margin (on the wrong side)



precisely, now we enforce:

$$y_i(\vec{w}^T\vec{x}_i + b) \geq \underbrace{1 - \xi_i}_{\text{the } \xi_i \geq 0 \text{ allow slack.}}$$

$\xi_i = 0 \Rightarrow \vec{x}_i$ on correct side of plane, outside margin

$0 < \xi_i \leq 1 \Rightarrow \vec{x}_i$ on correct side of dividing plane, but inside margin

$\xi_i > 1 \Rightarrow \vec{x}_i$ on wrong side of dividing plane, mis-classified

we impose an upper bound on $\sum \xi_i$.

this essentially limits the number of mis-classifications we allow in our attempt to maximize the margin.

e.g. enforce $\sum \xi_i < K \Rightarrow$ allow fewer than K misclassifications.

the Lagrangian form of the (convex) optimization problem:

$$\min_{\vec{w}, b, \xi_i} \frac{1}{2}\|\vec{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

maximize the margin but penalize size of slack variables

subject to $y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \xi_i$ $\quad i = 1, \ldots, n$
$\xi_i \geq 0$

C is a hyperparameter of the SVM.

$C \to \infty \Rightarrow$ disallow pts inside margins / misclassification