Imperfect separation

Soft-margin SVM

Definition: <u>Hinge loss</u> : cost of a violation.

$$\xi^{(i)} = \max\Big(0, \underbrace{1 - (\vec{x}^{(i)} \cdot \vec{w} + b)\, y^{(i)}}_{\text{size of violation}}\Big)$$

Soft-margin SVM optimisation:

$$\boxed{\begin{aligned} &\min_{\vec{w}} \frac{1}{2}\|\vec{w}\|^2 + \lambda \sum_{i=1}^{N} \xi^{(i)} \\ &\text{subject to} \quad 1 - y^{(i)}(\vec{w}\cdot\vec{x}^{(i)} + b) \leq \xi^{(i)} \\ &\qquad\qquad \xi^{(i)} \geq 0 \qquad i = 1, \dots, N \end{aligned}}$$

Remember that $y^{(i)}(\vec{x}^{(i)}\vec{w} + b) \geq 1$ for all points when there is no violation

Beyond linearity:

Kernelised SVM

based on the 'kernel trick'

Standard SVM (linear)

linear classifier $(\vec{w}^*, b)$

$$\begin{cases} \vec{w}^* = \sum_{i=1}^{N} \alpha_i \, y^{(i)} \vec{x}^{(i)} \\ \alpha_i = 0 \quad \forall i \text{ not a support vector} \end{cases}$$

Assume there are only two support vectors (generic case):

Then only $0 \neq \alpha_+$ and $\alpha_- \neq 0$: $\vec{w}^* = \alpha_+ \vec{x}_+ - \alpha_- \vec{x}_-$

$$b = 1 - \vec{x}_+ \cdot \left[\alpha_+ \vec{x}_+ - \alpha_- \vec{x}_- \right]$$

$$= 1 - \alpha_+ \boxed{\vec{x}_+ \cdot \vec{x}_+} + \alpha_- \boxed{\vec{x}_+ \cdot \vec{x}_-}$$

These define our model

Obtained from maximising of the dual: $L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j \, y^{(i)} y^{(j)} \left(\boxed{\vec{x}^{(i)} \cdot \vec{x}^{(j)}} \right)$

Given $\vec{x}^{in}$, $\begin{cases} \boxed{\vec{x}^{in} \cdot \vec{w}^*} + b \geq 0 \Rightarrow \hat{y} = +1 \\ \boxed{\vec{x}^{in} \cdot \vec{w}^*} + b \leq 0 \Rightarrow \hat{y} = -1 \end{cases}$

Kernel functions

Given $\vec{x}, \vec{y} \in \mathbb{R}^d$, consider $\vec{\phi}(\vec{x}) = \vec{z} \in \mathbb{R}^D$

potentially in $D \gg d$

In some cases, for the right ϕ we have nice properties:

e.g. $\vec{x} = (x_1, x_2) \in \mathbb{R}^2$

$$\vec{\phi}(\vec{x}) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \end{pmatrix} \in \mathbb{R}^3$$

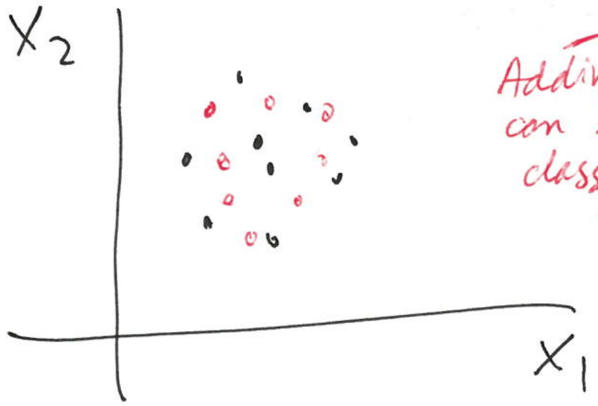
$$\vec{\phi}(\vec{x}) \cdot \vec{\phi}(\vec{y}) = x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 x_2 y_1 y_2$$

$$\vec{x} \cdot \vec{y} = x_1 y_1 + x_2 y_2$$

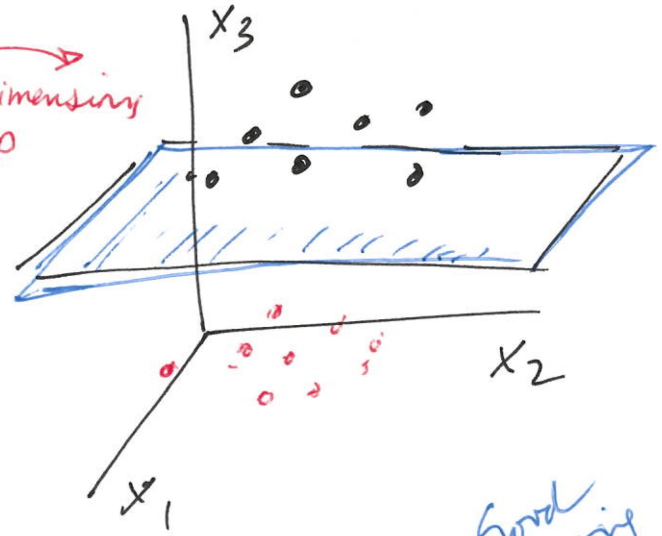
$$g(\underbrace{\vec{x} \cdot \vec{y}}_{2D \cdot 2D}) = (\vec{x} \cdot \vec{y})^2 = x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 y_1 x_2 y_2$$

$$= \underbrace{\vec{\phi}(\vec{x})}_{3D} \cdot \underbrace{\vec{\phi}(\vec{y})}_{3D}$$

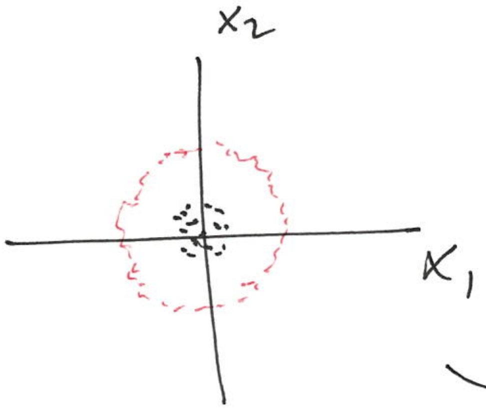
Idea: Nonlinearity and high-dimension can help with classification.



Adding dimensions can help classify



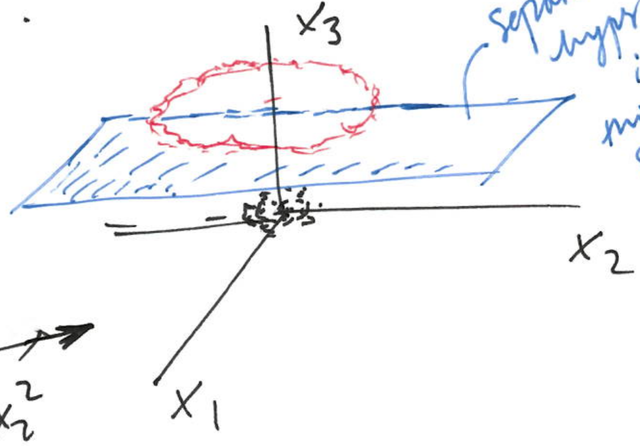
More interesting:



No separating hyperplane

$X_3 = X_1^2 + X_2^2$

Add nonlinear coordinate



Good separating hyperplane in this space

There exists a good separating hyperplane in this space:

$(x_1, x_2, x_1^2 + x_2^2)$

In general, a kernel function maps a pair of vectors in the lower dimensional space ($\mathbb{R}^d$) to give the dot product of transformed (nonlinear) vectors in the high-dimensional space ($\mathbb{R}^D$)

$$K(\vec{x},\vec{y}) = \vec{\phi}(\vec{x})\cdot\vec{\phi}(\vec{y}) \quad \text{where } \vec{x},\vec{y} \in \mathbb{R}^d \text{ and } \vec{\phi}(\vec{x}), \vec{\phi}(\vec{y}) \in \mathbb{R}^D$$

$\boxed{K \text{ is a kernel with associated } \vec{\phi} \text{ iff } K \text{ is positive semidefinite}}$

Several important kernels:

(1) $K(\vec{x},\vec{y}) = (\vec{x}\cdot\vec{y}+1)^n$ Polynomial

(2) $K(\vec{x},\vec{y}) = e^{-\|\vec{x}-\vec{y}\|^2/\sigma}$ Gaussian Radial basis

(3) $K(\vec{x},\vec{y}) = \tanh(\beta(\vec{x}\cdot\vec{y})+c)$ Sigmoid.

Choose a kernel and compute the corresponding Kernelised SVM.

More formally:

Choose a kernel $K(\bar{x},\bar{y})$

Kernelised SVM optimization:

$\vec{z} = \phi(\vec{x})$ associated with $\underline{K(\bar{x},\bar{y})}$

$$\min_{\vec{w}_z} \frac{1}{2} \|\vec{w}_z\|^2$$

subject to $1 - y^{(i)}(\underbrace{\vec{w}_z \cdot \vec{z}^{(i)}} + b) \leq 0$

$i = 1, \dots, N$

Looking back at the SVM expressions:

$$\vec{w}_z \cdot \vec{z}^{(i)} = \phi(\bar{w}) \cdot \phi(\vec{x}^{(i)}) = K(\bar{w}, \vec{x}^{(i)})$$

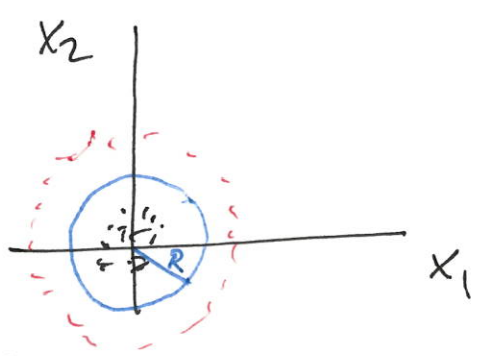
In the Lagrangian:

$$\vec{z}^{(i)} \cdot \vec{z}^{(j)} = K(\vec{x}^{(i)}, \vec{x}^{(j)})$$

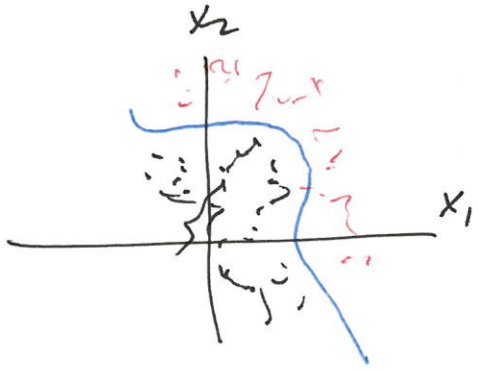
Decision:

Given $\vec{x}^{in}$

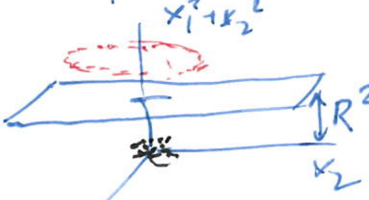
$$\begin{cases} K(\vec{x}^{in}, \bar{w}) + b > 0 \Rightarrow \hat{y} = +1 \\ K(\vec{x}^{in}, \bar{w}) + b < 0 \Rightarrow \hat{y} = -1 \end{cases}$$



K



This is equivalent to



The decision boundary is $\underline{\text{nonlinear}}$ in the $\vec{x}$ space