# # 6 Classification & Support Vector Machine

Arthur Charpentier (Université du Québec à Montréal)

Machine Learning & Econometrics

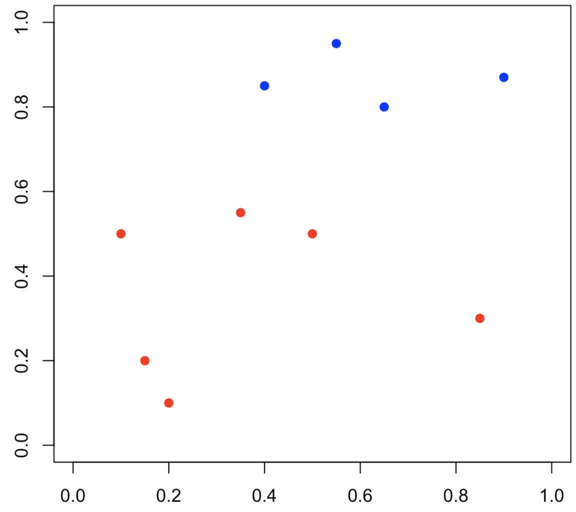
SIDE Summer School - July 2019

# SVM : Support Vector Machine

## Linearly Separable sample [econometric notations]

Data $(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)$ - with $y \in \{0, 1\}$ - are linearly separable if there are $(\beta_0, \boldsymbol{\beta})$ such that

- $y_i = 1$ if $\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} > 0$
- $y_i = 0$ if $\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} < 0$

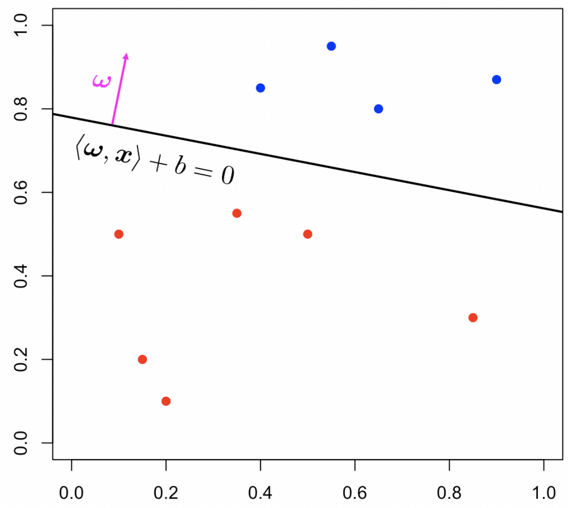



## Linearly Separable sample [ML notations]

Data $(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)$ - with $y \in \{-1, +1\}$ - are linearly separable if there are $(b, \boldsymbol{\omega})$ such that

- $y_i = +1$ if $b + \langle \boldsymbol{x}, \boldsymbol{\omega} \rangle > 0$
- $y_i = -1$ if $b + \langle \boldsymbol{x}, \boldsymbol{\omega} \rangle < 0$

or equivalently $y_i \cdot (b + \langle \boldsymbol{x}, \boldsymbol{\omega} \rangle) > 0$, $\forall i$.



## SVM : Support Vector Machine

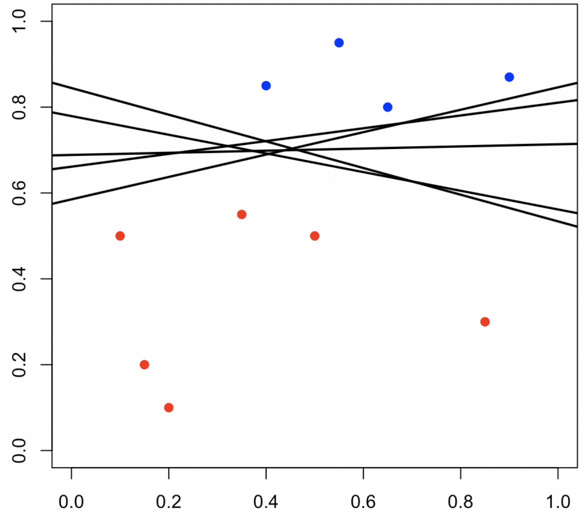
$y_i \cdot (b + \langle \boldsymbol{x}, \boldsymbol{\omega} \rangle) = 0$ is an hyperplane (in $\mathbb{R}^p$)
orthogonal with $\boldsymbol{\omega}$

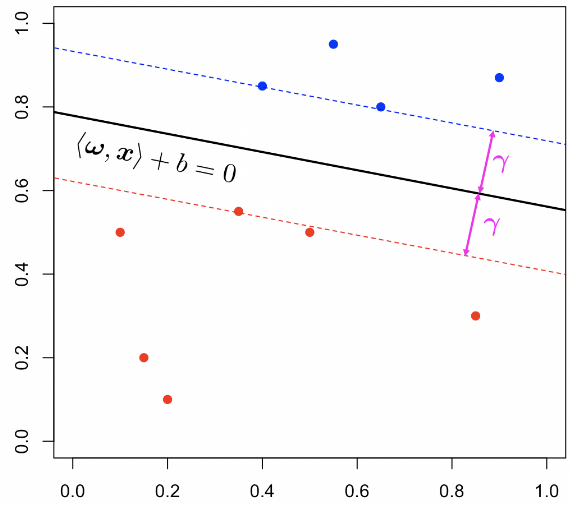
Use $m(\boldsymbol{x}) = \mathbf{1}_{b+\langle \boldsymbol{x}, \boldsymbol{\omega} \rangle \geq 0} - \mathbf{1}_{b+\langle \boldsymbol{x}, \boldsymbol{\omega} \rangle < 0}$ as classifier

Problem : equation (i.e. $(b, \boldsymbol{\omega})$) is not unique !

Canonical form : $\min_{i=1,\cdots,n} \left\{ |b + \langle \boldsymbol{x}_i, \boldsymbol{\omega} \rangle| \right\} = 1$

Problem : solution here is not unique !




Idea : use the widest (safety) margin $\gamma$

Vapnik & Lerner (1963, Pattern recognition using generalized portrait method) or Cover (1965, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition)

## SVM : Support Vector Machine

Consider two points, $\boldsymbol{\omega}_{-1}$ and $\boldsymbol{\omega}_{+1}$

$$\gamma = \frac{1}{2}\frac{\langle\omega,\boldsymbol{\omega}_{+1}-\boldsymbol{\omega}_{-1}\rangle}{\|\omega\|}$$

It is minimal when

$b+\langle\boldsymbol{x}_i,\boldsymbol{\omega}_{-1}\rangle=-1$ and

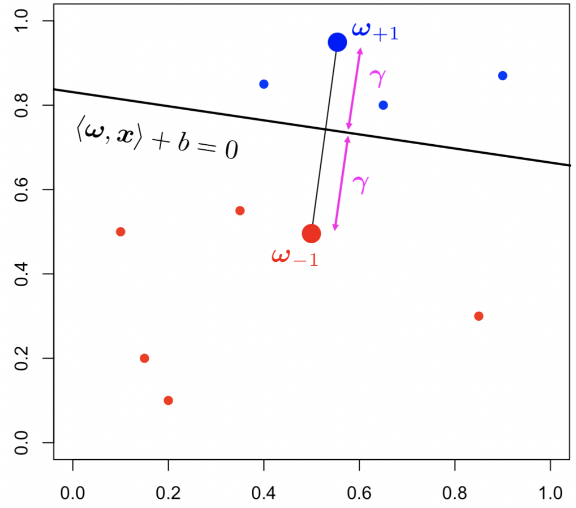
$b+\langle\boldsymbol{x}_i,\boldsymbol{\omega}_{+1}\rangle=+1$, and therefore

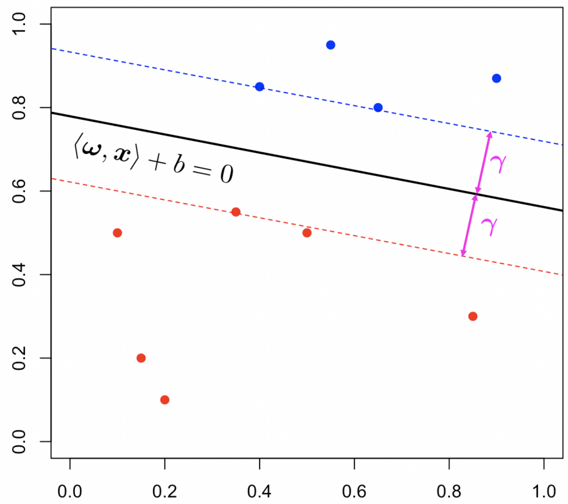
$$\gamma^\star=\frac{1}{\|\boldsymbol{\omega}\|}$$

Optimization problem becomes

$$\min_{(b\,\boldsymbol{\omega})}\left\{\frac{1}{2}\|\boldsymbol{\omega}\|_{\ell_2}^2\right\}\text{ s.t. } y_i\cdot(b+\langle\boldsymbol{x},\boldsymbol{\omega}\rangle)>0,\ \forall i.$$

convex optimization problem with linear constraints



## SVM : Support Vector Machine

Consider the following problem : $\min_{\boldsymbol{u}\in\mathbb{R}^p} h(\boldsymbol{u})$ s.t. $g_i(\boldsymbol{u}) \geq 0\ \forall i = 1, \cdots, n$

where $h$ is quadratic and $g_i$'s are linear.

Lagrangian is $L : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}$ defined as $L(\boldsymbol{u}, \boldsymbol{\alpha}) = h(\boldsymbol{u}) - \displaystyle\sum_{i=1}^n \alpha_i g_i(\boldsymbol{u})$

where $\boldsymbol{\alpha}$ are dual variables, and the dual function is

$$\Lambda : \boldsymbol{\alpha} \mapsto L(\boldsymbol{u}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) = \min\{L(\boldsymbol{u}, \boldsymbol{\alpha})\} \text{ where } \boldsymbol{u}_{\boldsymbol{\alpha}} = \text{argmin}\{L(\boldsymbol{u}, \boldsymbol{\alpha})\}$$

One can solve the dual problem, $\max\{\Lambda(\boldsymbol{\alpha})\}$ s.t. $\boldsymbol{\alpha} \geq \boldsymbol{0}$. Solution is $\boldsymbol{u} = \boldsymbol{u}_{\boldsymbol{\alpha}^\star}$.

Si $g_i(\boldsymbol{u}_{\boldsymbol{\alpha}^\star}) > 0$, then necessarily $\alpha_i^\star = 0$ (see Karush-Kuhn-Tucker (KKT) condition, $\alpha_i^\star \cdot g_i(\boldsymbol{u}_{\alpha^\star}) = 0$)

## SVM : Support Vector Machine

Here, $L(b,\boldsymbol{\omega},\boldsymbol{\alpha})=\dfrac{1}{2}\|\boldsymbol{\omega}\|^2-\displaystyle\sum_{i=1}^n\alpha_i\cdot\big(y_i\cdot(b+\langle\boldsymbol{x},\boldsymbol{\omega}\rangle)-1\big)$

From the first order conditions,

$$\frac{\partial L(b,\boldsymbol{\omega},\boldsymbol{\alpha})}{\partial\boldsymbol{\omega}}=\boldsymbol{\omega}-\sum_{i=1}^n\alpha_i\cdot y_i\boldsymbol{x}_i=\boldsymbol{0},\ \text{ i.e. }\ \boldsymbol{\omega}^\star=\sum_{i=1}^n\alpha_i^\star y_i\boldsymbol{x}_i$$

$$\frac{\partial L(b,\boldsymbol{\omega},\boldsymbol{\alpha})}{\partial b}=-\sum_{i=1}^n\alpha_i\cdot y_i=0,\ \text{ i.e. }\ \sum_{i=1}^n\alpha_i^\star\cdot y_i=0$$

and

$$\Lambda(\boldsymbol{\alpha})=\sum_{i=1}^n\alpha_i-\frac{1}{2}\sum_{i,j=1}^n\alpha_i\alpha_jy_iy_j\langle\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j\rangle$$

## SVM : Support Vector Machine

$$\min_{\boldsymbol{\alpha}}\left\{\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{\alpha}-\boldsymbol{1}^\top\boldsymbol{\alpha}\right\}\text{ s.t. }\begin{cases}\alpha_i\geq 0,\ \forall i\\ \boldsymbol{y}^\top\boldsymbol{1}=0\end{cases}$$

where $\boldsymbol{Q}=[\boldsymbol{Q}_{i,j}]$ and $\boldsymbol{Q}_{i,j}=y_iy_j\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$, and then

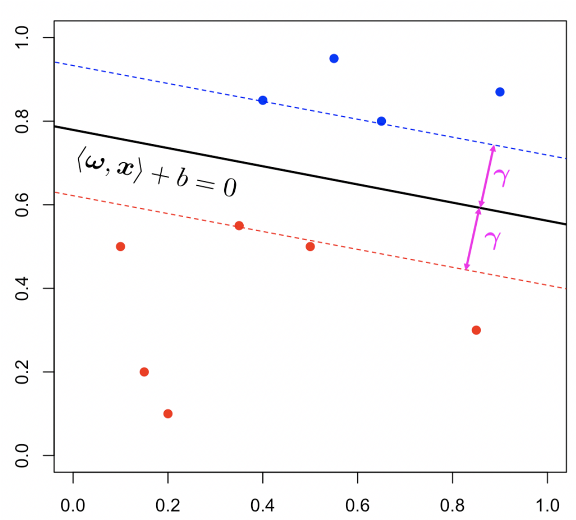
$$\boldsymbol{\omega}^\star=\sum_{i=1}^n\alpha_i^\star y_i\boldsymbol{x}_i\text{ and }b^\star=-\frac{1}{2}\left[\min_{i:y_i=+1}\{\langle\boldsymbol{x}_i,\boldsymbol{\omega}^\star\rangle\}+\min_{i:y_i=-1}\{\langle\boldsymbol{x}_i,\boldsymbol{\omega}^\star\rangle\}\right]$$

Points $\boldsymbol{x}_i$ such that $\alpha_i^\star>0$ are called support

$y_i\cdot(b^\star+\langle\boldsymbol{x}_i,\boldsymbol{\omega}^\star\rangle)=1$

Use $m^\star(\boldsymbol{x})=\boldsymbol{1}_{b^\star+\langle\boldsymbol{x},\boldsymbol{\omega}^\star\rangle\geq 0}-\boldsymbol{1}_{b^\star+\langle\boldsymbol{x},\boldsymbol{\omega}^\star\rangle<0}$ as classifier

Observe that $\gamma^\star=\left(\sum_{i=1}^n\alpha_i^{\star 2}\right)^{-1/2}$



## SVM : Support Vector Machine

Optimization problem was

$$\min_{(b,\boldsymbol{\omega})}\left\{\frac{1}{2}\|\boldsymbol{\omega}\|_{\ell_2}^2\right\}\text{ s.t. }y_i\cdot(b+\langle\boldsymbol{x},\boldsymbol{\omega}\rangle)>0,\ \forall i,$$

which became

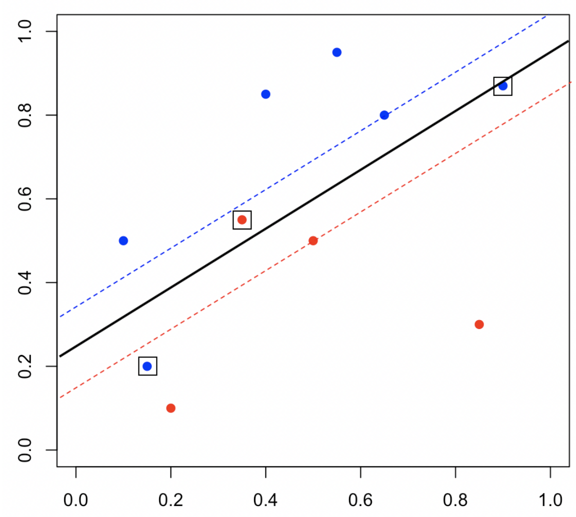
$$\min_{(b,\boldsymbol{\omega})}\left\{\frac{1}{2}\|\boldsymbol{\omega}\|_{\ell_2}^2+\sum_{i=1}^n\alpha_i\cdot\big(1-y_i\cdot(b+\langle\boldsymbol{x},\boldsymbol{\omega}\rangle)\big)\right\},$$

or

$$\min_{(b,\boldsymbol{\omega})}\left\{\frac{1}{2}\|\boldsymbol{\omega}\|_{\ell_2}^2+\text{penalty}\right\},$$

## SVM : Support Vector Machine

Consider here the more general case where the space is not linearly separable $(\langle\boldsymbol{\omega},\boldsymbol{x}_i\rangle+b)y_i\geq 1$ becomes $(\langle\boldsymbol{\omega},\boldsymbol{x}_i\rangle+b)y_i\geq 1-\xi_i$ for some slack variables $\xi_i$'s.




and penalize large slack variables $\xi_i$ (when $>0$) by solving (for some cost $C$)

$$\min_{\boldsymbol{\omega},b}\left\{\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}+C\sum_{i=1}^n\xi_i\right\}$$

subject to $\forall i$, $\xi_i\geq 0$ and $(\boldsymbol{x_i}^\top\boldsymbol{\omega}+b)y_i\geq 1-\xi_i$.

This is the soft-margin extension, see `e1071::svm()` or `kernlab::ksvm()`

## SVM : Support Vector Machine

The dual optimization problem is now

$$\min_{\boldsymbol{\alpha}}\left\{\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{\alpha}-\mathbf{1}^\top\boldsymbol{\alpha}\right\}\text{ s.t. }\begin{cases}0\leq\alpha_i\leq C, & \forall i\\ \boldsymbol{y}^\top\mathbf{1}=0\end{cases}$$

where $\boldsymbol{Q}=[\boldsymbol{Q}_{i,j}]$ and $\boldsymbol{Q}_{i,j}=y_iy_j\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$, and then

$$\boldsymbol{\omega}^\star=\sum_{i=1}^n\alpha_i^\star y_i\boldsymbol{x}_i\text{ and }b^\star=-\frac{1}{2}\left[\min_{i:y_i=+1}\{\langle\boldsymbol{x}_i,\boldsymbol{\omega}^\star\rangle\}+\min_{i:y_i=-1}\{\langle\boldsymbol{x}_i,\boldsymbol{\omega}^\star\rangle\}\right]$$

Note further that the (primal) optimization problem can be written

$$\min_{(b,\boldsymbol{\omega})}\left\{\frac{1}{2}\|\omega\|_{\ell_2}^2+\sum_{i=1}^n\big(1-y_i\cdot(b+\langle\boldsymbol{x},\boldsymbol{\omega}\rangle)\big)_+\right\},$$

where $(1-z)_+$ is a convex upper bound for empirical error $\mathbf{1}_{z\leq0}$

## SVM : Support Vector Machine

One can also consider the kernel trick : $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ is replace by $\varphi(\boldsymbol{x}_i)^\top \varphi(\boldsymbol{x}_j)$ for some mapping $\varphi$,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^\top \varphi(\boldsymbol{x}_j)$$

For instance $K(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a}^\top \boldsymbol{b})^3 = \varphi(\boldsymbol{a})^\top \varphi(\boldsymbol{b})$

where $\varphi(a_1, a_2) = (a_1^3\ ,\ \sqrt{3}a_1^2 a_2\ ,\ \sqrt{3}a_1 a_2^2\ ,\ a_2^3)$

Consider polynomial kernels

$$K(\boldsymbol{a}, \boldsymbol{b}) = (1 + \boldsymbol{a}^\top \boldsymbol{b})^p$$

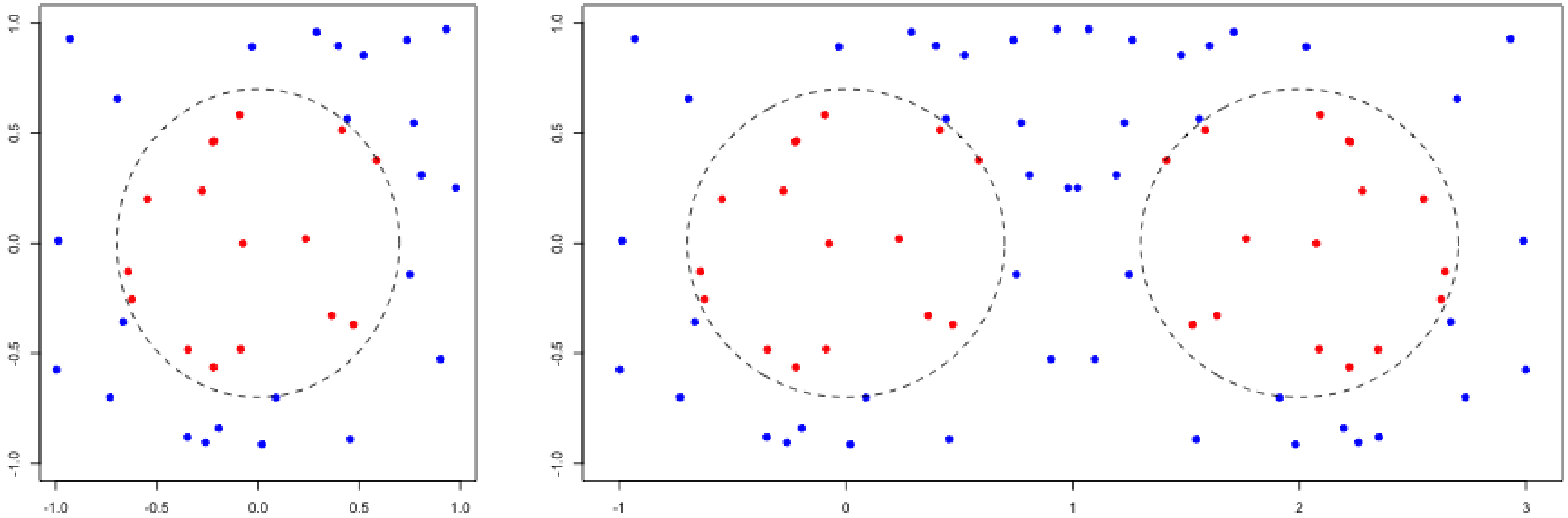
or a Gaussian kernel

$$K(\boldsymbol{a}, \boldsymbol{b}) = \exp(-(\boldsymbol{a} - \boldsymbol{b})^\top (\boldsymbol{a} - \boldsymbol{b}))$$

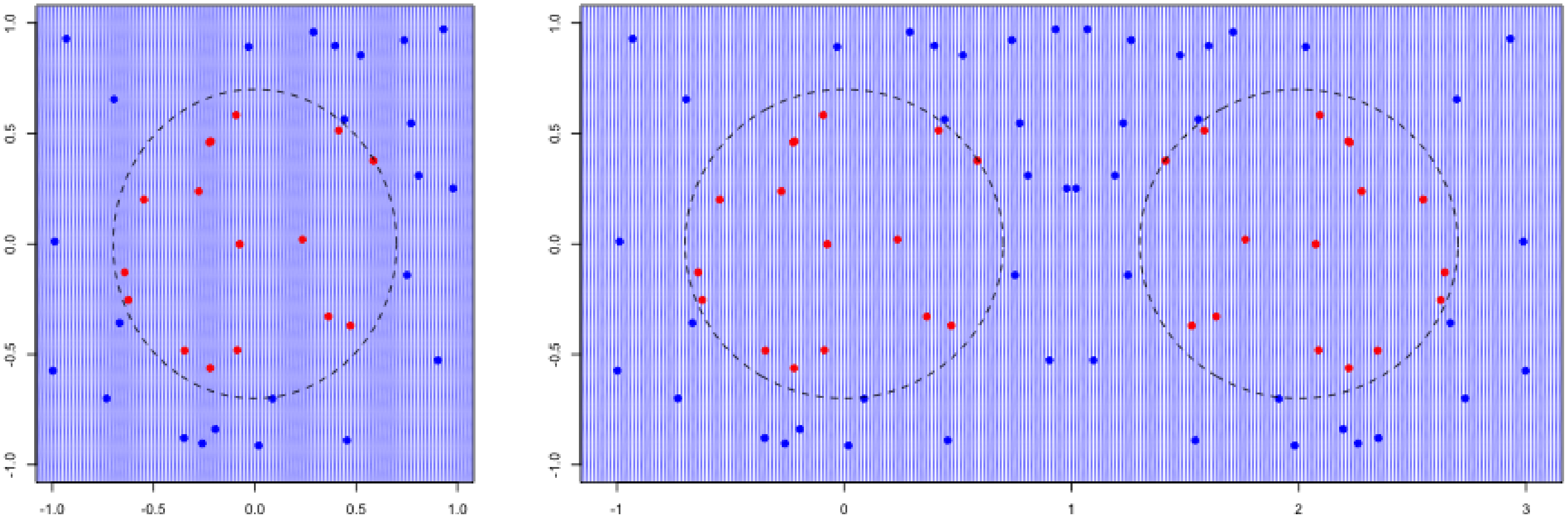
and solve $\displaystyle\max_{\alpha_i \geq 0} \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \right\}$

## SVM : Support Vector Machine

Consider the following training sample $\{(y_i, x_{1,i}, x_{2,i})\}$ with $y_i \in \{\bullet, \bullet\}$

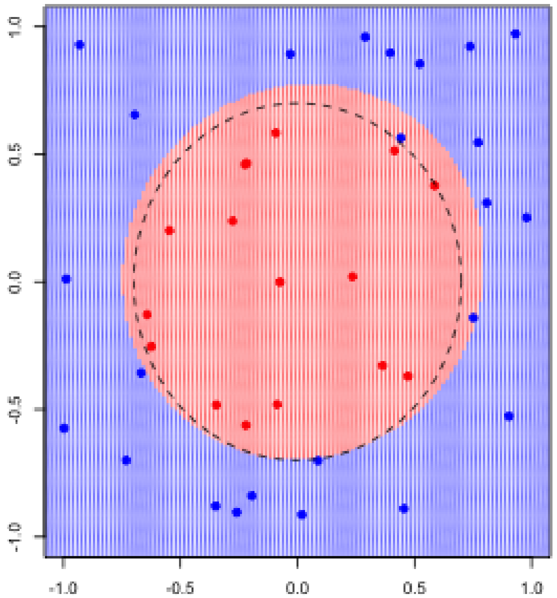


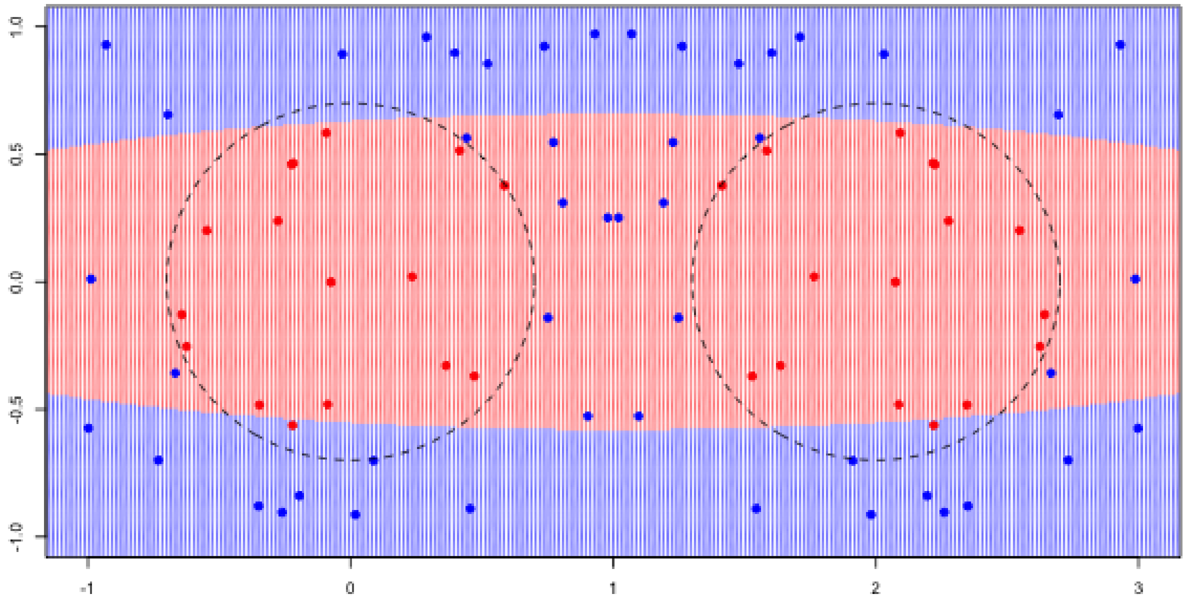
# SVM : Support Vector Machine




Linear kernel

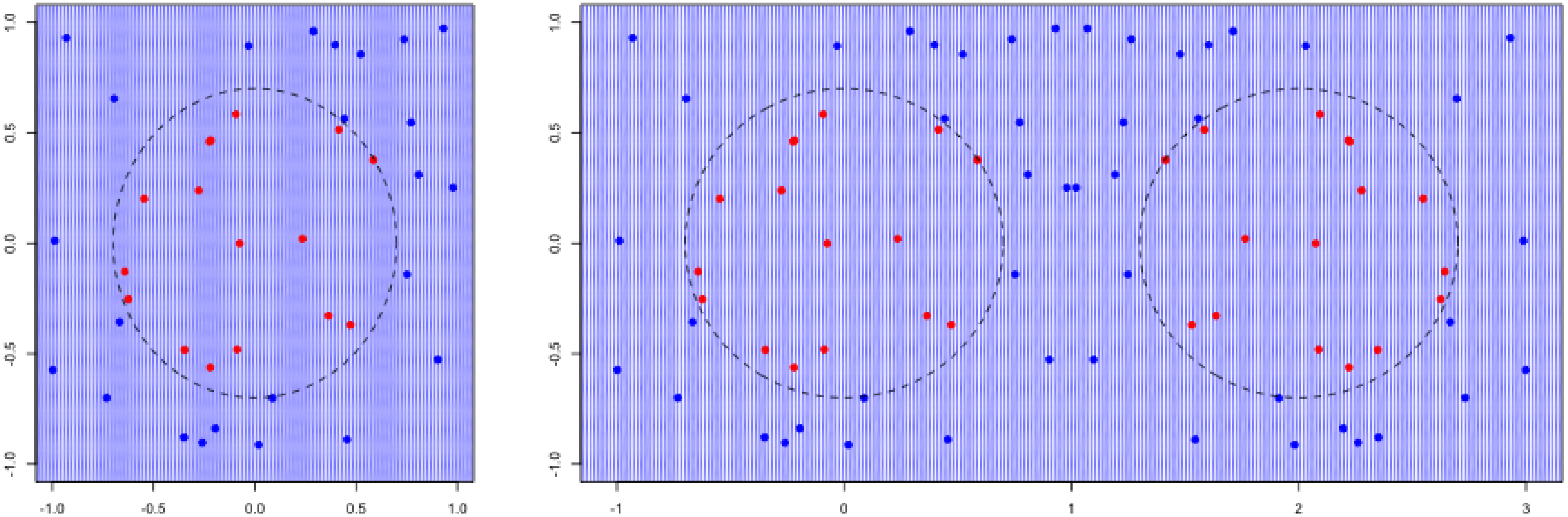
## SVM : Support Vector Machine



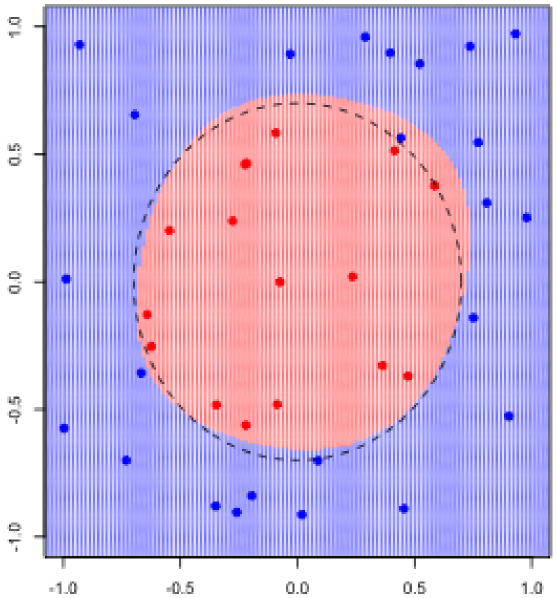

Polynomial kernel (degree 2)

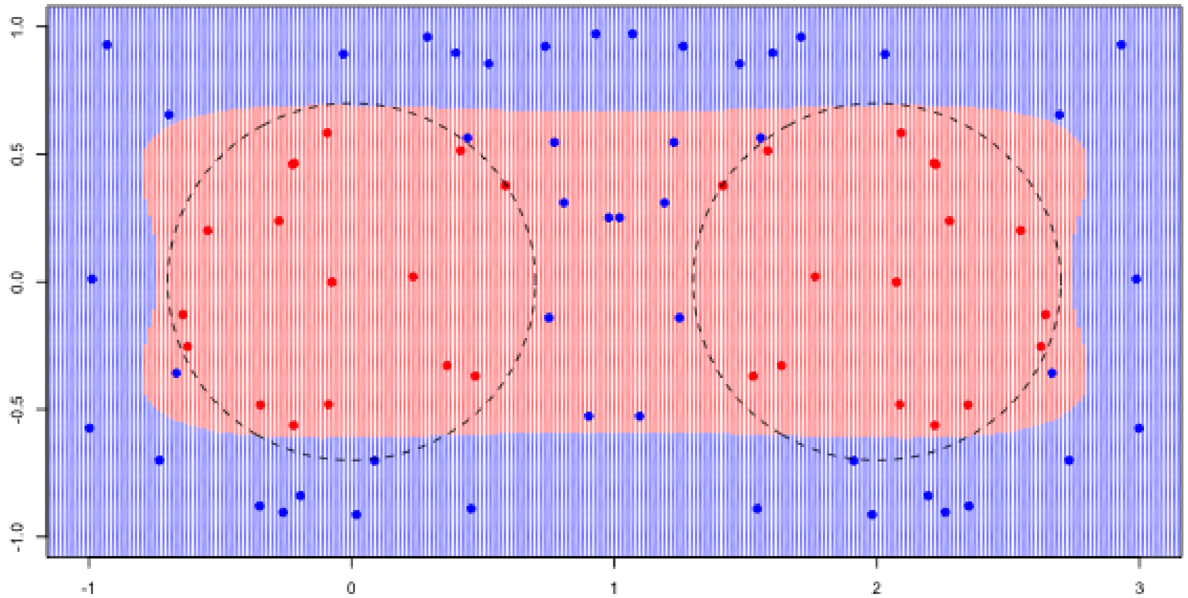
## SVM : Support Vector Machine




Polynomial kernel (degree 3)

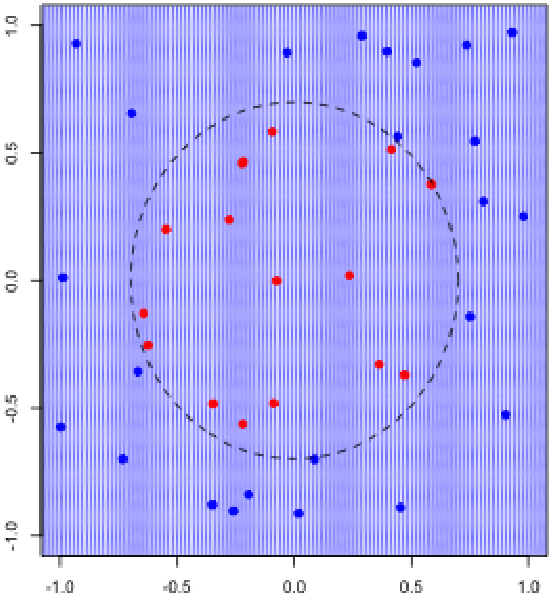
## SVM : Support Vector Machine

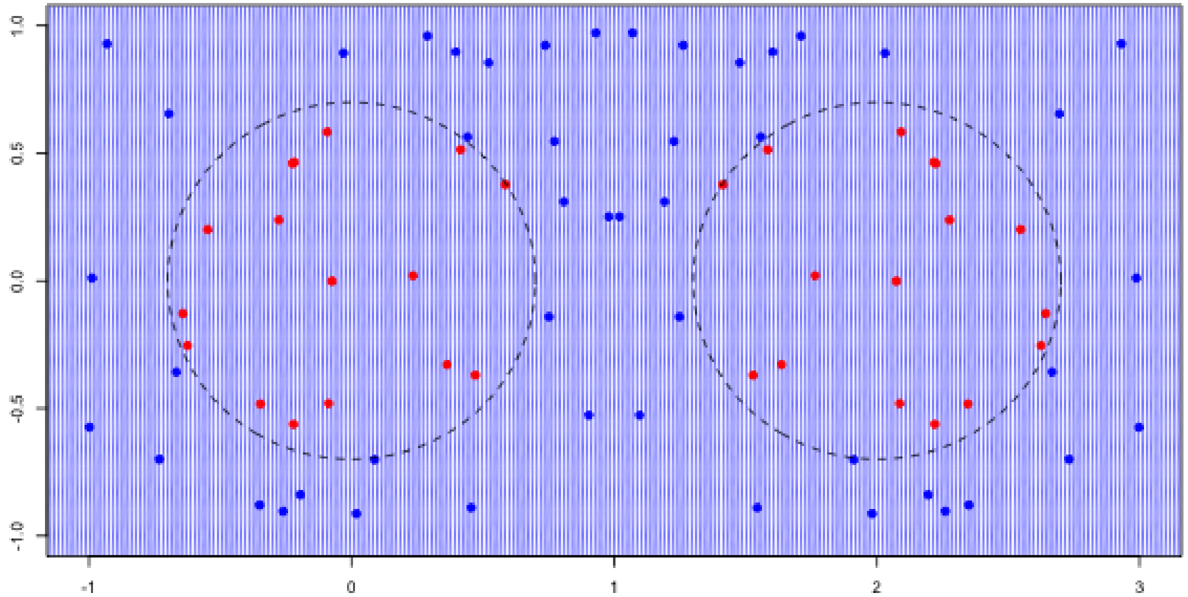



Polynomial kernel (degree 4)

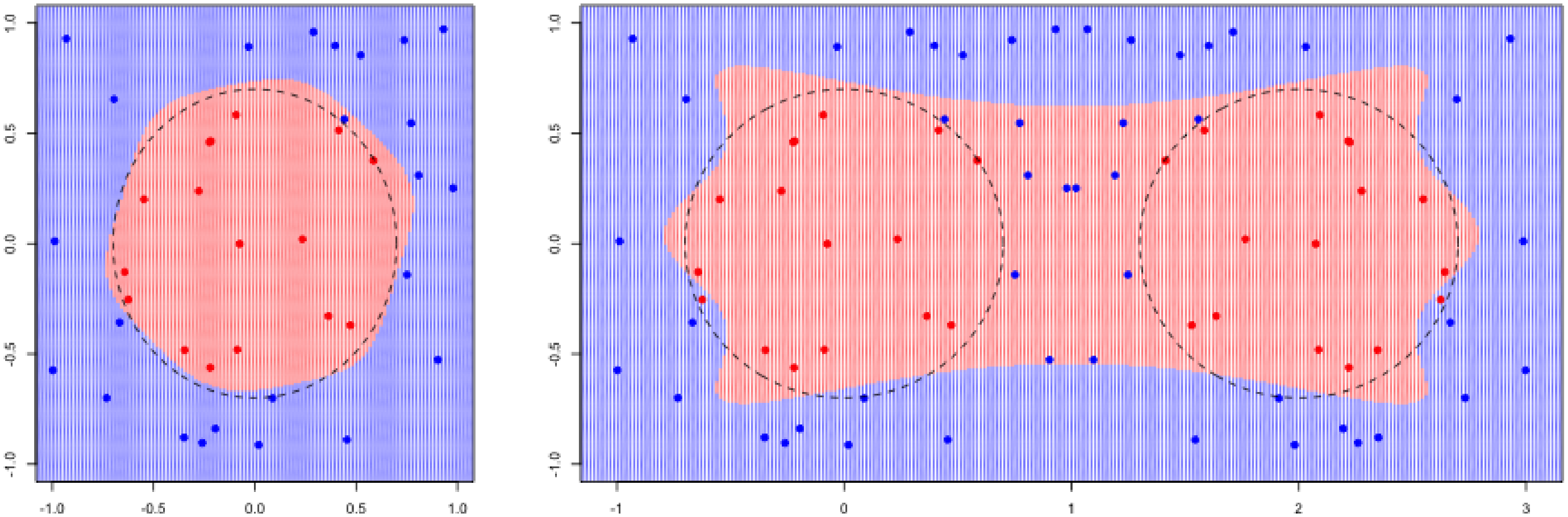
## SVM : Support Vector Machine



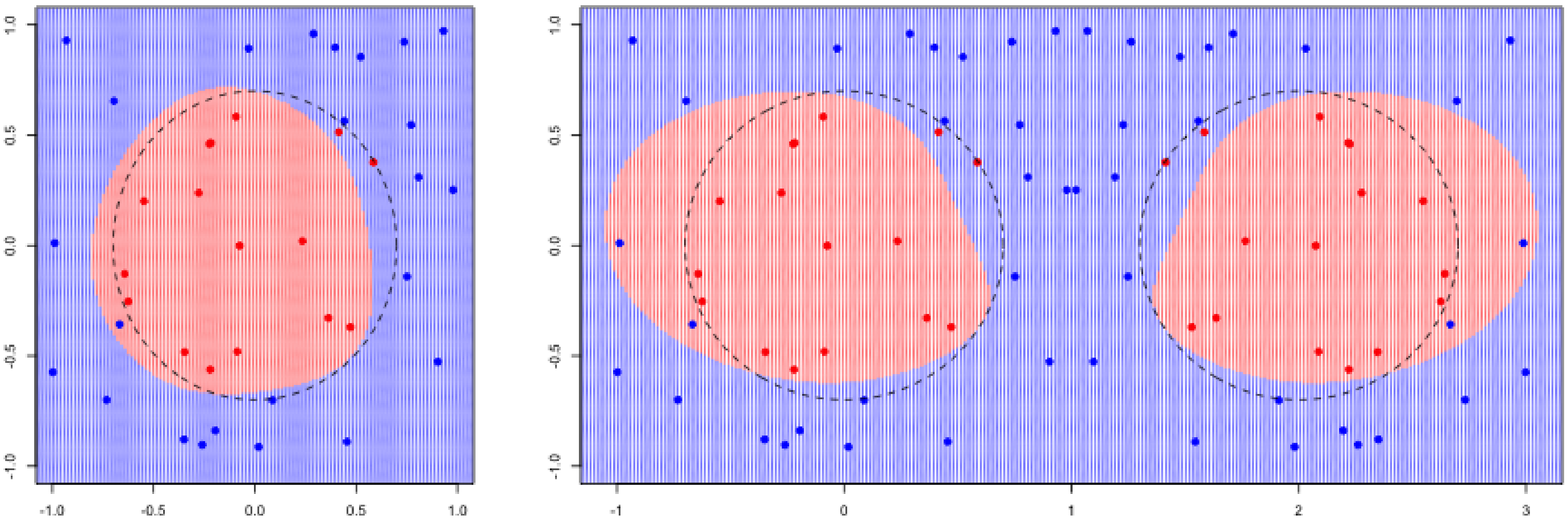

Polynomial kernel (degree 5)

## SVM : Support Vector Machine



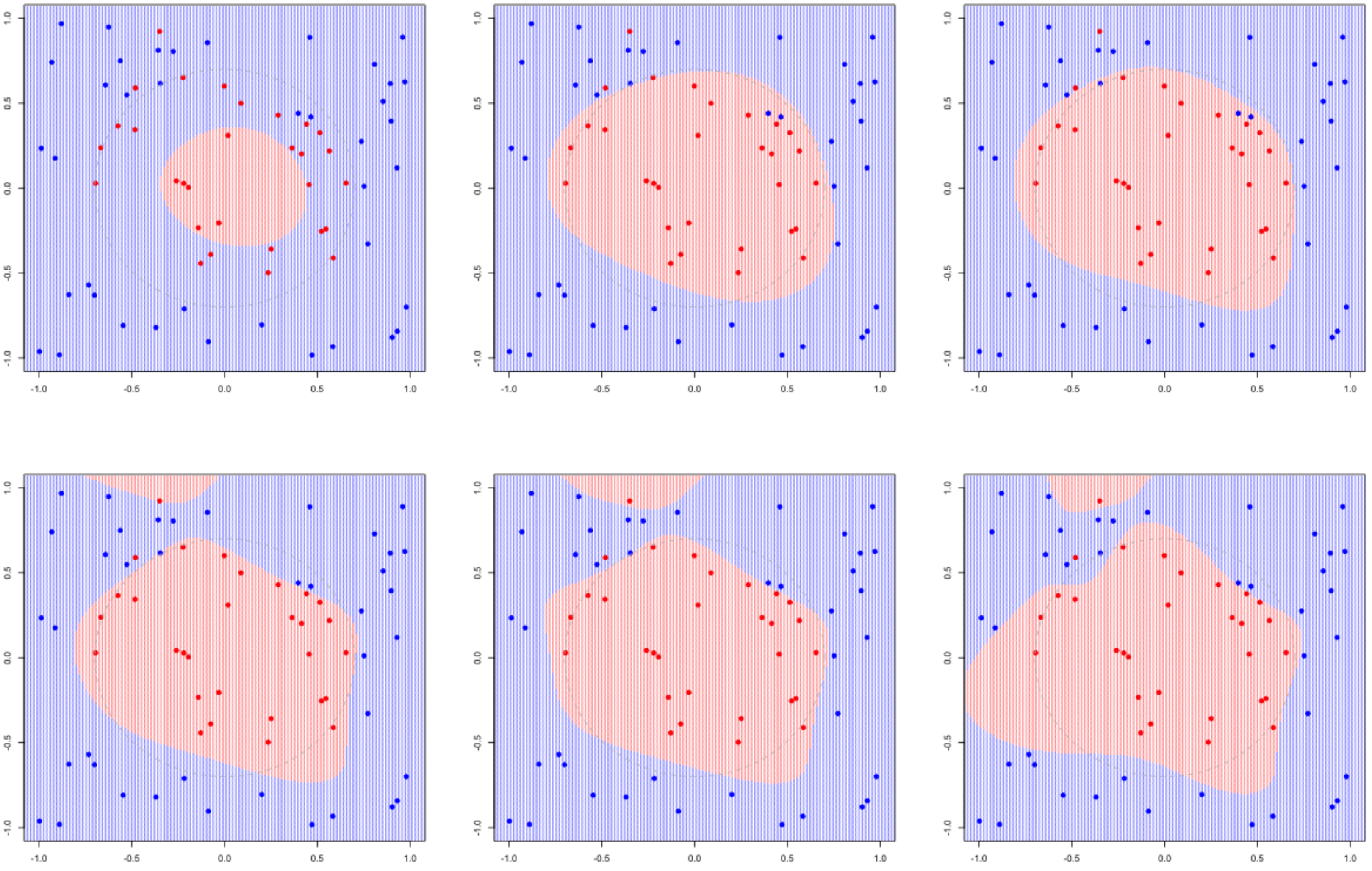

Polynomial kernel (degree 6)

## SVM : Support Vector Machine

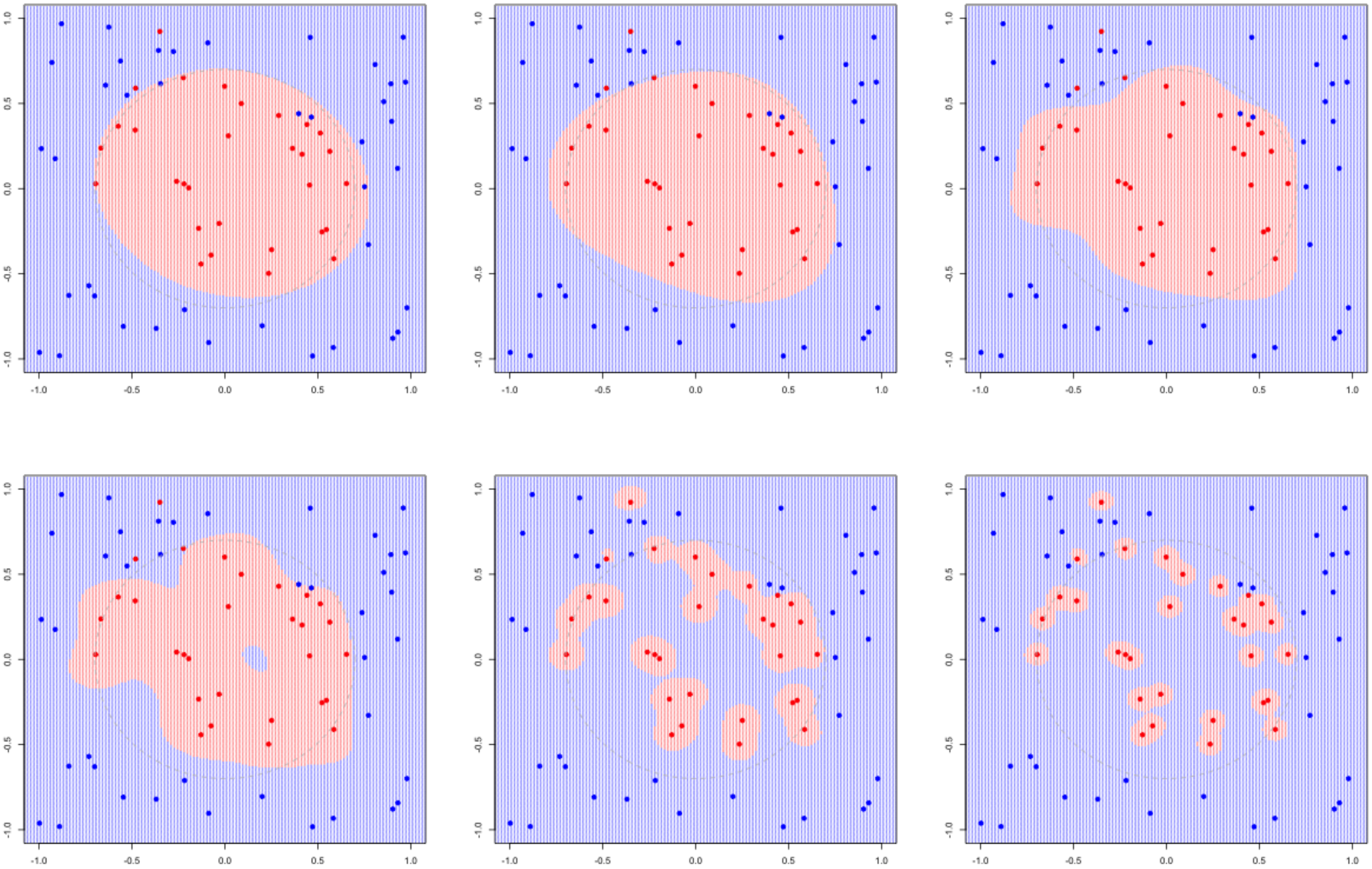



Radial kernel

## SVM : Support Vector Machine - Radial Kernel, impact of the cost $C$



# SVM : Support Vector Machine - Radial Kernel, tuning parameter $\gamma$



## SVM : Support Vector Machine

The radial kernel is formed by taking an infinite sum over polynomial kernels...

$$K(\boldsymbol{x},\boldsymbol{y}) = \exp\big(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2\big) = \langle\psi(\boldsymbol{x}),\psi(\boldsymbol{y})\rangle$$

where $\psi$ is some $\mathbb{R}^n \to \mathbb{R}^\infty$ function, since

$$K(\boldsymbol{x},\boldsymbol{y}) = \exp\big(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2\big) = \underbrace{\exp(-\gamma\|\boldsymbol{x}\|^2-\gamma\|\boldsymbol{y}\|^2)}_{=\text{constant}}\cdot\exp\big(2\gamma\langle\boldsymbol{x},\boldsymbol{y}\rangle\big)$$

i.e.

$$K(\boldsymbol{x},\boldsymbol{y}) \propto \exp\big(2\gamma\langle\boldsymbol{x},\boldsymbol{y}\rangle\big) = \sum_{k=0}^{\infty} 2\gamma\frac{\langle\boldsymbol{x},\boldsymbol{y}\rangle^k}{k!} = \sum_{k=0}^{\infty} 2\gamma K_k(\boldsymbol{x},\boldsymbol{y})$$

where $K_k$ is the polynomial kernel of degree $k$.

If $K = K_1 + K_2$ with $\psi_j : \mathbb{R}^n \to \mathbb{R}^{d_j}$ then $\psi : \mathbb{R}^n \to \mathbb{R}^d$ with $d \sim d_1 + d_2$

## SVM : Support Vector Machine

A kernel is a measure of similarity between vectors.

The smaller the value of $\gamma$ the narrower the vectors should be to have a small measure

Is there a probabilistic interpretation ?

Platt (2000, Probabilities for SVM) suggested to use a logistic function over the SVM scores,

$$p(\boldsymbol{x}) = \frac{\exp[b + \langle \boldsymbol{x}, \boldsymbol{\omega} \rangle]}{1 + \exp[b + \langle \boldsymbol{x}, \boldsymbol{\omega} \rangle]}$$