Advanced Machine learning Mastering Course

Introduced by

George Samuel

Master in computer science
Cairo University

Innovisionray.com

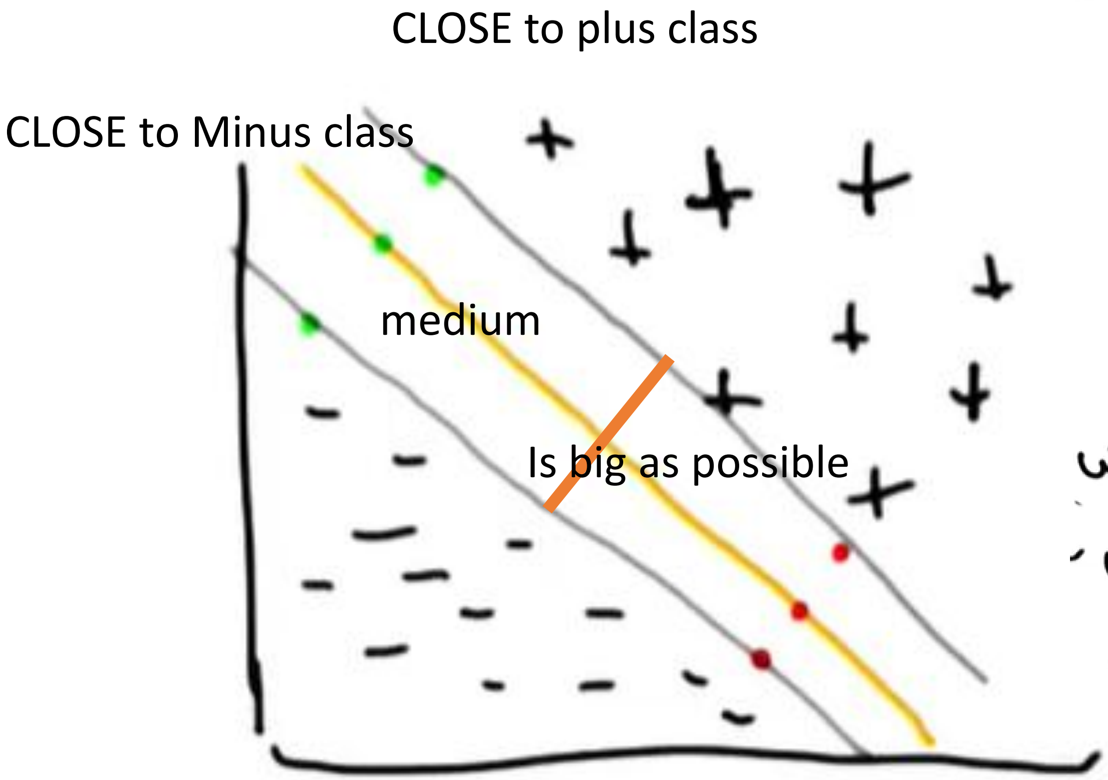
2024

Agenda

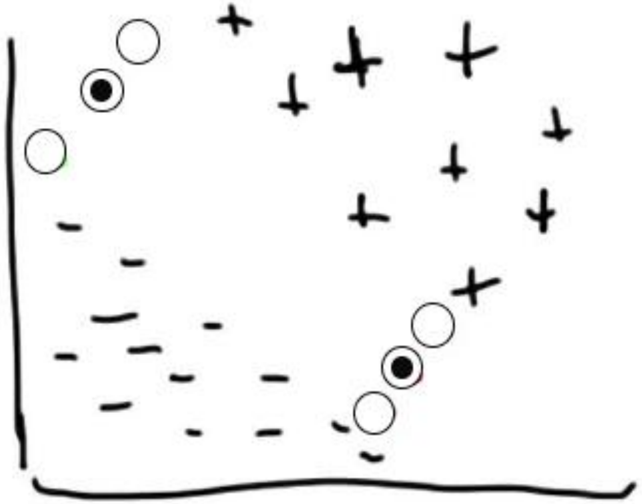
$$w^T x_1 + b = 1$$
$$w^T x_2 + b = -1$$

QUIZ !

WHICH IS THE BEST



$$w^T x + b = 1$$
$$w^T x + b = 0$$
$$w^T x + b = -1$$



SVM

Quiz

$$
\begin{aligned}
& \quad w^T x_1 + b = 1 \\
- & \quad w^T x_2 + b = -1 \\
\hline
\end{aligned}
$$

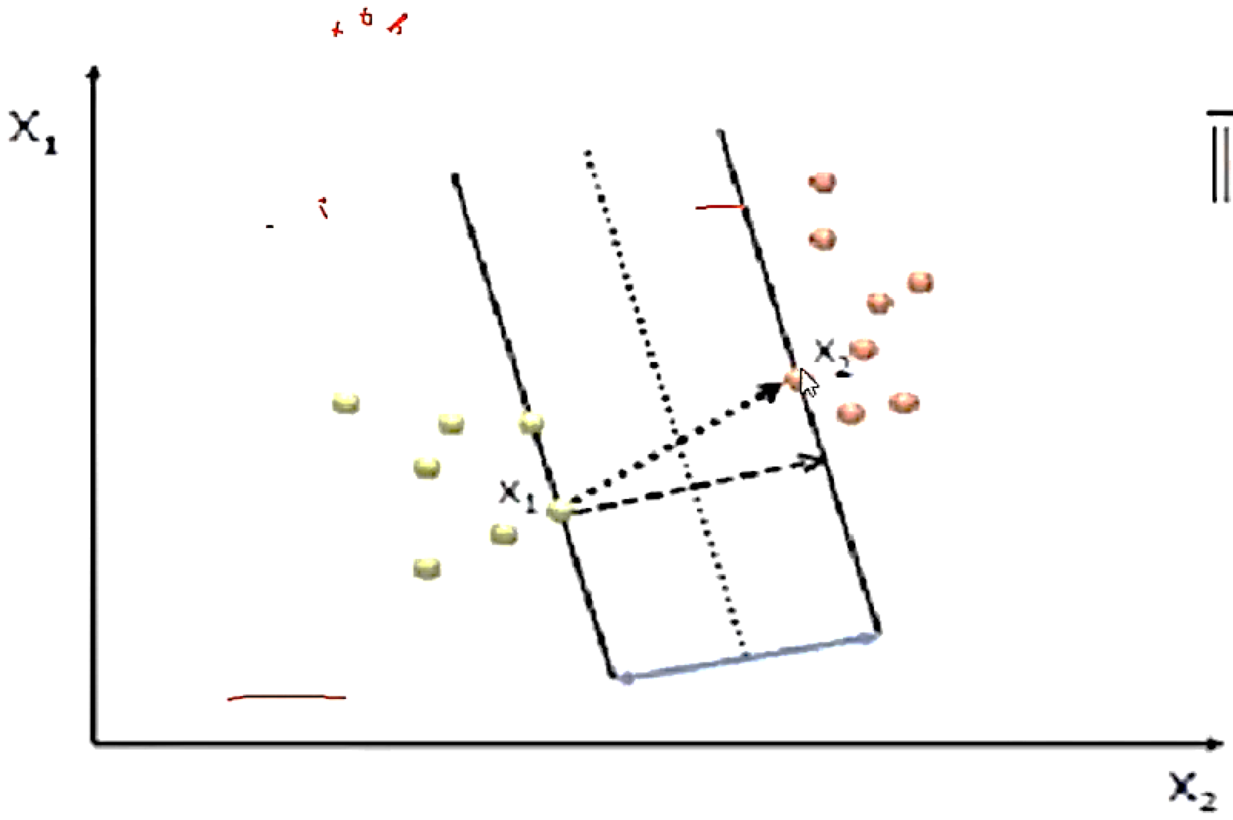
wtx1-wtx2=2

$$w^T(x_1 - x_2) = 2$$

$$\frac{w^T(x_1 - x_2)}{\|w\|} = \frac{2}{\|w\|}$$

$$\underbrace{\frac{w^T(x_1 - x_2)}{\|w\|}}_{\text{margin}} = \frac{2}{\|w\|}$$

SVM



$$\frac{w}{\|w\|}\cdot(x_2 - x_1) = \text{width} = \frac{2}{\|w\|}$$

$$w\cdot x_2 + b = 1$$

$$w\cdot x_1 + b = -1$$

$$w\cdot x_2 + b - w\cdot x_1 - b = 1-(-1)$$

$$w\cdot x_2 - w\cdot x_1 = 2$$

$$\frac{w}{\|w\|}(x_2 - x_1) = \frac{2}{\|w\|}$$

SVM

$\max \frac{2}{\|w\|}$ while classifying everything correctly

$y_i(w^T x_i + b) \geq 1 \quad \forall i$

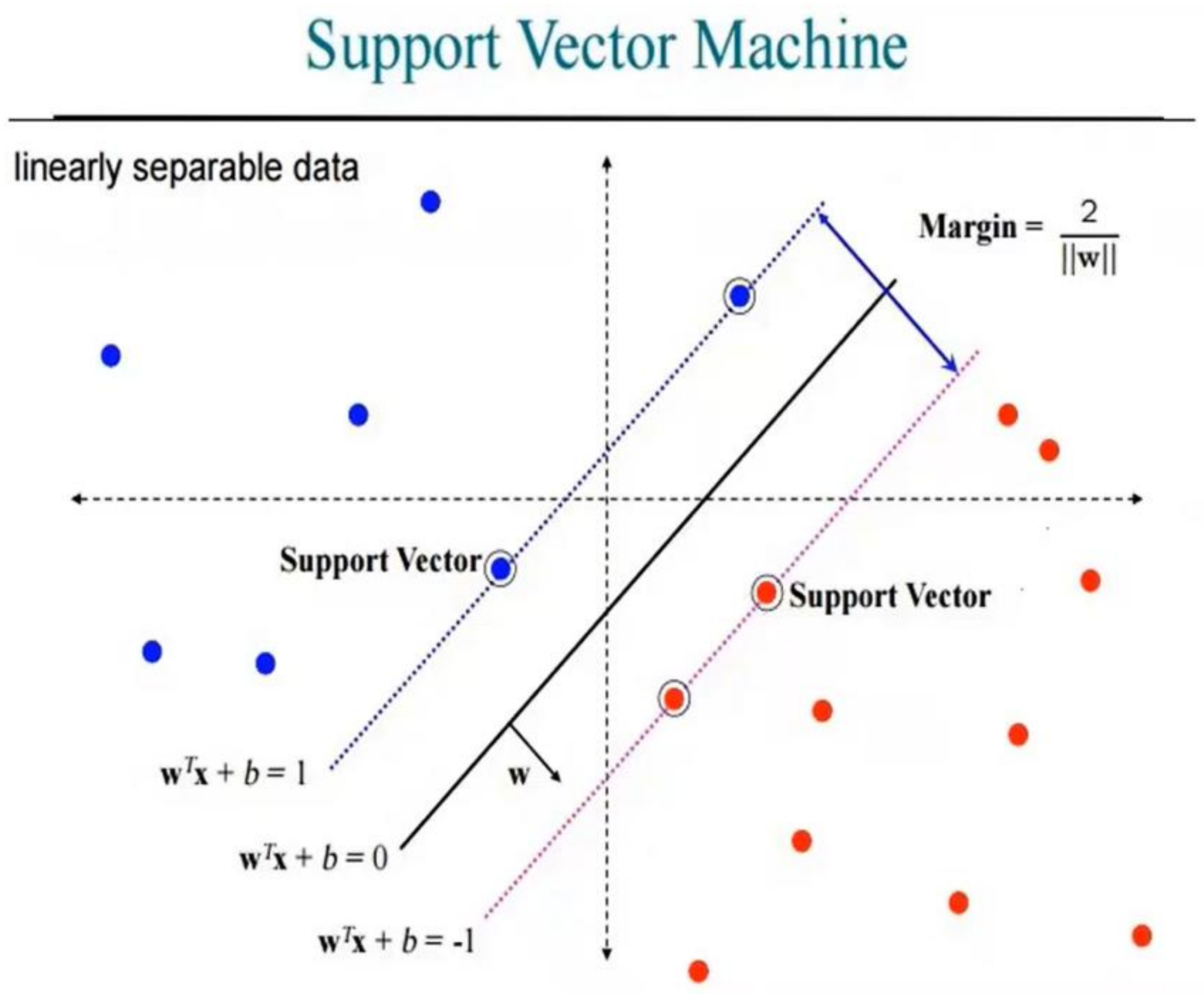
- $\min \frac{1}{2} \|w\|^2$

quadratic programming

- $\max W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j$

s.t. $\alpha_i \geq 0, \quad \sum_i \alpha_i y_i = 0$

SVM


Support Vector Machine
linearly separable data
Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$
w

SVM

Support Vector Machines (SVM) can be framed as a minimization problem in the context of both classification and regression. The objective is to find the optimal hyperplane that separates data points of different classes with the maximum margin in the case of classification, or to fit the best possible regression line with minimal error in the case of regression. Here's how SVM can be formulated as a minimization problem for both cases:

1. SVM for Classification

Primal Form

The primal form of the SVM optimization problem can be written as:

Minimize:
$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$

Subject to:
$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$
$\xi_i \geq 0, \quad \forall i$

where:

- $\mathbf{w}$ is the weight vector.
- b is the bias term.
- C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error.
- ξ_i are the slack variables that allow for some misclassification.
- y_i are the class labels ($+1$ or -1).
- $\mathbf{x}_i$ are the feature vectors of the training samples.

Dual Form

The dual form of the SVM optimization problem can be written as:

Maximize:
$\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j$

Subject to:
$\sum_{i=1}^{n}\alpha_i y_i = 0$
$0 \leq \alpha_i \leq C, \quad \forall i$

where:

- α_i are the Lagrange multipliers.

SVM

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{s.t.} \qquad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall_i$$

In order to cater for the constraints in this minimization, we need to allocate them Lagrange multipliers $\boldsymbol{\alpha}$, where $\alpha_i \geq 0 \ \forall_i$:

$$\begin{aligned} L_P &\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \boldsymbol{\alpha}\left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \ \forall_i\right] \\ &\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{L} \alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1\right] \\ &\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{L} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{L} \alpha_i \end{aligned}$$

We wish to find the **w** and **b** which minimizes, and the **α** which maximizes LP(whilst keeping αi ≥ 0 ∀i We can do this by differentiating LP with respect to w and b and setting the derivatives to zero:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0$$

2. SVM for Regression (Support Vector Regression, SVR)

In the case of SVR, the objective is to find a function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ that deviates from the actual observed values y_i by at most ϵ for all training data points.

Primal Form

The primal form of the SVR optimization problem can be written as:

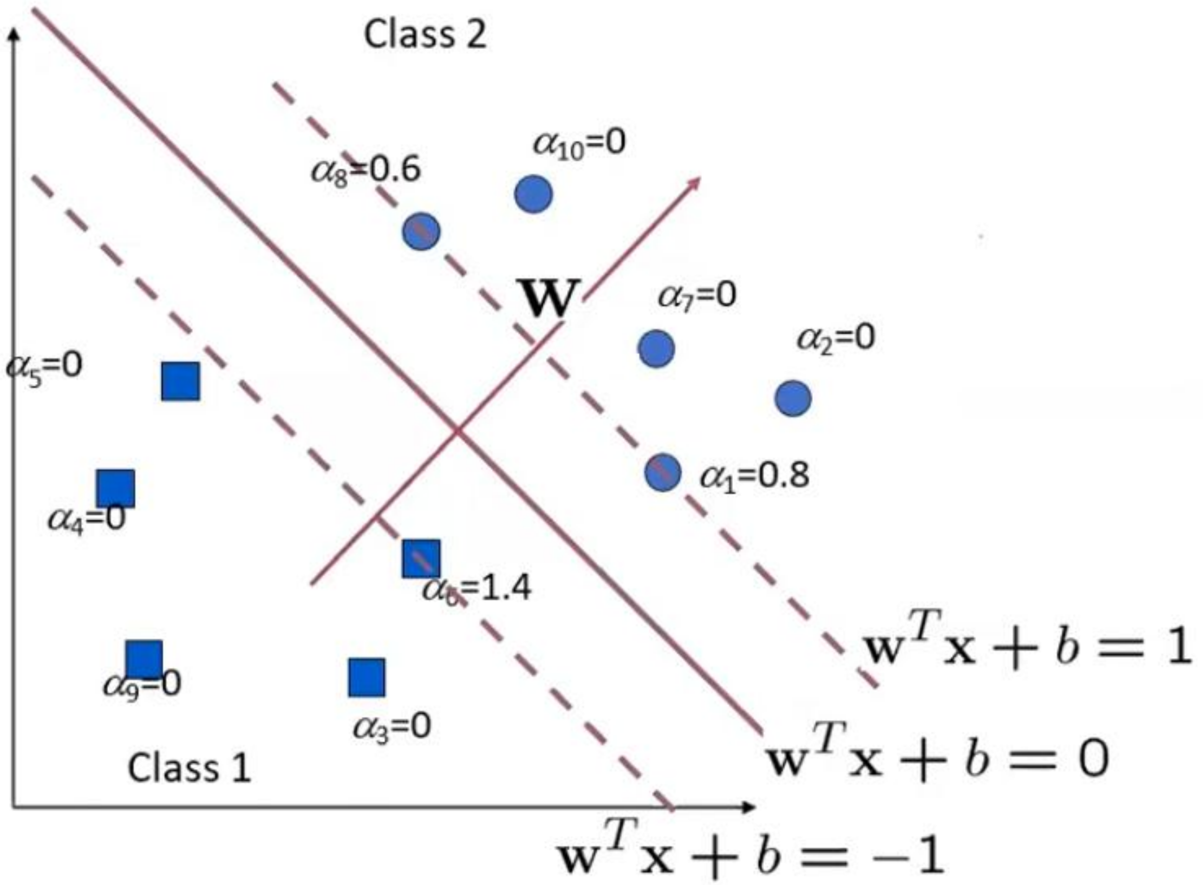
Minimize:
$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$

Subject to:
$y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \leq \epsilon + \xi_i$
$(\mathbf{w} \cdot \mathbf{x}_i + b) - y_i \leq \epsilon + \xi_i^*$
$\xi_i, \xi_i^* \geq 0, \quad \forall i$

where:

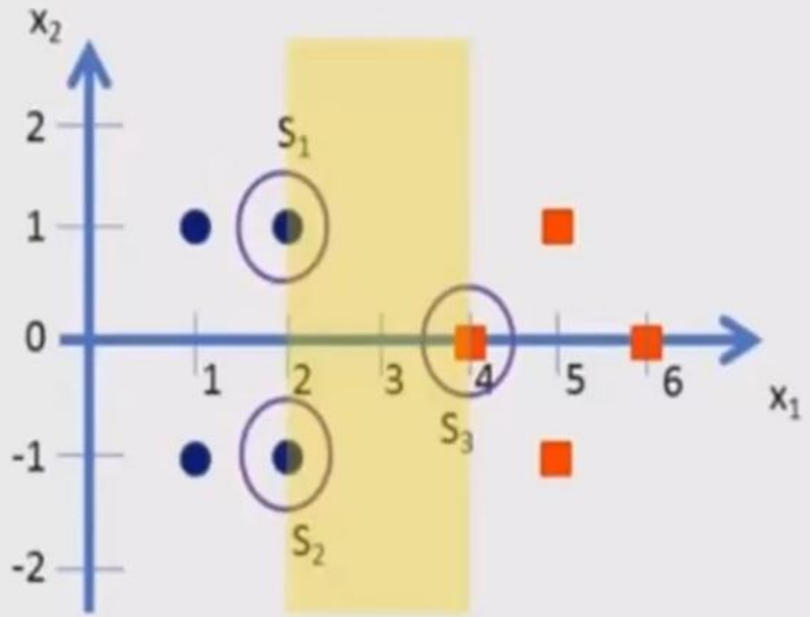
- $\mathbf{w}$ is the weight vector.
- b is the bias term.
- C is the regularization parameter.
- ξ_i and ξ_i^* are the slack variables that measure the deviations of predictions from the actual values.
- ϵ is the margin of tolerance.

A Geometrical Interpretation



Example

- Here we select 3 Support Vectors to start with.
- They are S_1, S_2 and S_3.



$$S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

SVM

$$S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

$$\widetilde{S_1} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

$$\widetilde{S_2} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

$$\widetilde{S_3} = \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

Now we need to find 3 parameters α1, α2 and α3 based on the following 3 linear equations:

$$\alpha_1 \widetilde{S_1}.\widetilde{S_1} + \alpha_2 \widetilde{S_2}.\widetilde{S_1} + \alpha_3 \widetilde{S_3}.\widetilde{S_1} = -1 \ \ (-ve\ class)$$

$$\alpha_1 \widetilde{S_1}.\widetilde{S_2} + \alpha_2 \widetilde{S_2}.\widetilde{S_2} + \alpha_3 \widetilde{S_3}.\widetilde{S_2} = -1 \ (-ve\ class)$$

$$\alpha_1 \widetilde{S_1}.\widetilde{S_3} + \alpha_2 \widetilde{S_2}.\widetilde{S_3} + \alpha_3 \widetilde{S_3}.\widetilde{S_3} = +1 \ \ (+ve\ class)$$

Let's substitute the values for $\widetilde{S_1}$, $\widetilde{S_2}$ and $\widetilde{S_3}$ in the above equations.

SVM

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix} = +1$$

After simplification we get

$$6\alpha_1 + 4\alpha_2 + 9\alpha_3 = -1$$

$$4\alpha_1 + 6\alpha_2 + 9\alpha_3 = -1$$

$$9\alpha_1 + 9\alpha_2 + 17\alpha_3 = +1$$

SVM

Simplifying the above 3 simultaneous equation we get

α1= -3.25 α2= -3.25 and α3=3.5

The hyperplane that discriminates the positive class from negative class is given by:

$$\tilde{w} = \sum_i \alpha_i \tilde{S}_i$$

SVM

$$\tilde{w} = \alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$

$$\tilde{w} = (-3.25).\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + (-3.25).\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + (3.5).\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$$

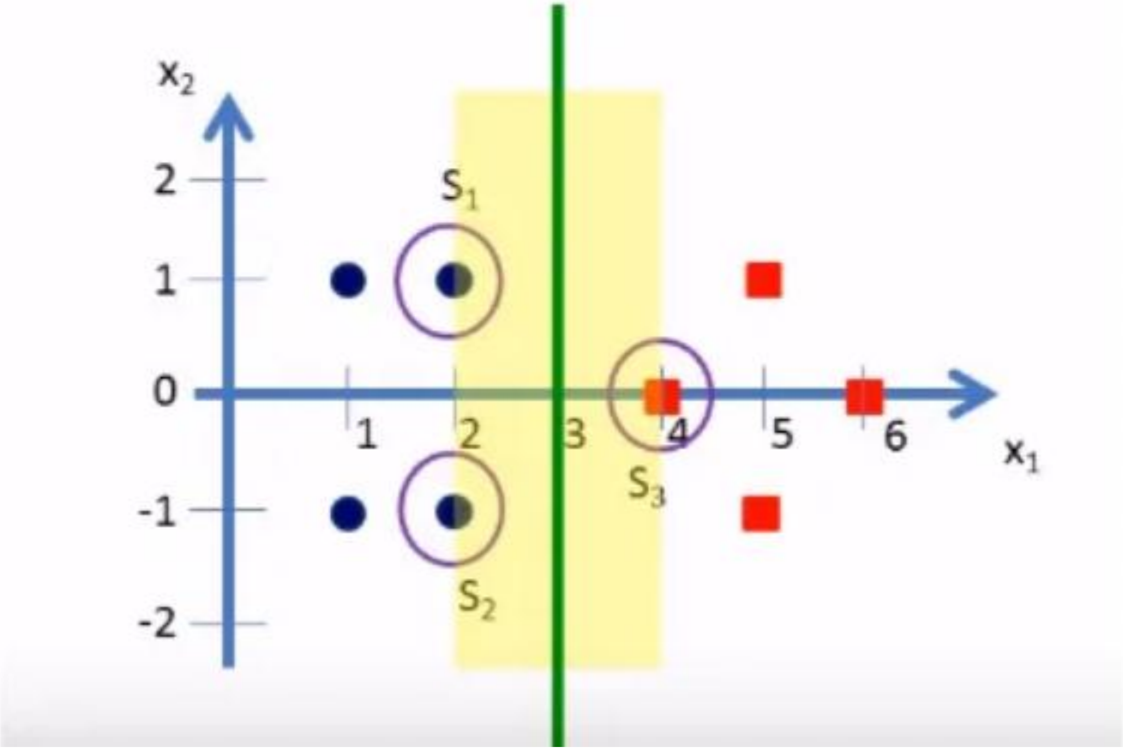
Our Vectors are augmented with a bias.

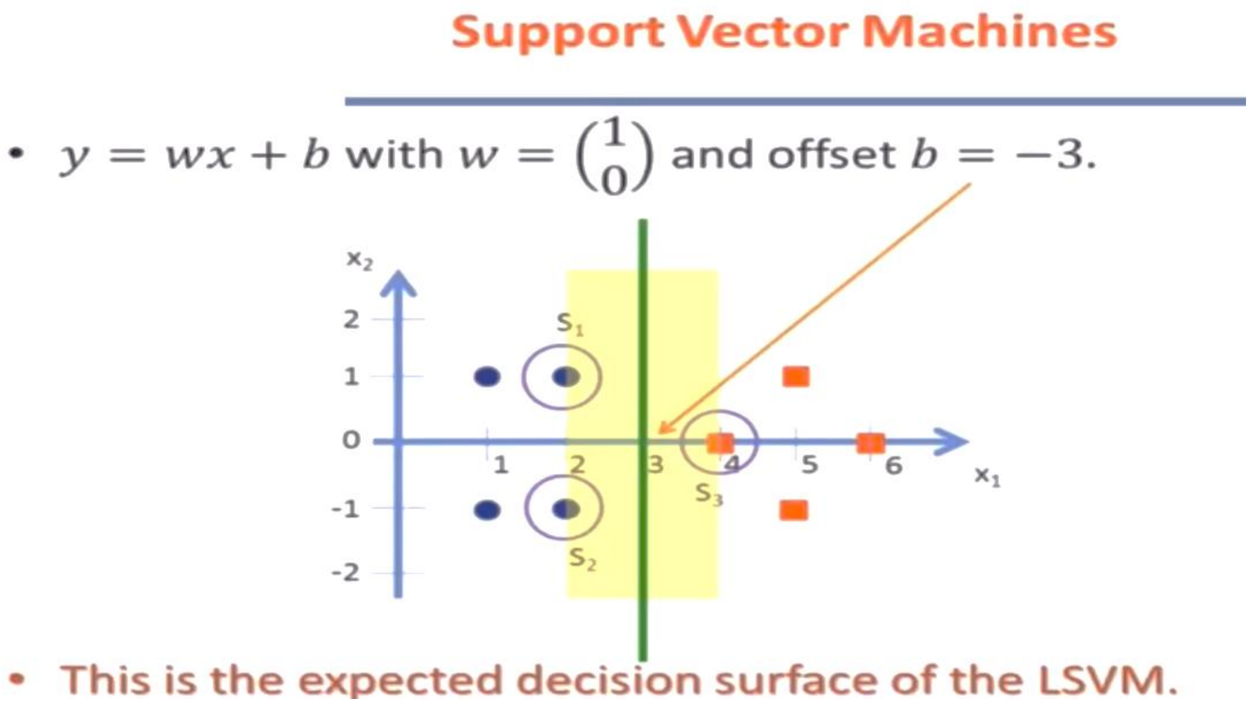
Hence we can equate the entry in $\tilde{w}$ as the hyperplane with an offset b.

Therefore the separating hyperplane equation

$$\mathbf{y = wx + b} \qquad \mathbf{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \textbf{and offset b} = -3$$

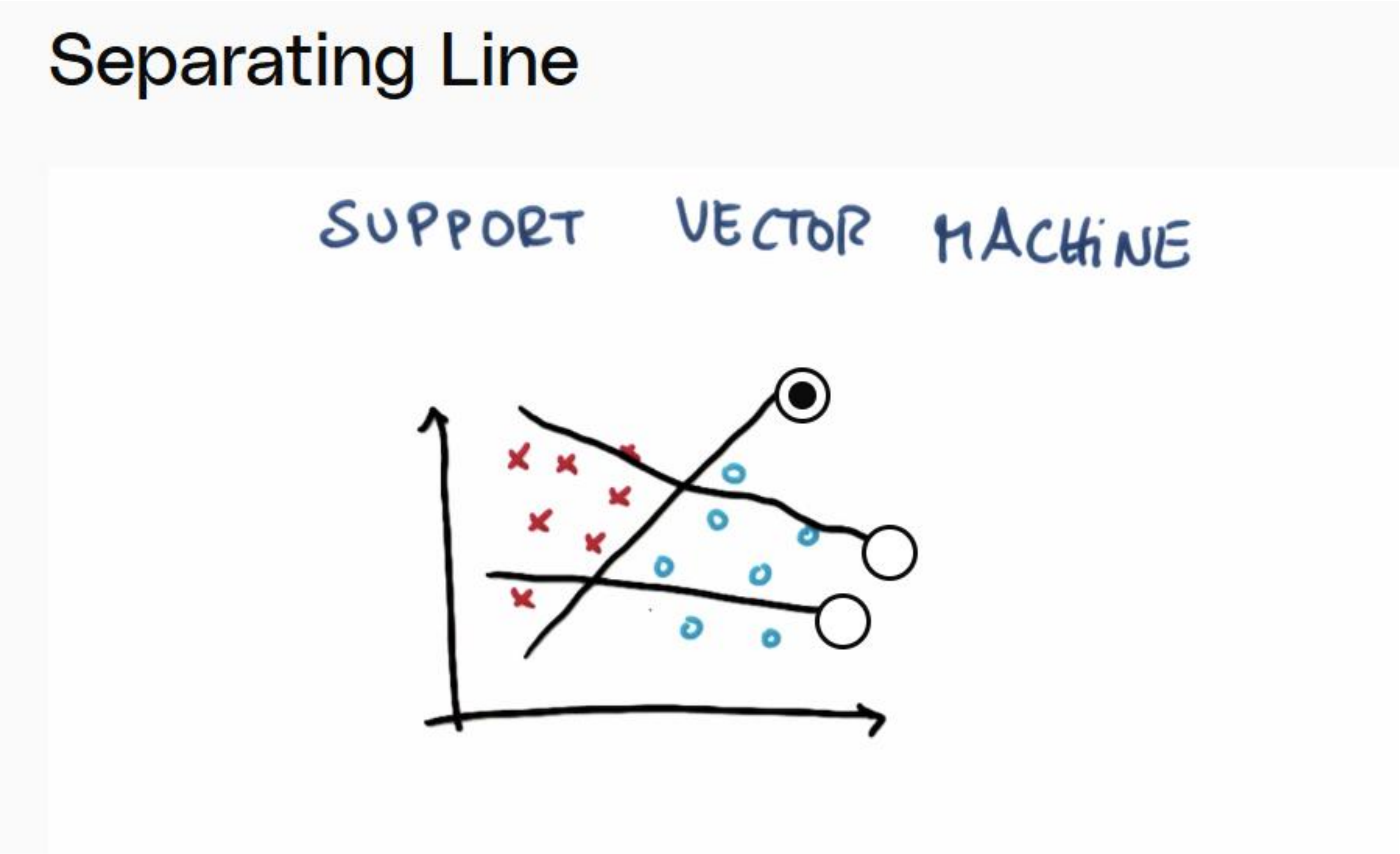
SVM



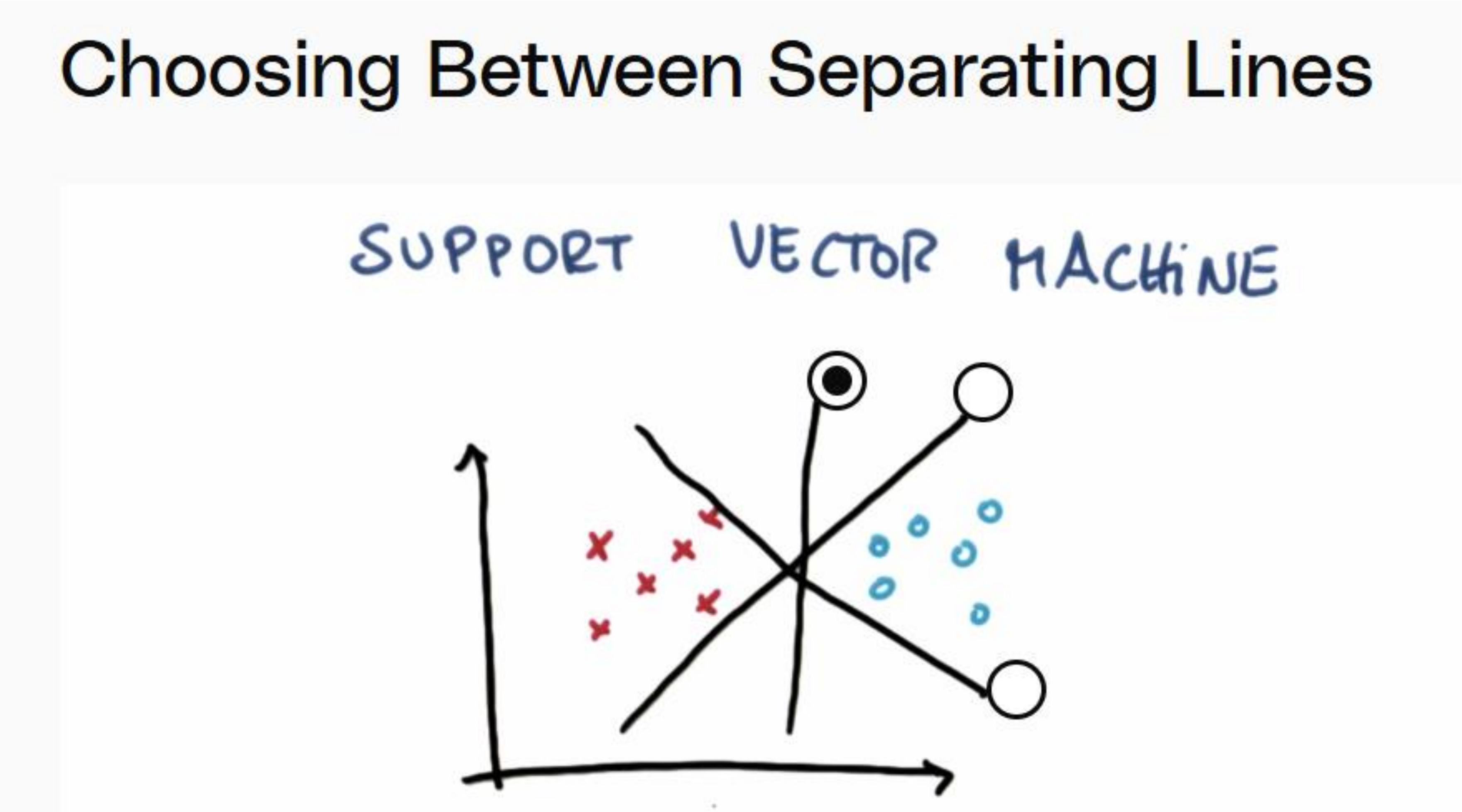


This is the expected decision surface of the LSVM.

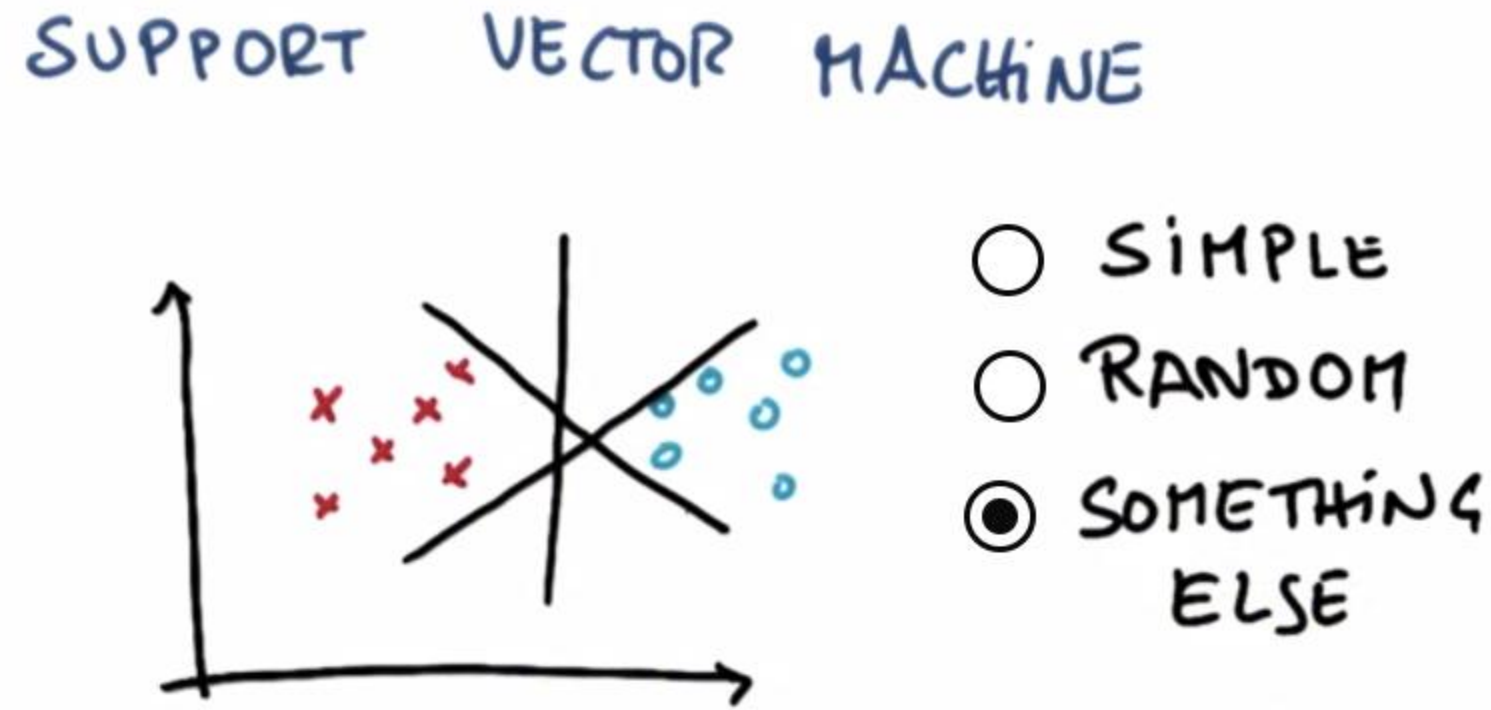
Quiz

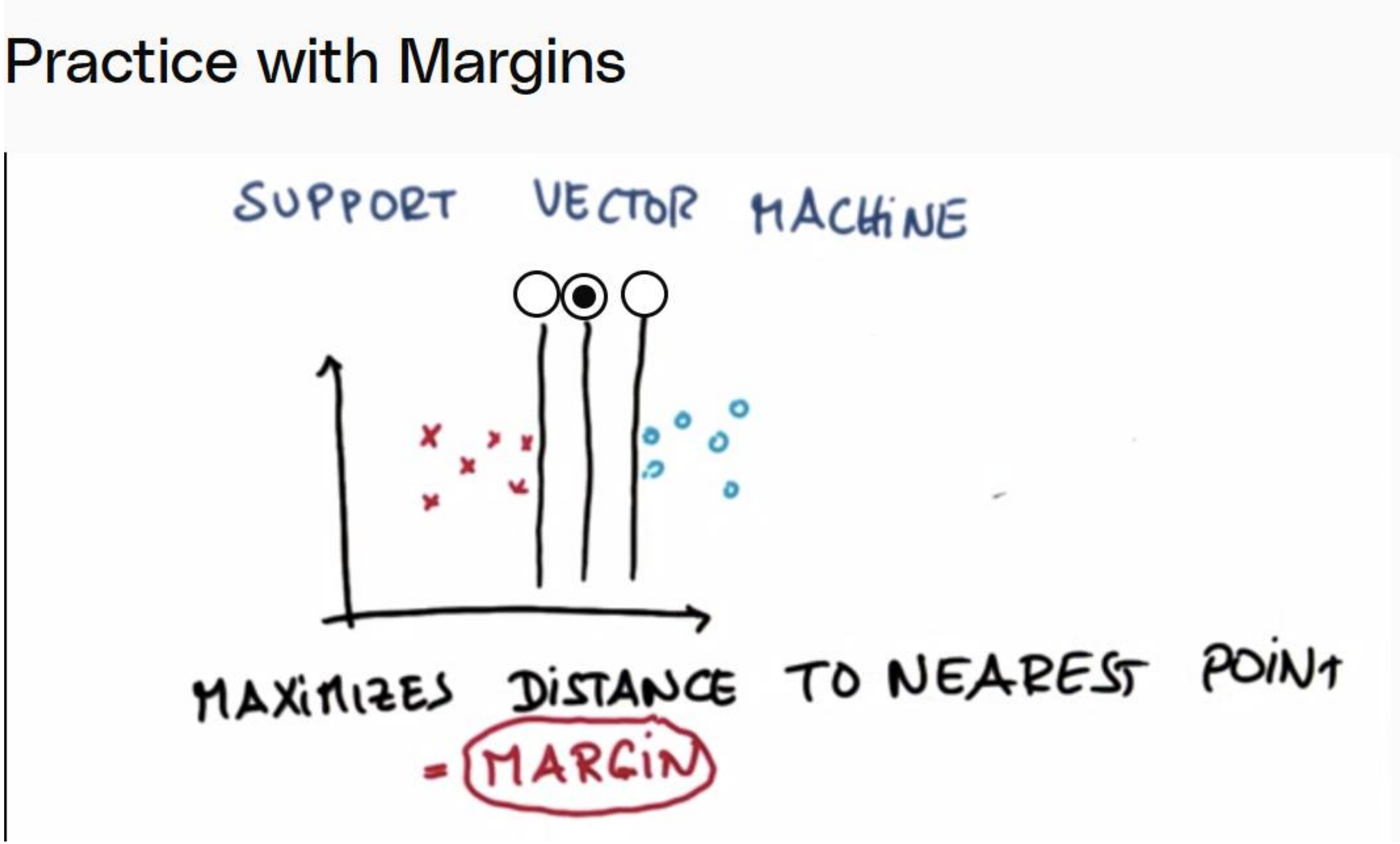


SVM


Choosing Between Separating Lines
SUPPORT VECTOR MACHINE

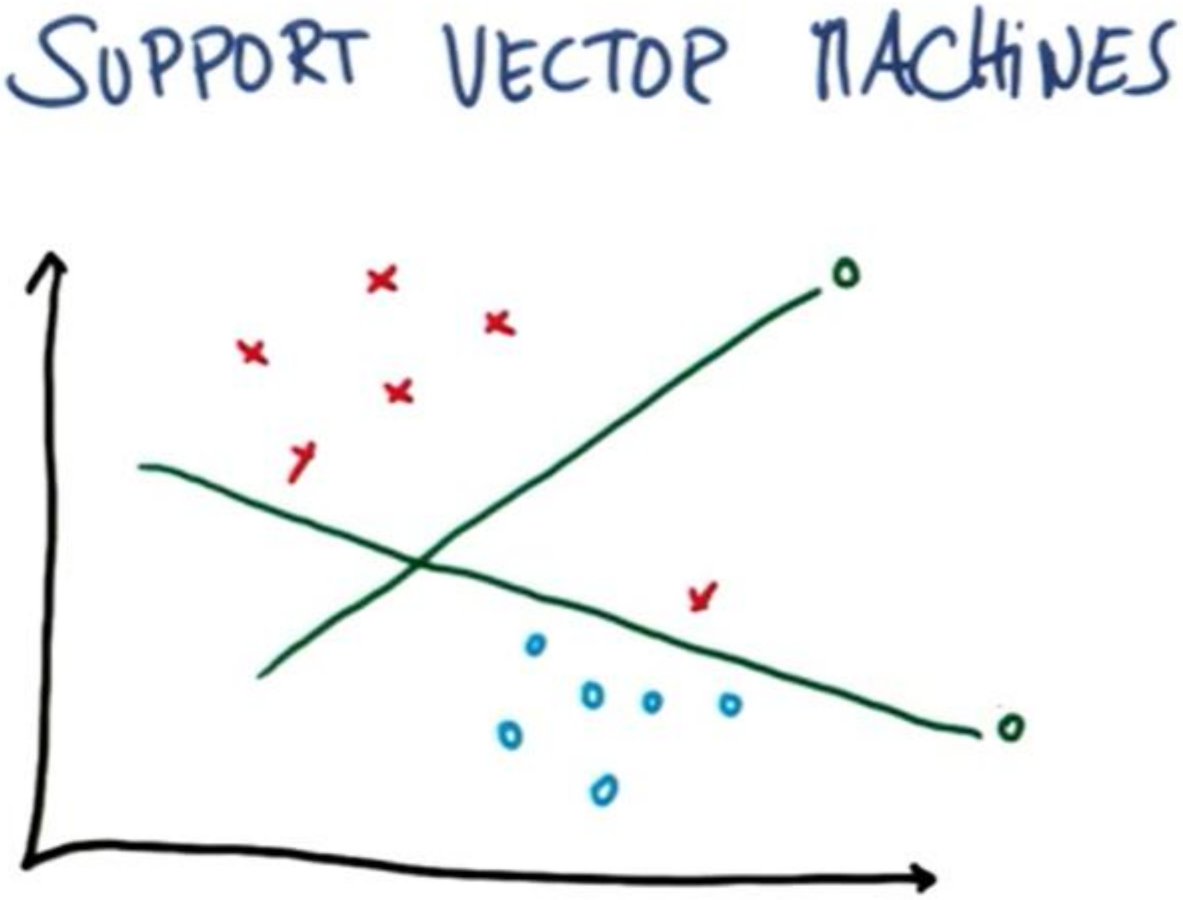
What Makes A Good Separating Line




Practice with Margins
SUPPORT VECTOR MACHINE
MAXIMIZES DISTANCE TO NEAREST POINT
= MARGIN

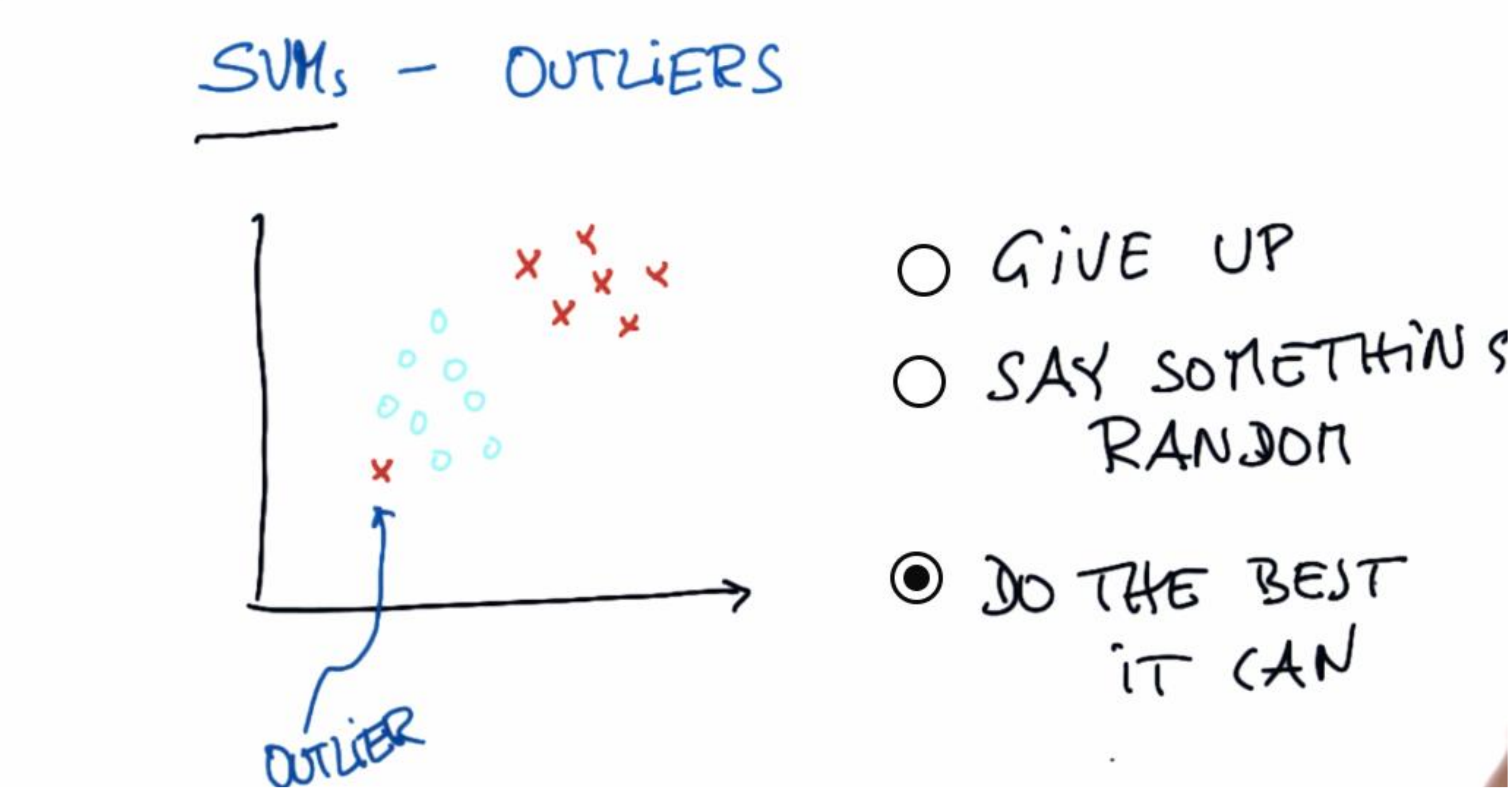
SVM

Which is the best line



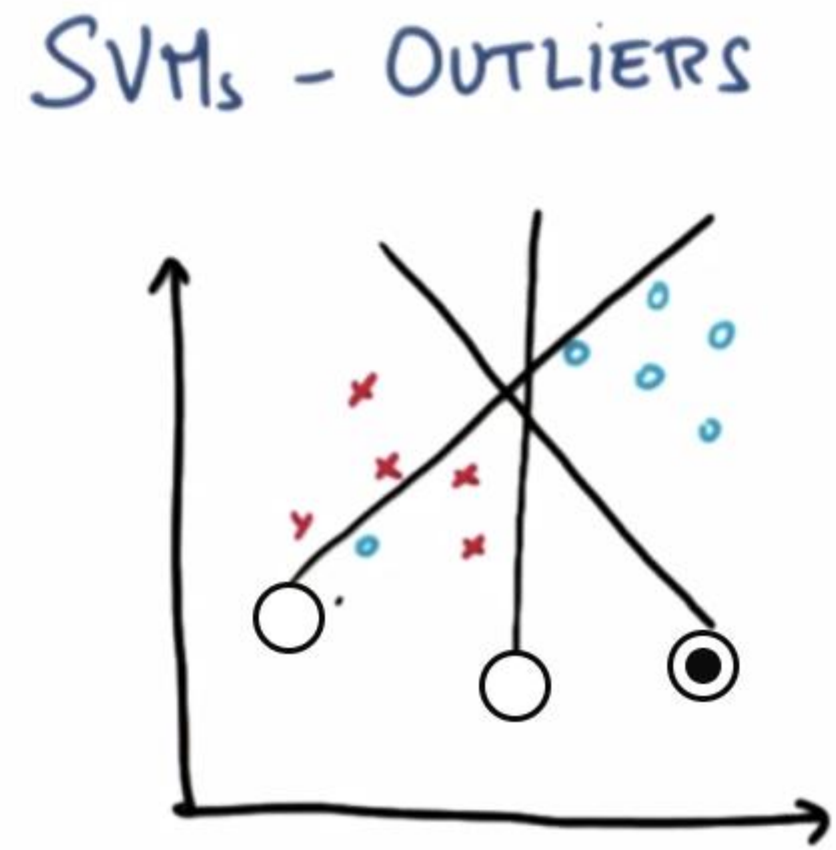
svm puts first and foremost the correct classification
of the labels and then maximize the margin
so for svm you are trying to classify correctly and subject
to that constrain you maximize the margin

SVM Response to Outliers



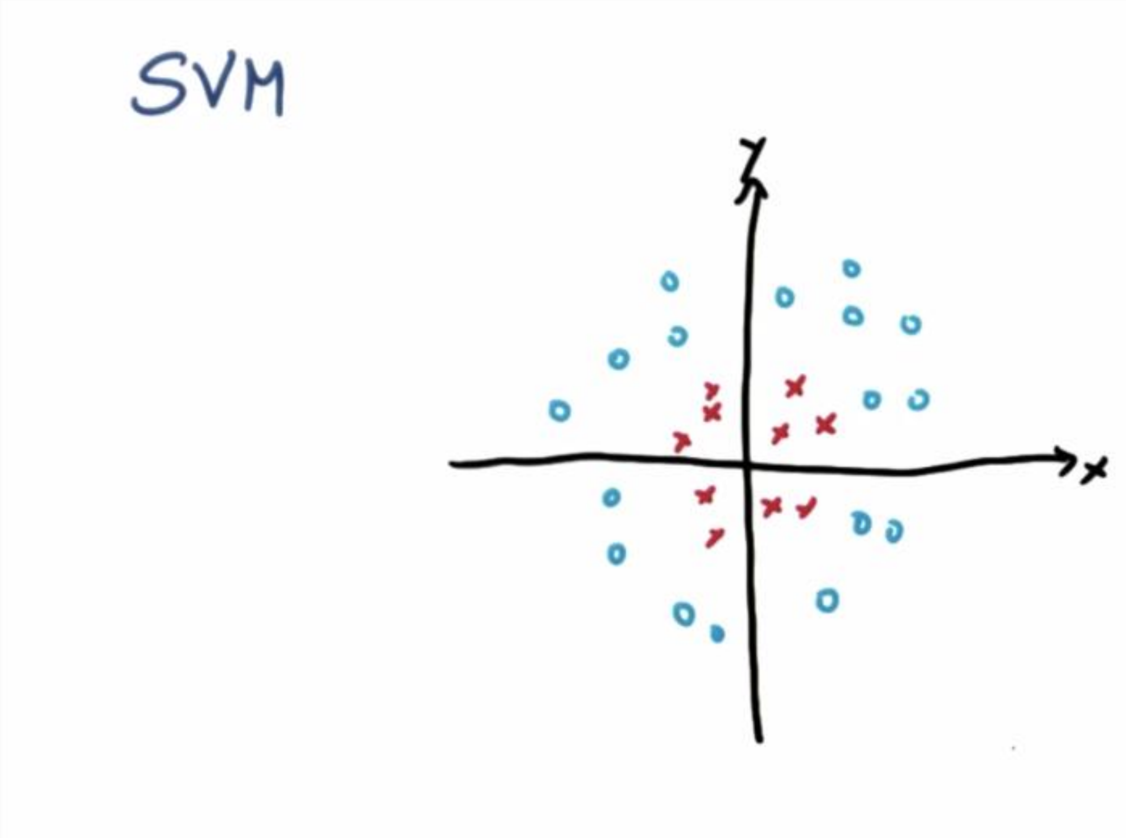
SVM

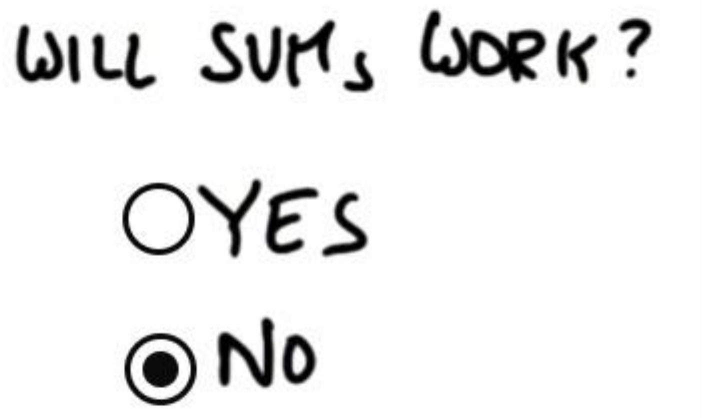
SVM Outlier Practice



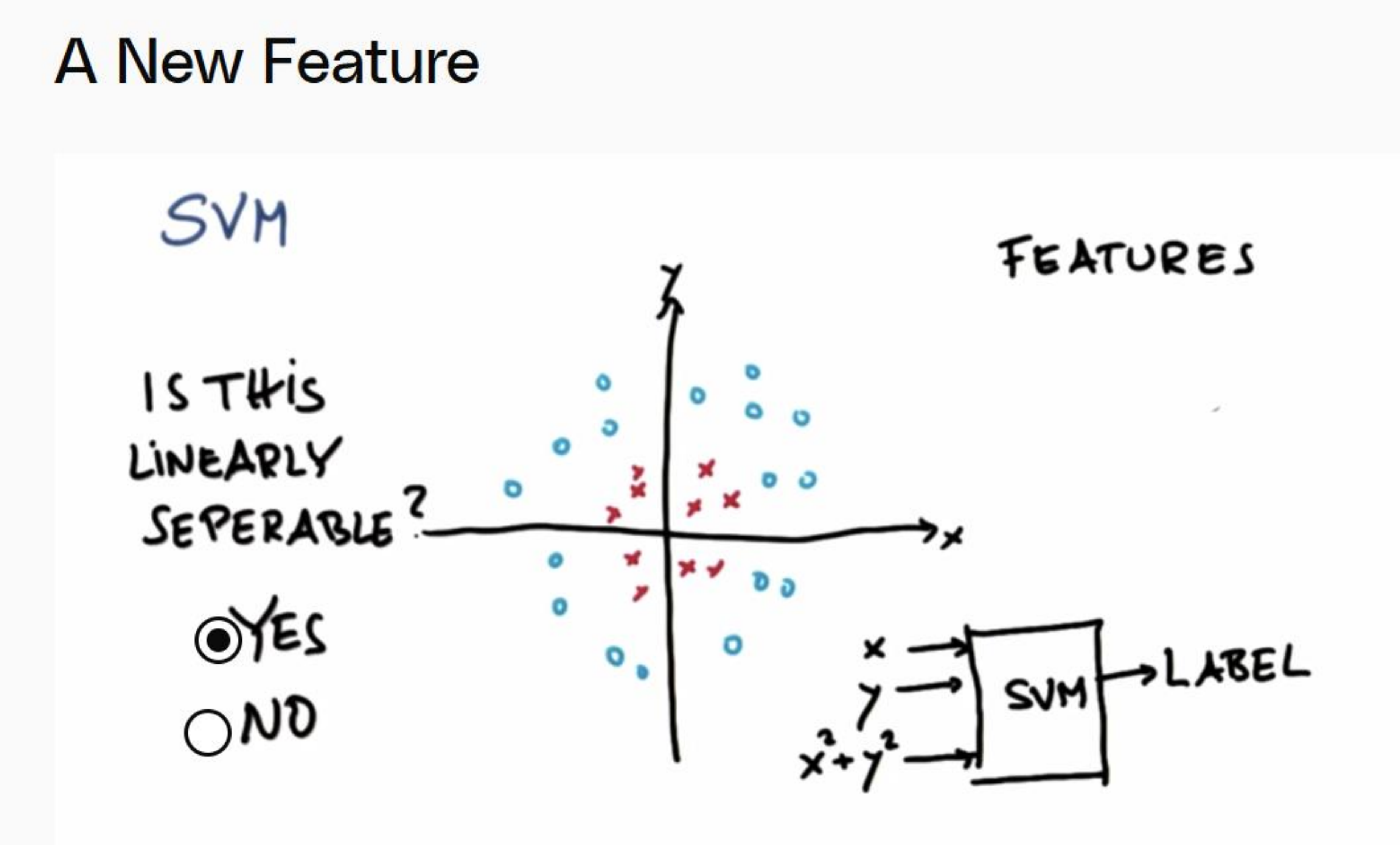
SVM

Nonlinear Data



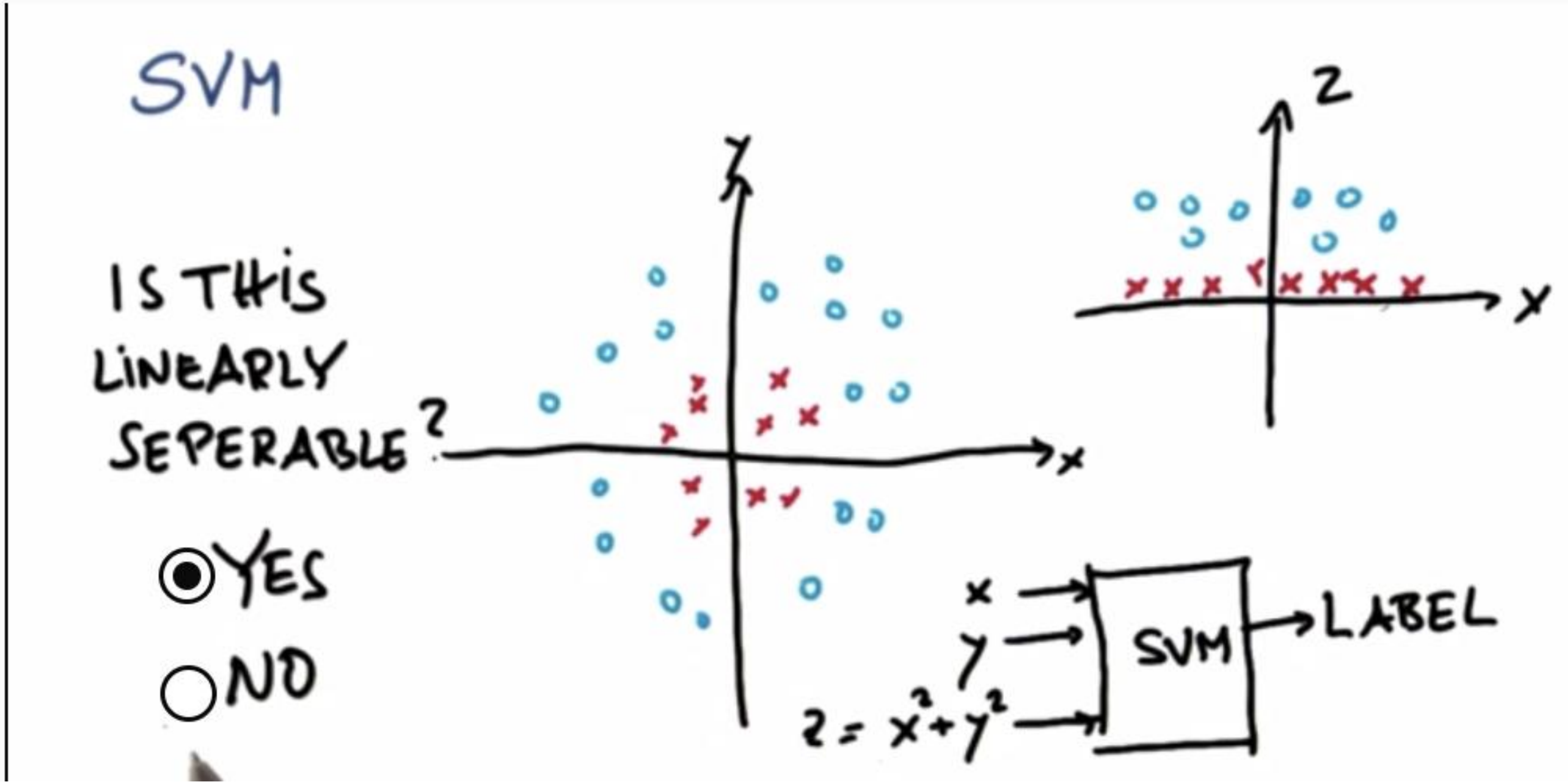


SVM


A New Feature
SVM
IS THIS LINEARLY SEPERABLE?
YES
NO
FEATURES
y
x
x
y
x²+y²
SVM
LABEL

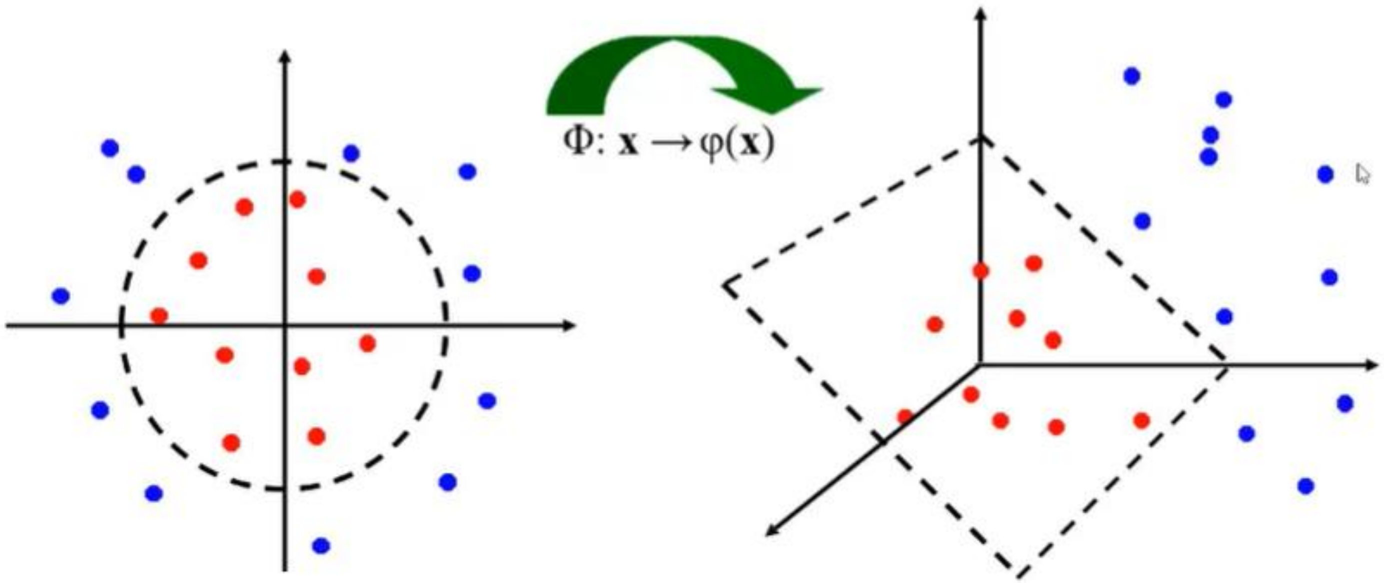
SVM

Separating with the New Feature

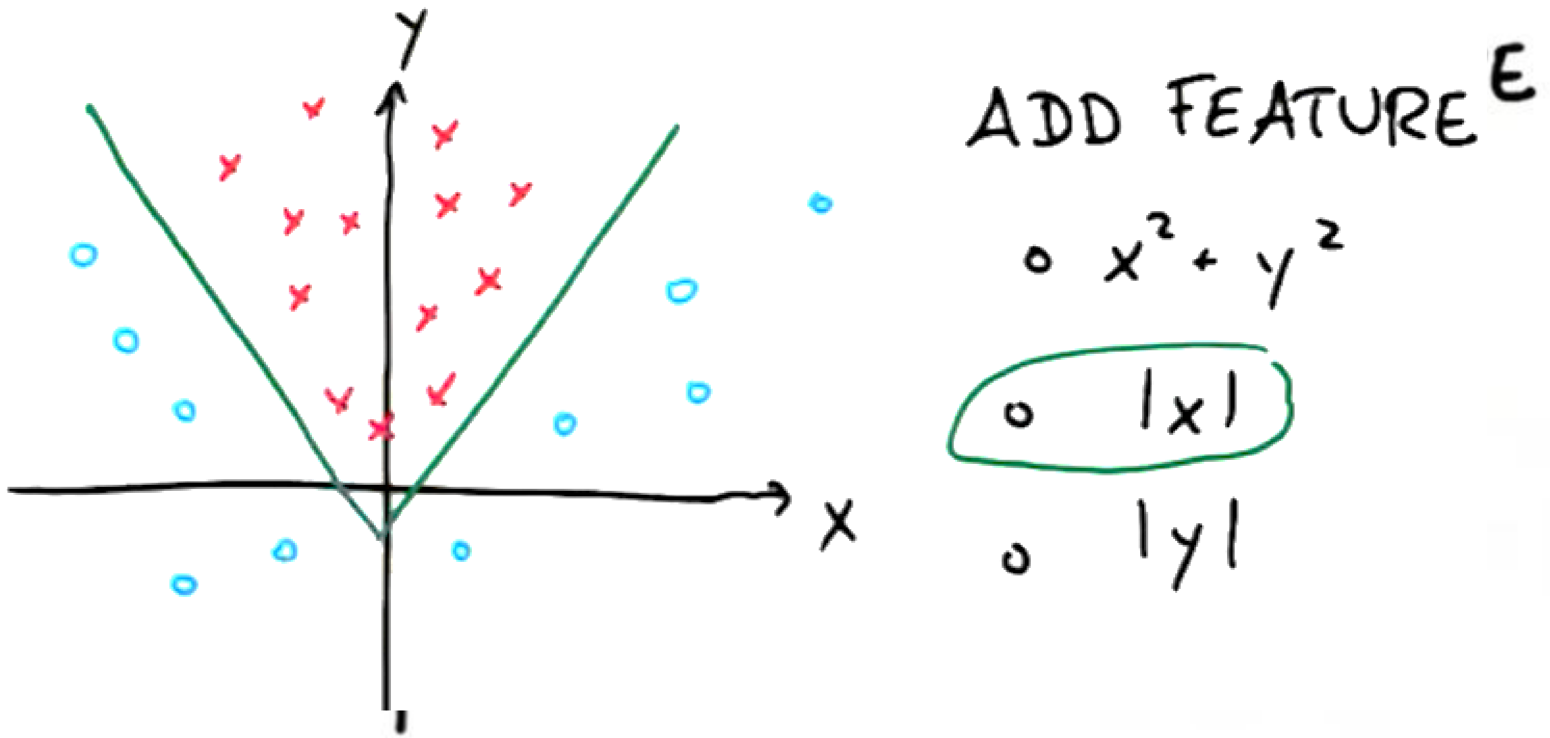


Non-linear SVMs: Feature spaces

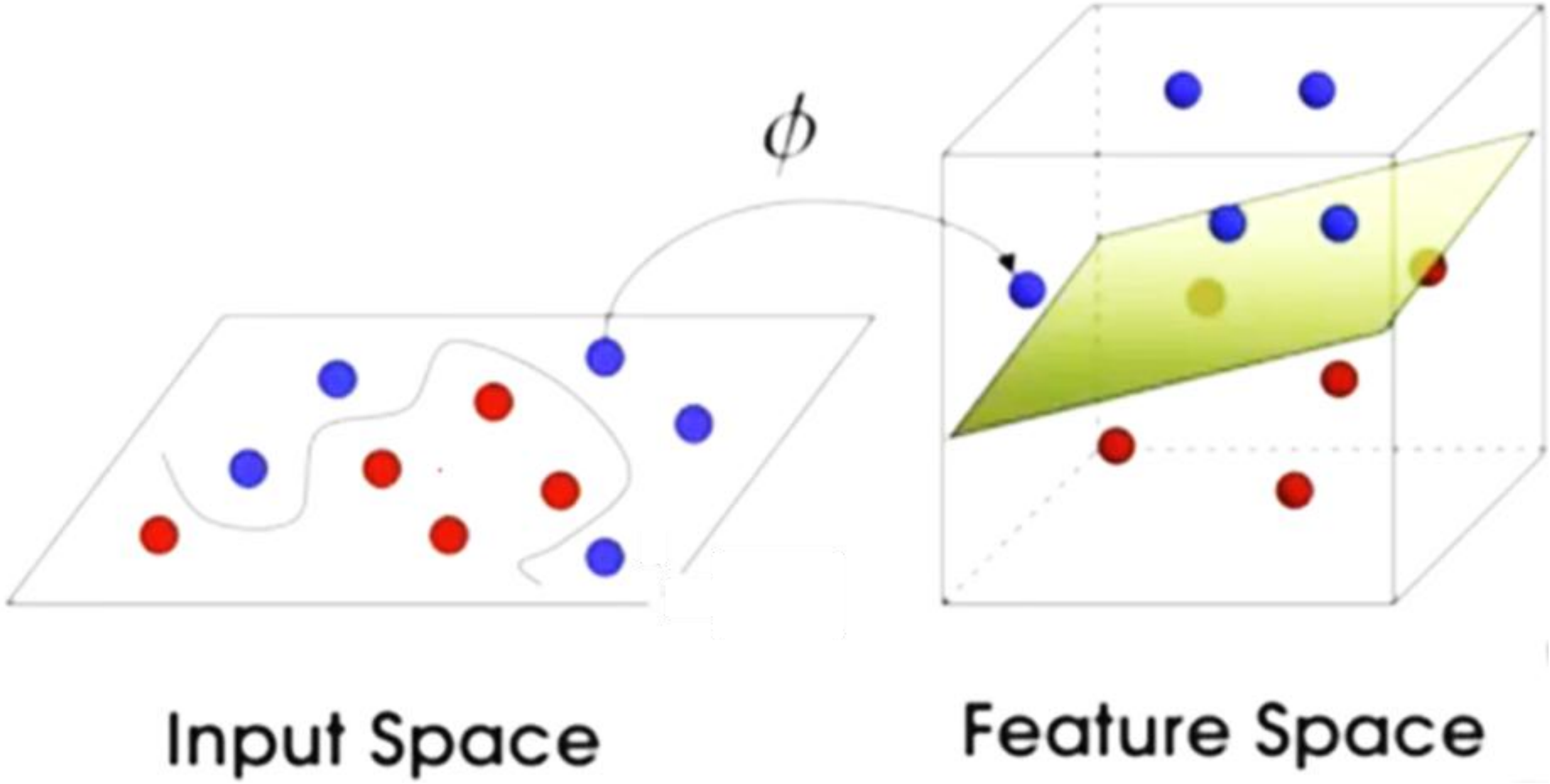
- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



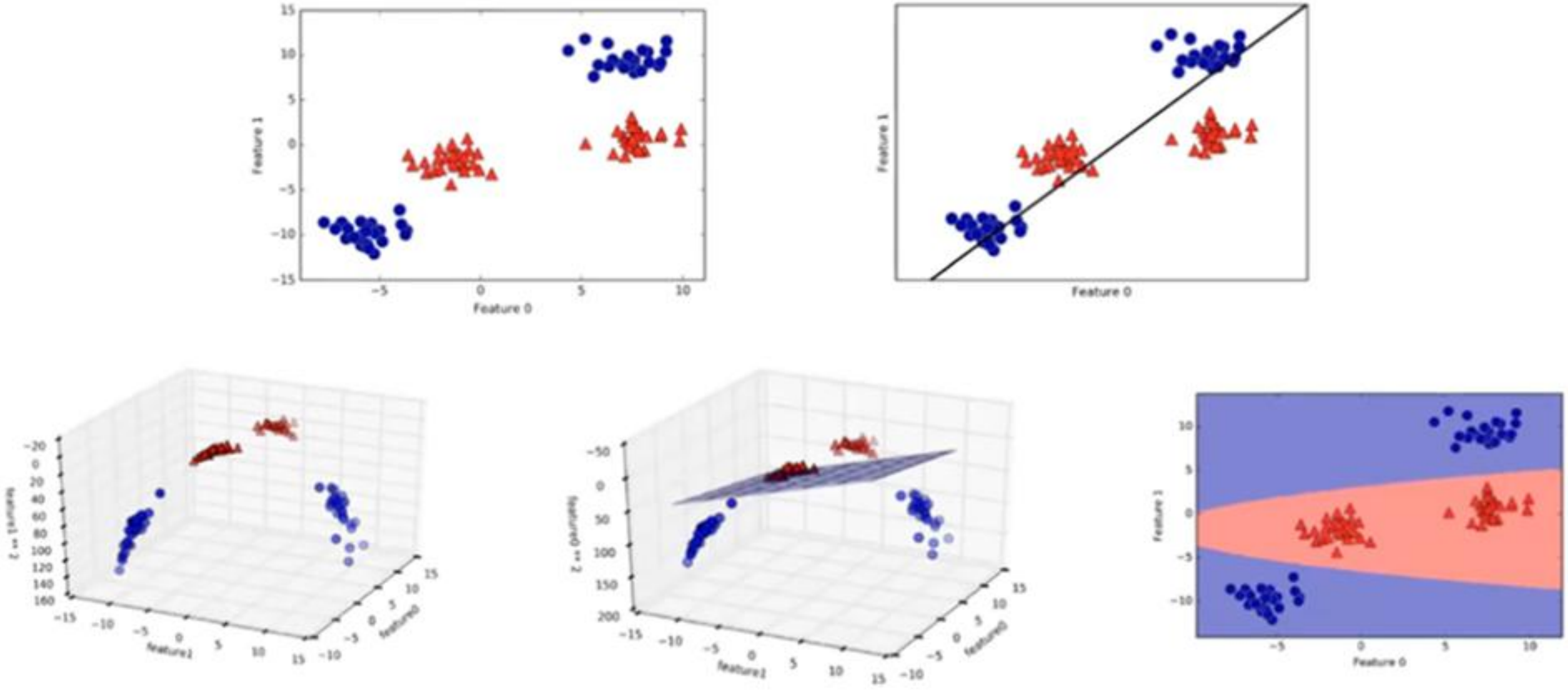
Practice Making a New Feature

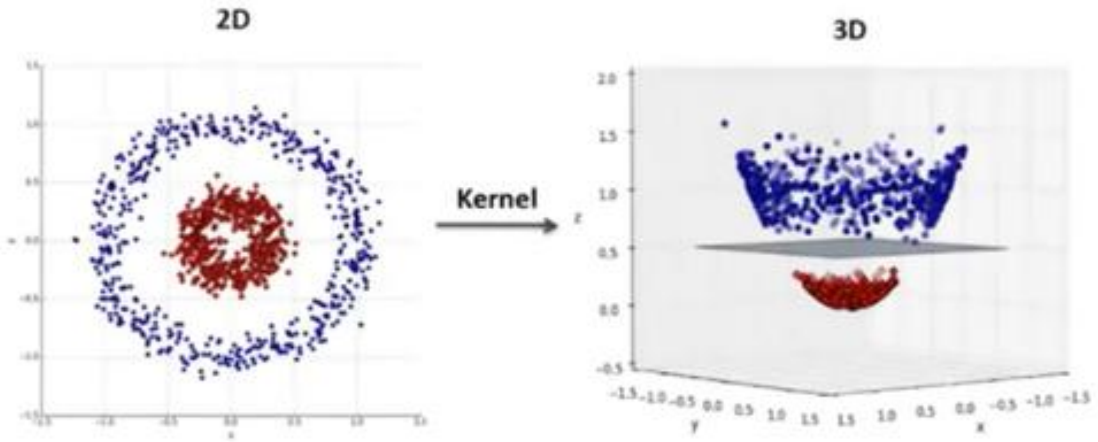


Overfitting


ϕ
Input Space
Feature Space

(Applying Kernel Function)





No	Kernel function	Formula	Optimization parameter
1	Dot -product	$K(x_n,x_i) = (x_n,x_i)$	C
2	RBF	$K(x_n,x_i) =$ $\exp(-\gamma \parallel x_n - x_i \parallel^2 + C$	C and γ
3	Sigmoid	$K(x_n,x_i) =$ $\tanh(\gamma(x_n,x_i) + r)$	C, γ, and r
4	Poly-nomial	$K(x_n,x_i) =$ $(\gamma(x_n,x_i) + r)^d$	C, γ, r, d

Thank You