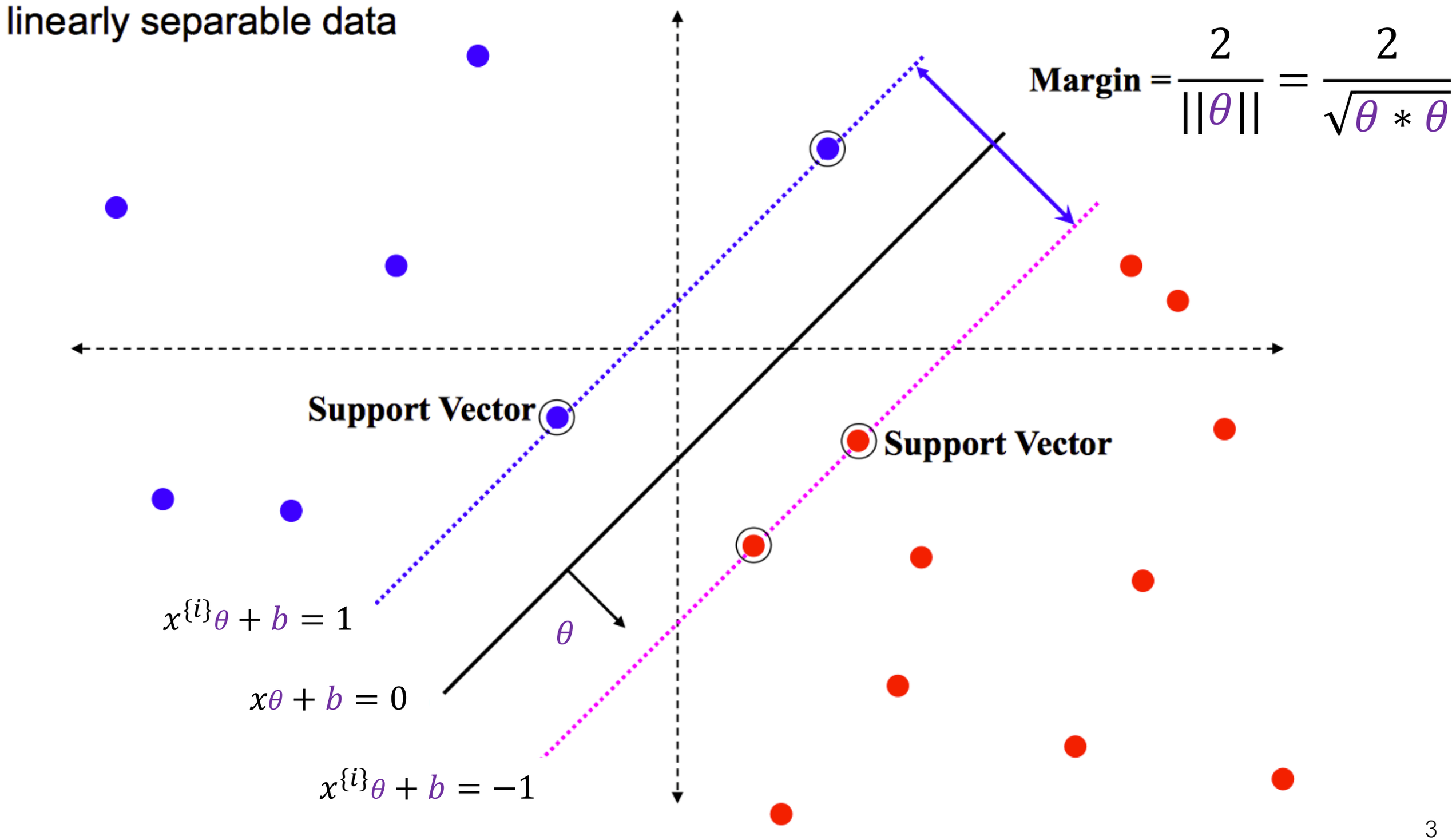
Georgia Tech

Kernel SVM

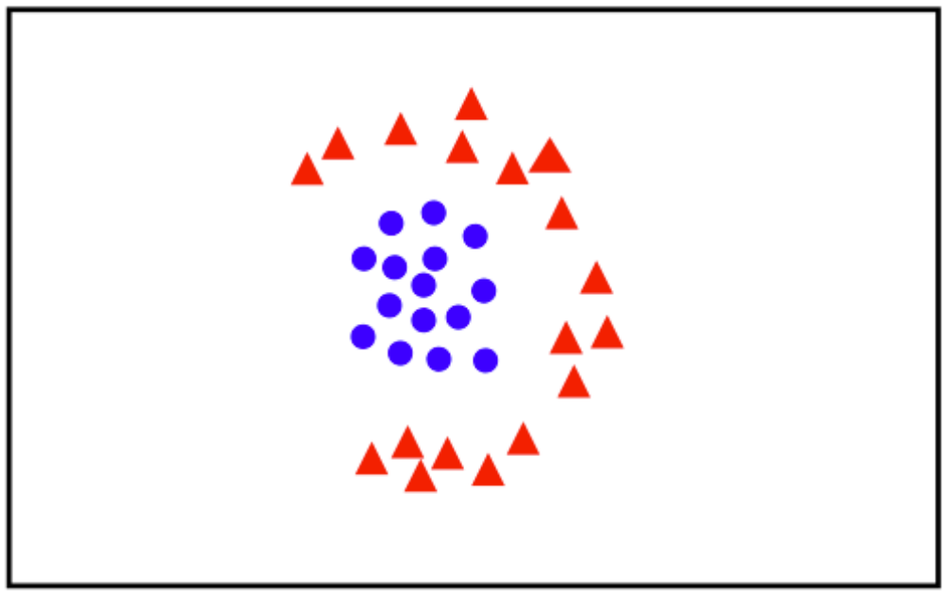
Mahdi Roozbahani
Georgia Tech

These slides are inspired based on slides from Yaser Mostafa, Le Song and Eric Eaton.

Geometric Interpretation



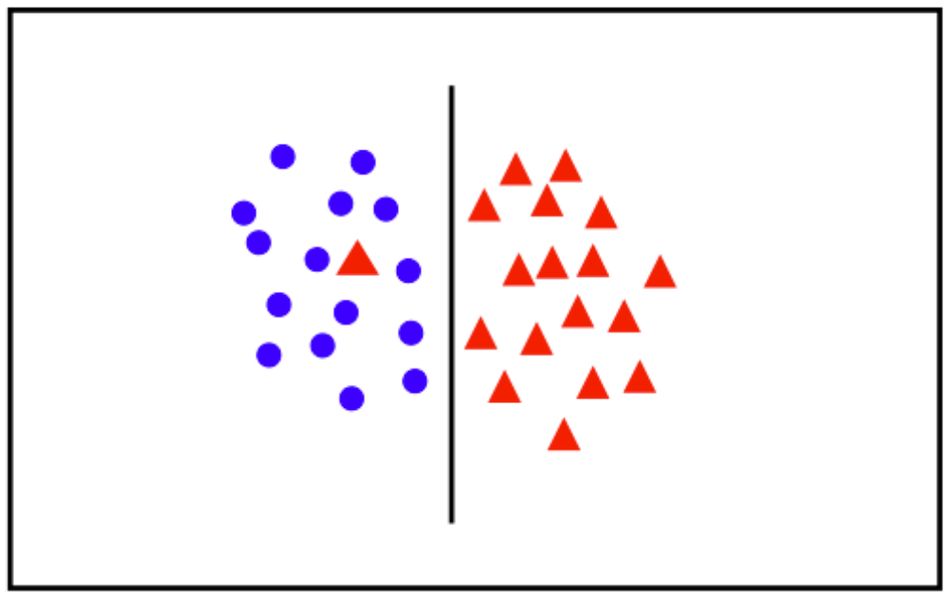
Handling Data that are Not Linearly Separable



- linear classifier not appropriate

??

Kernel trick

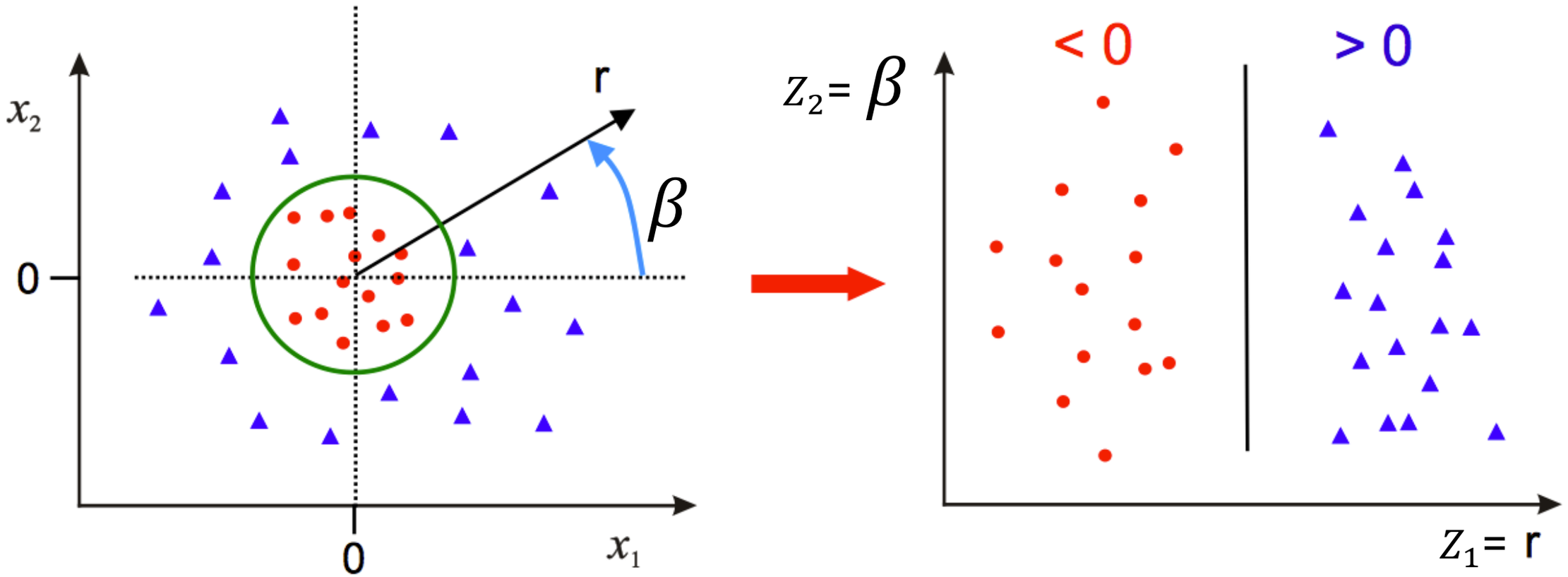


- introduce slack variables

Soft Margin SVM

(allowing ourselves to make errors)

Idea 1: Use Polar Coordinates to go to z space

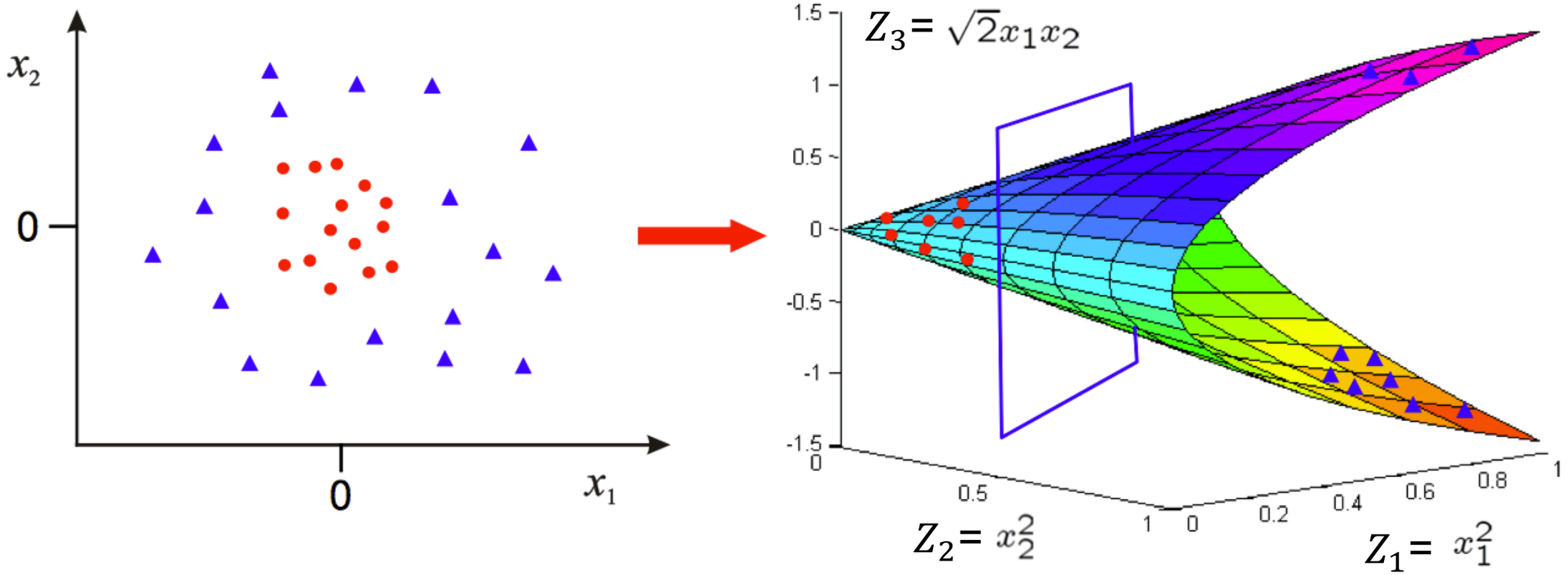


- Data is linearly separable in polar coordinates
- Acts non-linearly in original space

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} r \\ \beta \end{pmatrix} \qquad \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

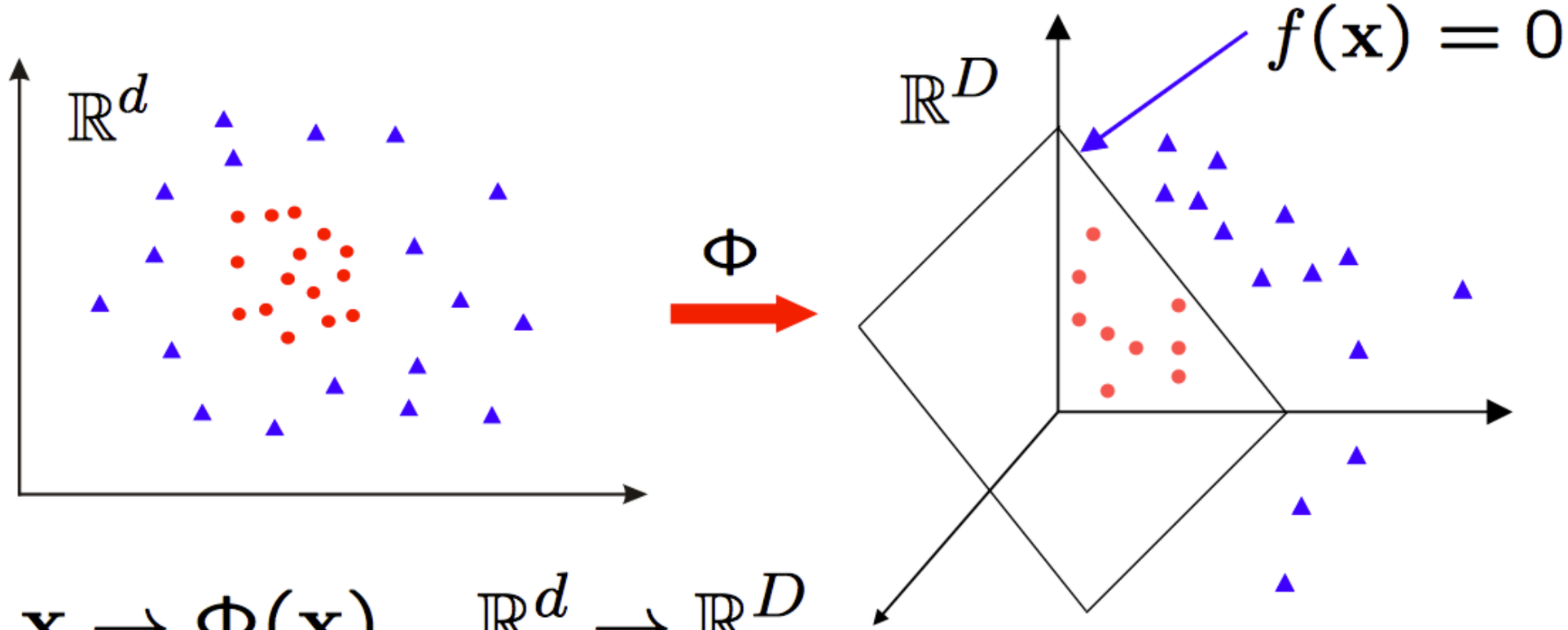
Idea 1: Map Data to Higher Dimension $\mathbb{Z}$ space

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \qquad \mathbb{R}^2 \rightarrow \mathbb{R}^3$$



- Data is linearly separable in 3D
- This means that the problem can still be solved by a linear classifier

SVM in a Transformed Feature Space



$\Phi : \mathbf{x} \to \Phi(\mathbf{x})$ $\quad \mathbb{R}^d \to \mathbb{R}^D$

Learn classifier linear in $\mathbf{w}$ for $\mathbb{R}^D$:

$$f(x) = \phi(x)\theta + \theta_0 = z\theta + \theta_0$$

$\Phi(\mathbf{x})$ is a feature map

Kernel trick – what do we need from $\mathbb{Z}$ space

$$l(\alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y^{\{i\}}y^{\{j\}}\alpha^{\{i\}}\alpha^{\{j\}}\underbrace{z^{\{i\}}z^{\{j\}^T}}_{\text{Inner products}}$$

Constraints: $\alpha^{\{i\}} \geq 0$ for $i = 1, \dots, N$ and $\sum_{i=1}^{N} \alpha^{\{i\}}y^{\{i\}} = 0$

We already have this: $\begin{bmatrix} y^{\{1\}}y^{\{1\}}z^{\{1\}}z^{\{1\}^T} & \cdots & y^{\{1\}}y^{\{N\}}z^{\{1\}}z^{\{N\}^T} \\ \vdots & \ddots & \vdots \\ y^{\{N\}}y^{\{1\}}z^{\{N\}}z^{\{1\}^T} & \cdots & y^{\{N\}}y^{\{N\}}z^{\{N\}}z^{\{N\}^T} \end{bmatrix}$

Same result as hard SVM:

Solve $\alpha^{\{i\}}$ using quadratic programming and predict a test data point z in z space➔

$$sign(\mathrm{z}\theta + b) = \sum_{z_i in\ SV} \alpha^{\{i\}}y^{\{i\}}z^{\{i\}}\mathrm{z} + b$$

Generalized inner product

Given **two points** $x^{\{1\}}$ and $x^{\{2\}}$

$$\begin{bmatrix} y^{\{1\}}y^{\{1\}}K(x^{\{1\}},x^{\{1\}}) & y^{\{1\}}y^{\{2\}}K(x^{\{1\}},x^{\{2\}}) \\ y^{\{2\}}y^{\{1\}}K(x^{\{2\}},x^{\{1\}}) & y^{\{2\}}y^{\{2\}}K(x^{\{2\}},x^{\{2\}}) \end{bmatrix}$$

How can we calculate $z^{\{1\}}{z^{\{2\}}}^T$

Let $K\left(x^{\{1\}},x^{\{2\}}\right) = z^{\{1\}}{z^{\{2\}}}^T$ **The kernel** inner product of $x^{\{1\}}$and $x^{\{2\}}$

Example:

$x^{\{1\}} = \left(1, h^{\{1\}}, w^{\{1\}}\right) \rightarrow 2nd - order\ \Phi$ here $x^{\{1\}}\ and\ x^{\{2\}}$ have two dimensions

$x^{\{2\}} = \left(1, h^{\{2\}}, w^{\{2\}}\right) \rightarrow 2nd - order\ \Phi$

$z^{\{1\}} = \Phi(\boldsymbol{x}) = (1, h^{\{1\}}, w^{\{1\}}, {h^{\{1\}}}^2, {w^{\{1\}}}^2, h^{\{1\}}w^{\{1\}})$

$z^{\{2\}} = \Phi(\boldsymbol{x}') = (1, h^{\{2\}}, w^{\{2\}}, {h^{\{2\}}}^2, {w^{\{2\}}}^2, h^{\{2\}}w^{\{2\}})$

How many dimensions z has?

$K\left(x^{\{1\}},x^{\{2\}}\right) = z^{\{1\}}{z^{\{2\}}}^T = 1 + h^{\{1\}}h^{\{2\}} + w^{\{1\}}w^{\{2\}} + {h^{\{1\}}}^2{h^{\{2\}}}^2 + {w^{\{1\}}}^2{w^{\{2\}}}^2 + h^{\{1\}}h^{\{2\}}w^{\{1\}}w^{\{2\}}$

We can also calculate: $K\left(x^{\{1\}},x^{\{1\}}\right)K\left(x^{\{2\}},x^{\{1\}}\right)$ $K\left(x^{\{2\}},x^{\{2\}}\right)$

The trick

Can we compute $K(x^{\{1\}}, x^{\{2\}})$ **without** transforming $x^{\{1\}}$and $x^{\{2\}}$?

Example:

Datapoint 1 in x space $x^{\{1\}} = (1, h^{\{1\}}, w^{\{1\}})$

Datapoint 2 in x space $x^{\{2\}} = (1, h^{\{2\}}, w^{\{2\}})$

Datapoint 1 and 2 have 2 dimensions in x space and 1 is for the biased term

Datapoint 1 in z space $\phi(x^{\{1\}}) = z^{\{1\}} = (1, h^{\{1\}^2}, w^{\{1\}^2}, \sqrt{2}h^{\{1\}}, \sqrt{2}w^{\{1\}}, \sqrt{2}h^{\{1\}}w^{\{1\}})$

Datapoint 2 in z space $\phi(x^{\{2\}}) = z^{\{2\}} = (1, h^{\{2\}^2}, w^{\{2\}^2}, \sqrt{2}h^{\{2\}}, \sqrt{2}w^{\{2\}}, \sqrt{2}h^{\{2\}}w^{\{2\}})$

Datapoint 1 and 2 have 5 dimensions in z space and 1 is for the biased term

We need to calculate the dot product for the kernel

$$K\left(x^{\{1\}}, x^{\{2\}}\right) = \phi\left(x^{\{1\}}\right)\,\phi\left(x^{\{2\}}\right)^T = z^{\{1\}}{z^{\{2\}}}^T$$

$$z^{\{1\}} = (1, {h^{\{1\}}}^2, {w^{\{1\}}}^2, \sqrt{2}h^{\{1\}}, \sqrt{2}w^{\{1\}}, \sqrt{2}h^{\{1\}}w^{\{1\}})$$

$$z^{\{2\}} = (1, {h^{\{2\}}}^2, {w^{\{2\}}}^2, \sqrt{2}h^{\{2\}}, \sqrt{2}w^{\{2\}}, \sqrt{2}h^{\{2\}}w^{\{2\}})$$

$$z^{\{1\}}{z^{\{2\}}}^T = K\left(x^{\{1\}}, x^{\{2\}}\right) =$$

$$1 + {h^{\{1\}}}^2{h^{\{2\}}}^2 + {w^{\{1\}}}^2{w^{\{2\}}}^2 + 2h^{\{1\}}h^{\{2\}} + 2w^{\{1\}}w^{\{2\}} + 2h^{\{1\}}h^{\{2\}}w^{\{1\}}w^{\{2\}}$$

$$K\left(x^{\{1\}}, x^{\{2\}}\right) = \left(1 + h^{\{1\}}h^{\{2\}} + w^{\{1\}}w^{\{2\}}\right)^2$$

$$x^{\{1\}} = (1, h^{\{1\}}, w^{\{1\}}) \qquad x^{\{2\}} = (1, h^{\{2\}}, w^{\{2\}})$$

$$z^{\{1\}}{z^{\{2\}}}^T = K\left(x^{\{1\}}, x^{\{2\}}\right) = \left(x^{\{1\}}{x^{\{2\}}}^T\right)^2$$ Homogeneous kernel

Example: Let's say we have two datapoints

We need the build the inner product matrix to optimize SVM parameters.

$$l(\alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y^{\{i\}}y^{\{j\}}\alpha^{\{i\}}\alpha^{\{j\}} z^{\{i\}} z^{\{j\}^T} = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y^{\{i\}}y^{\{j\}}\alpha^{\{i\}}\alpha^{\{j\}} K(x^{\{i\}}, x^{\{j\}})$$

$$X = \begin{matrix} & \text{weight} & \text{Height} \\ \begin{matrix}1\\1\end{matrix} & \begin{matrix}1\\-1\end{matrix} & \begin{matrix}2\\3\end{matrix} \end{matrix} \quad Y = \begin{bmatrix} cat \\ dog \end{bmatrix}$$

$$K\left(x^{\{1\}}, x^{\{2\}}\right) = \begin{bmatrix} y^{\{1\}}y^{\{1\}}K(x^{\{1\}}, x^{\{1\}}) & y^{\{1\}}y^{\{2\}}K(x^{\{1\}}, x^{\{2\}}) \\ y^{\{2\}}y^{\{1\}}K(x^{\{2\}}, x^{\{1\}}) & y^{\{2\}}y^{\{2\}}K(x^{\{2\}}, x^{\{2\}}) \end{bmatrix}$$

The polynomial kernel

$\chi = \mathcal{R}^d \; and \; \Phi\colon \; \chi \; \rightarrow \mathbb{Z}$ is polynomial of order Q

The "equivalent kernel $= K\left(x^{\{1\}}, x^{\{2\}}\right) = \left(1 + x^{\{1\}} x^{\{2\}^T}\right)^Q$ Inhomogeneous kernel

$$= \left(1 + x_1^{\{1\}} x_1^{\{2\}} + x_2^{\{1\}} x_2^{\{2\}} + \cdots + x_d^{\{1\}} x_d^{\{2\}}\right)^Q$$

Does it matter if Q is 2 or 1000?

What will happen if we have $d = 10$ and $Q = 100$ and we want to compute the inner product explicitly?

We need to calculate the inner product of two big huge ugly vectors

We only need $\mathbb{Z}$ space to exist

If $K\left(x^{\{1\}}, x^{\{2\}}\right)$ is an inner product in some space $\mathbb{Z}$, we are doing good

Example:

$$K\left(x^{\{1\}}, x^{\{2\}}\right) = \exp\left(-\gamma \left\| x^{\{1\}} - x^{\{2\}} \right\|^2\right)$$ Radial basis kernel

First thing first, this is a function of $x \; and \; x'$

This function will take us to infinite-dimensional $\mathbb{Z}$ → CONGRATULATIONS

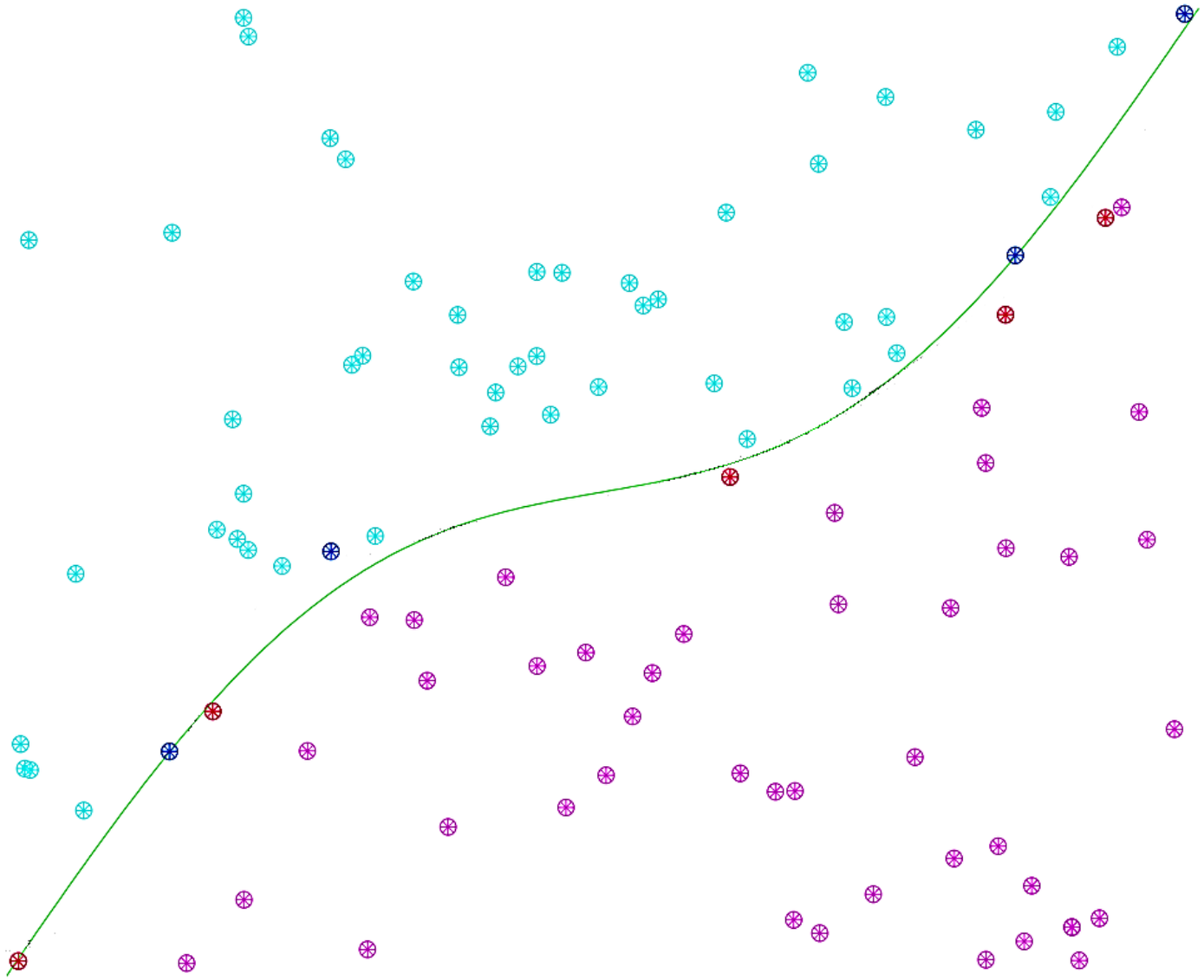
For d and $\gamma = 1 \Rightarrow \quad K\left(x^{\{1\}}, x^{\{2\}}\right) = \exp\left(-\left(x^{\{1\}} - x^{\{2\}}\right)^2\right)$

$$= \exp\left(-{x^{\{1\}}}^2\right) \exp\left(-{x^{\{2\}}}^2\right) \sum_{k=0}^{\infty} \frac{2^k {x^{\{1\}}}^k {x^{\{2\}}}^k}{k!}$$

exp Taylor expansion for $\exp(2xx')$

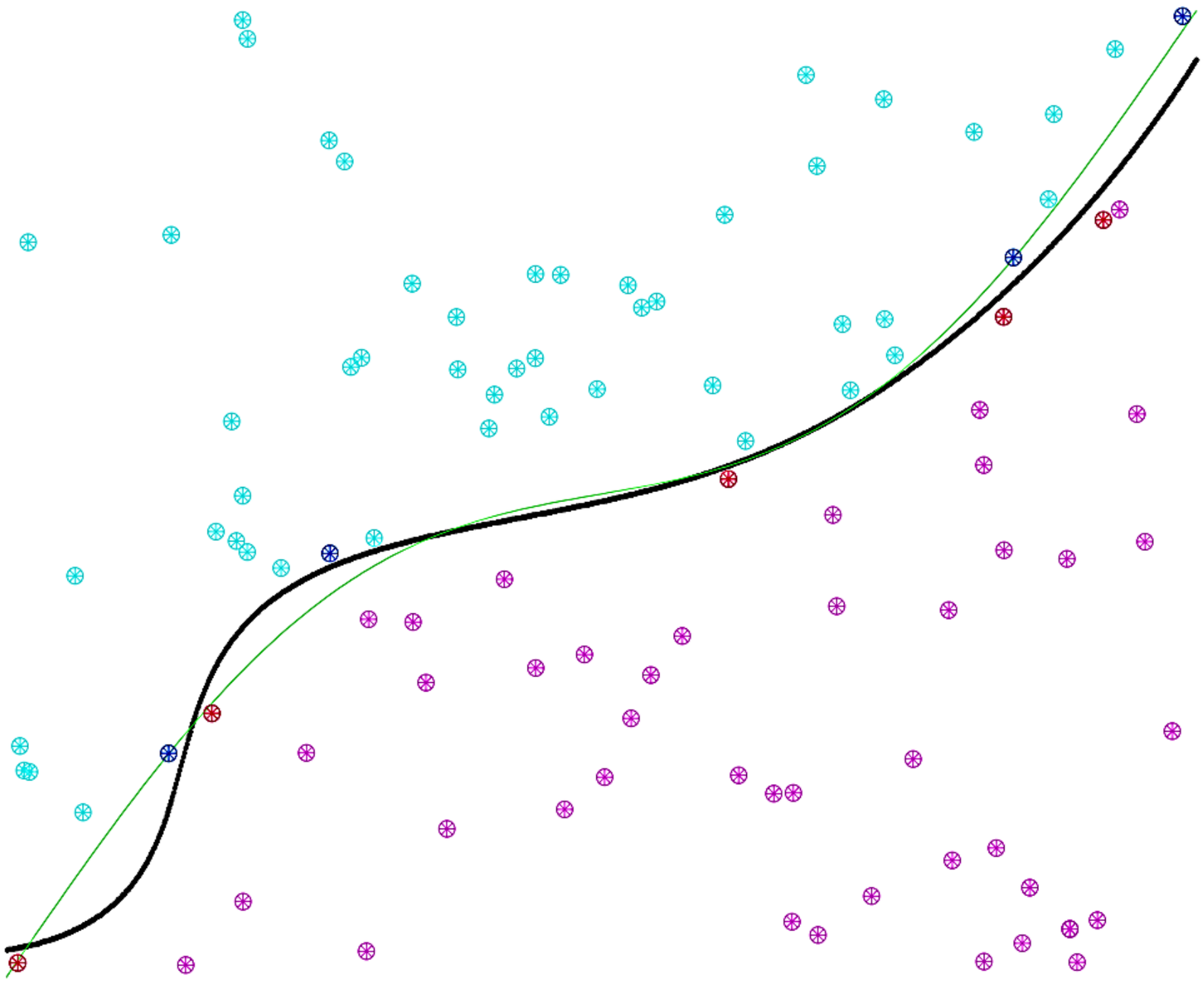
Radial basis kernel in action

Slightly non-linearly separable case for 100 datapoints:



Transforming χ

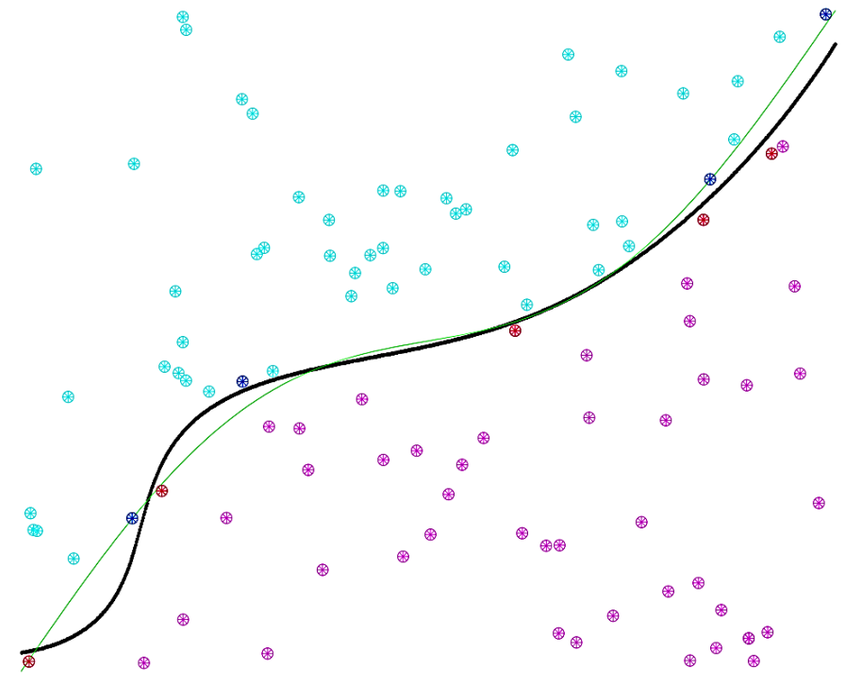
into ∞-dimensional $\mathbb{Z}$ space



Generalization

Are we killing the generalization by going to infinite-dimension? (overfitting)

I am going to answer this with a question. In this, example how many support vectors, we have?



What will happen if we have many support vectors?

The decision boundary line (plane) will be super wiggly ➔ overfitting alarm

$$\mathrm{E}[E_{out}] \leq \frac{\mathrm{E}[Number\ of\ support\ vectors]}{N-1}$$

N is number of datapoints

Kernel formulation of SVM

Remember quadratic programming?

$$\begin{bmatrix} y^{\{1\}}y^{\{1\}}x^{\{1\}}x^{\{1\}^T} & y^{\{1\}}y^{\{2\}}x^{\{1\}}x^{\{2\}^T} & \dots & y^{\{1\}}y^{\{N\}}x^{\{1\}}x^{\{N\}^T} \\ y^{\{2\}}y^{\{1\}}x^{\{2\}}x^{\{1\}^T} & y^{\{2\}}y^{\{2\}}x^{\{2\}}x^{\{2\}^T} & \dots & y^{\{2\}}y^{\{N\}}x^{\{2\}}x^{\{N\}^T} \\ \dots & \dots & \dots & \dots \\ y^{\{N\}}y^{\{1\}}x^{\{N\}}x^{\{1\}^T} & y^{\{N\}}y^{\{2\}}x^{\{N\}}x^{\{2\}^T} & \dots & y^{\{N\}}y^{\{N\}}x^{\{N\}}x^{\{N\}^T} \end{bmatrix}$$

Quadratic coefficients

In $\mathbb{Z}$ space, the only thing you need:

$$\begin{bmatrix} y^{\{1\}}y^{\{1\}}K(x^{\{1\}},x^{\{1\}^T}) & y^{\{1\}}y^{\{2\}}K(x^{\{1\}},x^{\{2\}^T}) & \dots & y^{\{1\}}y^{\{N\}}K(x^{\{1\}},x^{\{N\}^T}) \\ y^{\{2\}}y^{\{1\}}K(x^{\{2\}},x^{\{1\}^T}) & y^{\{2\}}y^{\{2\}}K(x^{\{2\}},x^{\{2\}^T}) & \dots & y^{\{2\}}y^{\{N\}}K(x^{\{2\}},x^{\{N\}^T}) \\ \dots & \dots & \dots & \dots \\ y^{\{N\}}y^{\{1\}}K(x^{\{N\}},x^{\{1\}^T}) & y^{\{N\}}y^{\{2\}}K(x^{\{N\}},x^{\{2\}^T}) & \dots & y^{\{N\}}y^{\{N\}}K(x^{\{N\}},x^{\{N\}^T}) \end{bmatrix}$$

Final stage:

$$g(x) = sign(z\theta + b) \qquad in\ terms\ of\ K(-,-)$$

$$where \rightarrow \theta = \sum_{z^{\{i\}} in\ SV} \alpha^{\{i\}} y^{\{i\}} z^{\{i\}} \rightarrow \quad g(x) = sign(\sum_{z^{\{i\}} in\ SV} \alpha^{\{i\}} y^{\{i\}} z^{\{i\}}\ z^T + b)$$

$$g(x) = sign(\sum_{\alpha_i > 0} \alpha^{\{i\}} y^{\{i\}} K(x^{\{i\}}, x) + b)$$

$$and\ b: \quad b = y^{\{j\}} - \sum_{\alpha^{\{i\}} > 0, \alpha^{\{j\}} > 0} \alpha^{\{i\}} y^{\{i\}} K(x^{\{i\}}, x^{\{j\}})$$

$$for\ any\ SV\ x^{\{i\}} and\ x^{\{j\}}$$

How do we know that the kernel is valid?

$For\ a\ given\ K(x^{\{i\}}, x^{\{j\}})$ → We can check the validity

Three approaches:

1. By construction (Polynomial one)

2. Math properties (Mercer's condition)

3. Who cares? ☺

Design your kernel

$K(x^{\{1\}}, x^{\{2\}})$ is valid iff

1. It is symmetric ➔ $K(x^{\{1\}}, x^{\{2\}}) = K(x^{\{2\}}, x^{\{1\}})$

2. The matrix ➔ $\begin{bmatrix} y^{\{1\}}y^{\{1\}}K(x^{\{1\}}, x^{\{1\}^T}) & y^{\{1\}}y^{\{2\}}K(x^{\{1\}}, x^{\{2\}^T}) & \dots & y^{\{1\}}y^{\{N\}}K(x^{\{1\}}, x^{\{N\}^T}) \\ y^{\{2\}}y^{\{1\}}K(x^{\{2\}}, x^{\{1\}^T}) & y^{\{2\}}y^{\{2\}}K(x^{\{2\}}, x^{\{2\}^T}) & \dots & y^{\{2\}}y^{\{N\}}K(x^{\{2\}}, x^{\{N\}^T}) \\ \dots & \dots & \dots & \dots \\ y^{\{N\}}y^{\{1\}}K(x^{\{N\}}, x^{\{1\}^T}) & y^{\{N\}}y^{\{2\}}K(x^{\{N\}}, x^{\{2\}^T}) & \dots & y^{\{N\}}y^{\{N\}}K(x^{\{N\}}, x^{\{N\}^T}) \end{bmatrix}$

is positive-semi definite

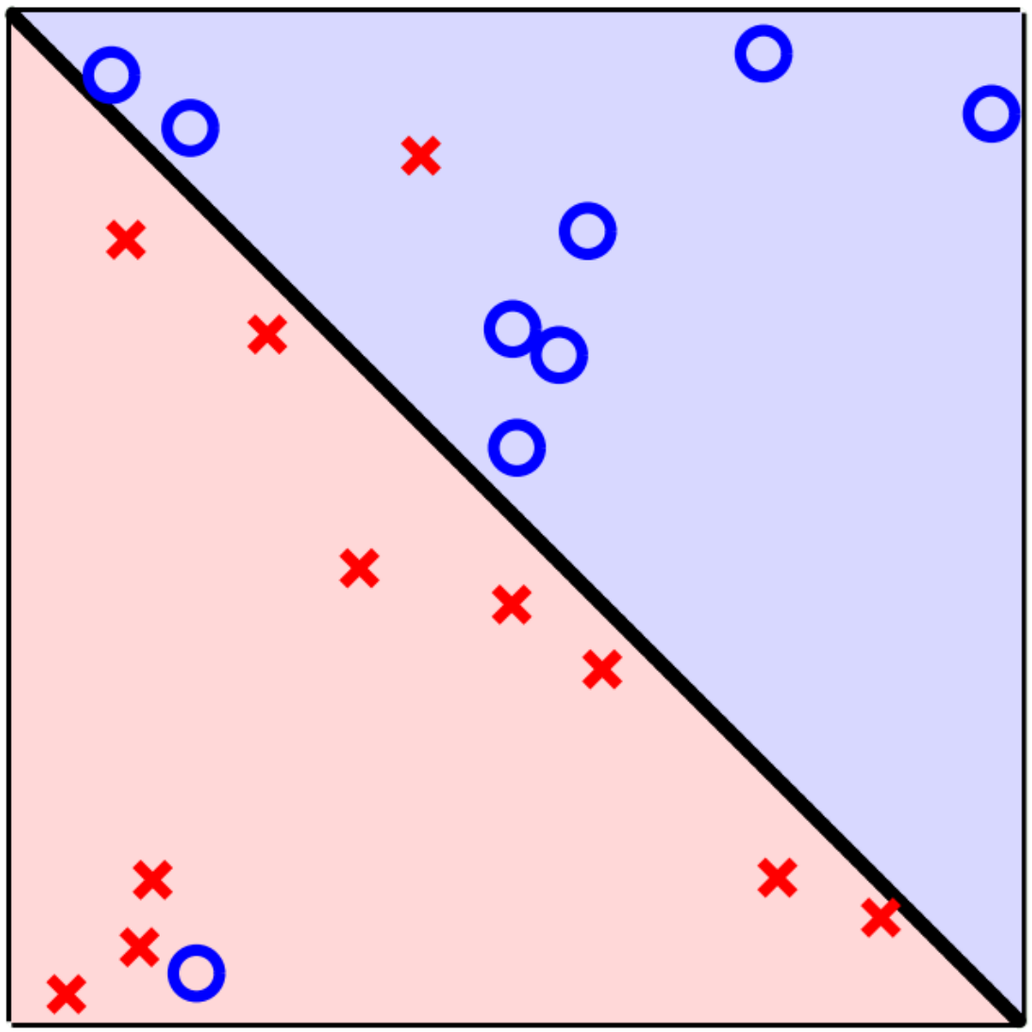
For any $x^{\{1\}}, \dots, x^{\{N\}}$ (Mercer's condition)

Common Kernels

- Linear kernels $K\left(x^{\{1\}}, x^{\{2\}}\right) = x^{\{1\}} x^{\{2\}^T}$
- Polynomial kernels $K\left(x^{\{1\}}, x^{\{2\}}\right) = \left(1 + x^{\{1\}} x^{\{2\}^T}\right)^T$ for any d>0
  - Contains all polynomials terms up to degree d
- Gaussian kernels $K\left(x^{\{1\}}, x^{\{2\}}\right) = \exp\left(-\frac{\left\|x^{\{1\}} - x^{\{2\}}\right\|^2}{2 * \sigma^2}\right) for\ \sigma > 0$
  - Infinite dimensional feature space

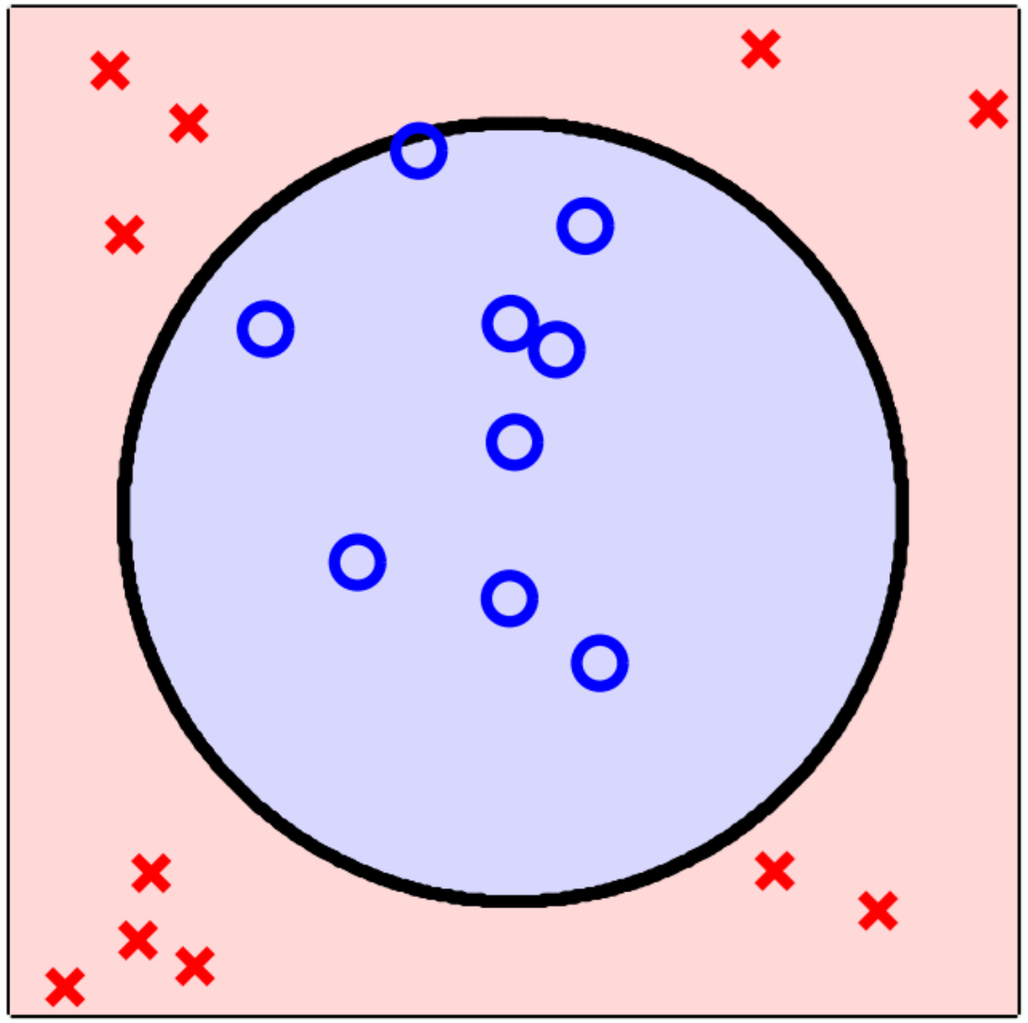
Soft SVM – Two types of non-separable

slightly:



Soft SVM will deal with this

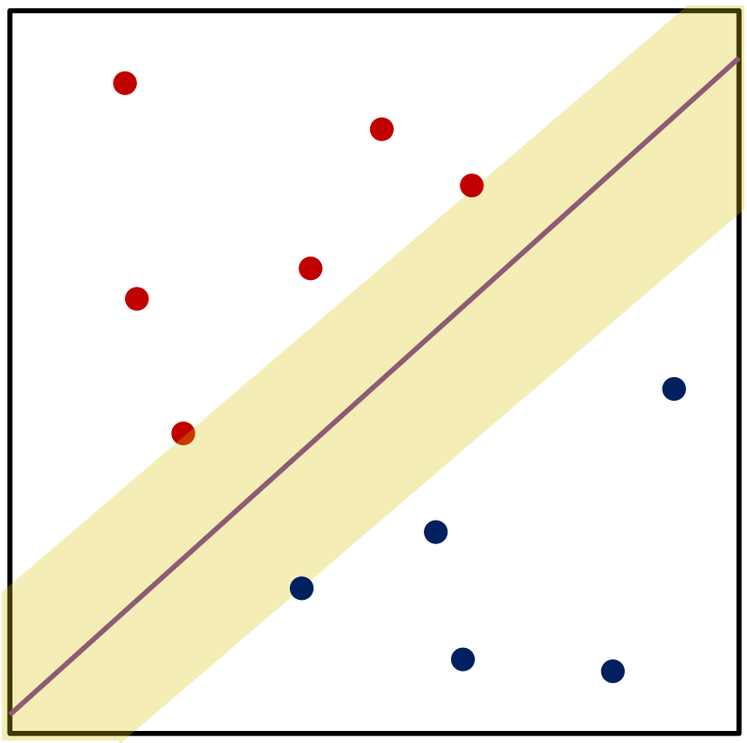
seriously:



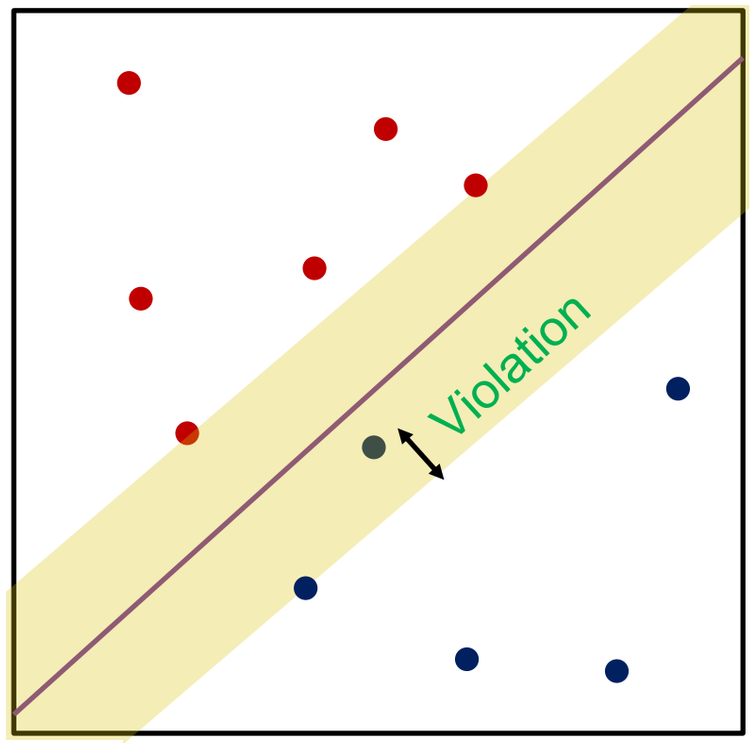
Kernel will deal with this

Error measure

Non-violated case:



Margin violation:



if $y^{\{i\}}\left(x^{\{i\}}\theta + b\right) > 1 \Rightarrow$ Non SV

Let's introcduce a slack variable: $y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \geq 1 - \xi^{\{i\}} \qquad \xi^{\{i\}} \geq 0$

$$\text{Total violation} = \sum_{i=1}^{N} \xi^{\{i\}}$$

The new optimization

Minimize $\frac{1}{2}\theta\theta^T + \mathrm{C}\sum_{i=1}^{N}\xi^{\{i\}}$

C will define the relative importance of the first or second term

$\mathrm{C} = inf$ is equal to Hard SVM (will see soon)

$s.t.\ \ y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \geq 1 - \xi^{\{i\}} \quad \text{for } i = 1, \ldots, N$

$and \xi^{\{i\}} \geq 0 \qquad \text{for } i = 1, \ldots, N$

$\theta \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^N$

The Lagrange formulation

Hard svm: Minimize $\frac{1}{2}\theta\theta^T$ $s.t.\ \ y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \geq 1$

$$\mathcal{L}(\theta, b, \alpha) = \frac{1}{2}\theta\theta^T - \sum_{i=1}^{N} \alpha^{\{i\}}(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1)$$

Soft svm: Minimize $\frac{1}{2}\theta\theta^T + \mathrm{C}\sum_{i=1}^{N} \xi^{\{i\}}$ $s.t.\ \ y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \geq 1 - \xi^{\{i\}} and$ $\xi^{\{i\}} \geq 0$

$$\mathcal{L}(\theta, b, \xi, \alpha) = \frac{1}{2}\theta\theta^t \ + C\sum_{i=1}^{N} \xi^{\{i\}} - \sum_{i=1}^{N} \alpha^{\{i\}}\left(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1 \ + \xi^{\{i\}}\right) - \sum_{i=1}^{N} \beta^{\{i\}}\xi^{\{i\}}$$

PLEASE do not scare, terms will be dropping fast

$$\mathcal{L}(\theta, b, \xi, \alpha) = \frac{1}{2}\theta\theta^T + C\sum_{i=1}^{N} \xi^{\{i\}} - \sum_{i=1}^{N} \alpha^{\{i\}}\left(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1 + \xi^{\{i\}}\right) - \sum_{i=1}^{N} \beta^{\{i\}}\xi^{\{i\}}$$

Minimize w.r.t $\boldsymbol{\theta}, b, and\ \xi$ and maximize w.r.t $\alpha^{\{i\}} \geq 0\ and \beta^{\{i\}} \geq 0$

KKT condition for inequality constraints

Let's do the minimization:

$$\nabla_\theta\, \mathcal{L}(\theta, b, \xi, \alpha) = \theta - \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}} = 0$$

If we substitute $\beta^{\{i\}}$ up there, the whole formulation will get back to hard svm

$$\nabla_b\, \mathcal{L}(\theta, b, \xi, \alpha) = -\sum_{n=1}^{N} \alpha^{\{i\}} y^{\{i\}} = 0$$

$$\nabla_\xi\, \mathcal{L}(\theta, b, \xi, \alpha) = C - \alpha^{\{i\}} - \beta^{\{i\}}$$

We should say thank you $\beta^{\{i\}}$ for the great service

The solution

$$\beta^{\{i\}} = C - \alpha^{\{i\}}$$

$$\beta^{\{i\}} \geq 0 \quad \rightarrow \quad C - \alpha^{\{i\}} \geq 0 \rightarrow 0 \leq \alpha^{\{i\}} \leq C \qquad \text{for } i = 1, \dots, N$$

$$Maximize\ \mathcal{L}(\alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{\{i\}} y^{\{j\}} \alpha^{\{i\}} \alpha^{\{j\}} x^{\{i\}} x^{\{j\}^T} \qquad \text{w.r.t} \quad \alpha$$

$$s.t. \qquad 0 \leq \alpha^{\{i\}} \leq C \qquad \text{for } i = 1, \dots, N \qquad \boldsymbol{and} \qquad \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} = 0$$

$$\Rightarrow \theta = \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}} \qquad \text{will minimize} \qquad \frac{1}{2} \theta \theta^T + C \sum_{i=1}^{N} \xi^{\{i\}}$$

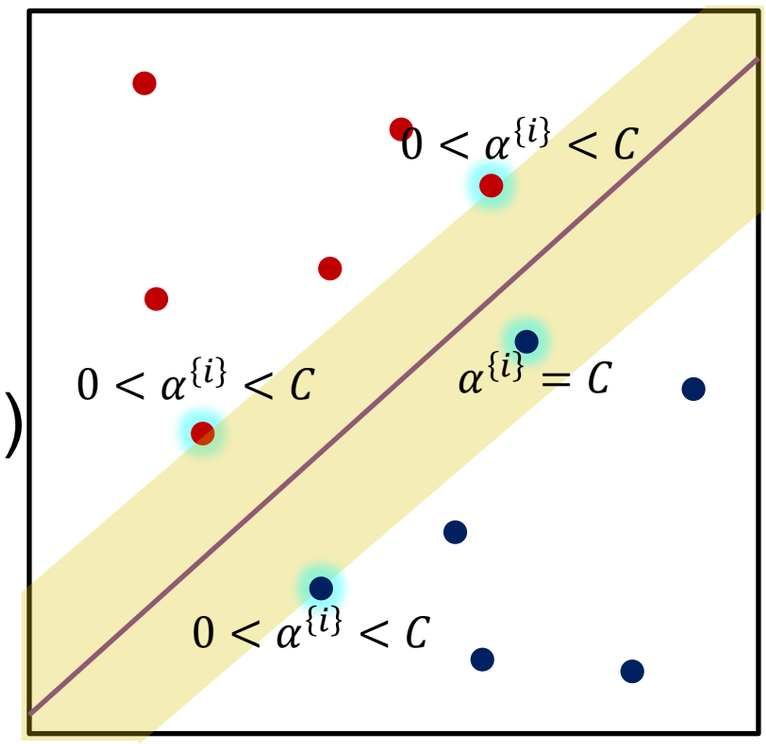
Type of support vectors

We call the three points as **margin** support vectors

$$0 < \alpha^{\{i\}} < C$$

$$\beta^{\{i\}} = C - \alpha^{\{i\}}$$

$$y^{\{i\}}\left(x^{\{i\}}\theta + b\right) = 1 \Rightarrow \beta^{\{i\}} > 0 \Rightarrow \xi^{\{i\}} = 0 \text{ (KKT condition)}$$



non-margin support vectors $(\alpha^{\{i\}} = C)$

$$\beta^{\{i\}} = 0 \Rightarrow \xi^{\{i\}} > 0 \text{ (KKT condition)}$$

$$y^{\{i\}}\left(x^{\{i\}}\theta + b\right) > 1 - \xi^{\{i\}} \text{ if } \xi^{\{i\}} > 0$$

$$y^{\{i\}}\left(x^{\{i\}}\theta + b\right) < 1$$

$$\alpha^{\{i\}} = 0 \Rightarrow y^{\{i\}}\left(x^{\{i\}}\theta + b\right) > 1$$

Non SV

$$\alpha^{\{i\}} = C \Rightarrow y^{\{i\}}\left(x^{\{i\}}\theta + b\right) < 1$$

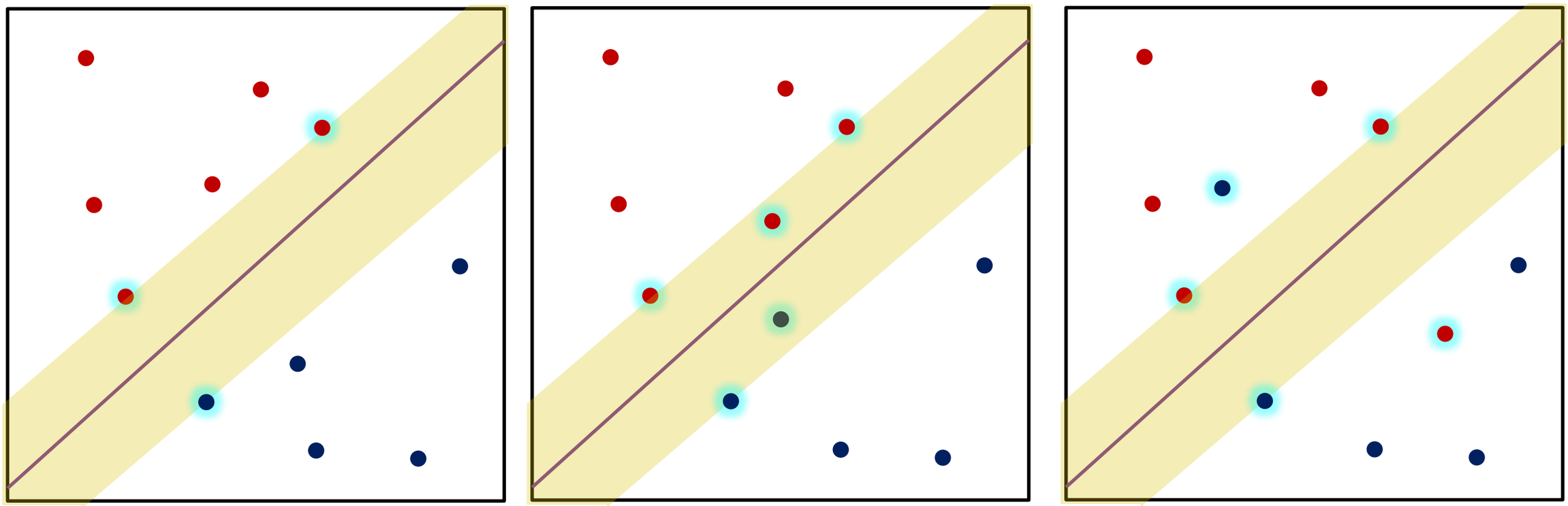
SV on the wrong side

$$0 < \alpha^{\{i\}} < C \Rightarrow y^{\{i\}}\left(x^{\{i\}}\theta + b\right) = 1$$

SV on the margin

Any violating points become support vectors

How to choose C?



violating points become
support vectors

How to define the hyper-parameter C: Cross Validation

Primal and Dual Forms of SVM

Primal version of classifier:

$$f(x_{test}) = x_{test}\theta + \theta_0$$

Dual version of classifier:

$$f(x_{test}) = \sum_{x^{\{i\}} in\ SV} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}} {x_{test}}^T + b$$

Kernel SVM: Summary

- Classifiers can be learnt for high dimensional features spaces, without actually having to map the points into the high dimensional space

- Data may be linearly separable in the high dimensional space, but not linearly separable in the original feature space

- Kernels can be used for an SVM because of the scalar product in the dual form, but can also be used elsewhere – they are not tied to the SVM formalism