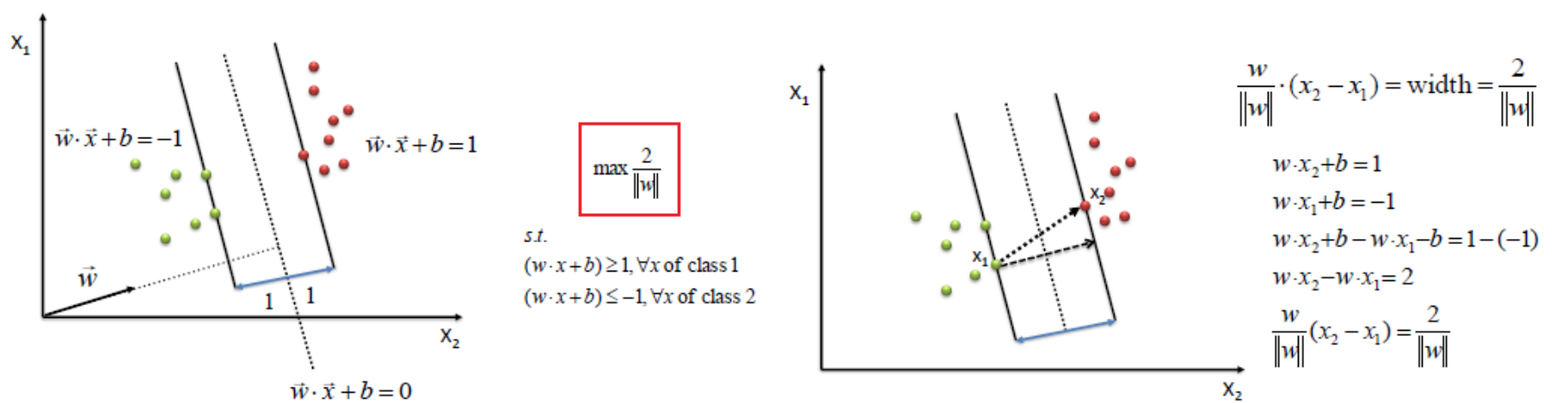# How to find the optimal hyperplane?

A hyperplane is a decision boundary that differentiates the two classes in SVM, and the hyperplane for which the margin is maximum is the optimial hyperplane.

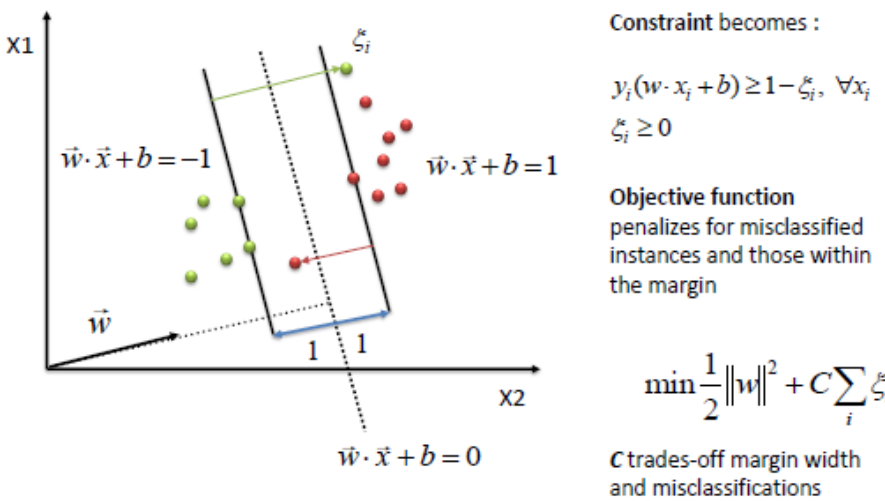To define an optimal hyperplane we need to maximize the width of the margin (w).

We find w and b by solving the following objective function using Quadratic Programming.

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

$$s.t.\ y_i(w \cdot x_i + b) \geq 1,\ \forall x_i$$

In SVM if the data is linearly separable, there is a unique global minimum value. An ideal SVM analysis should produce a hyperplane that completely separates two classes. However, perfect separation may not be possible, or it may result in a model with so many cases that the model does not classify correctly. In this situation SVM finds the hyperplane that maximizes the margin and minimizes the misclassifications.

slack variable:

ξ_i

Allow some instances to fall off the margin, but penalize them
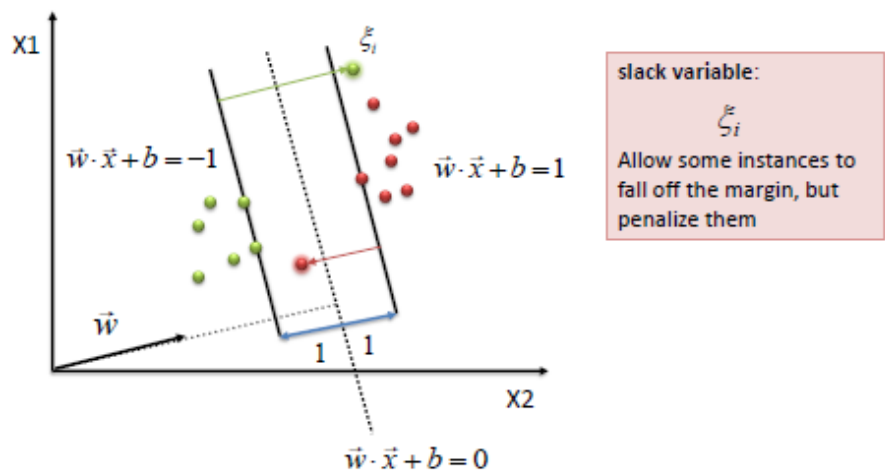
SVM tries to maintain the **slack variable** to zero while maximizing margin. However, it does not minimize the number of misclassifications (NP-complete problem) but the sum of distances from the margin hyperplanes.

Constraint becomes :

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i,\ \forall x_i$$
$$\xi_i \geq 0$$

Objective function penalizes for misclassified instances and those within the margin

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

C trades-off margin width and misclassifications

There are situations when it is required to separate two classes in nonlinear region. SVM handles this by using a kernel function (for nonlinearity) to map the data into a different space where a linear functionality can not be used to do the separation. Kernel function is used to separate the data into a higher dimensional feature space to make it possible to perform the linear separation.

Linear SVM $\quad x_i \cdot x_j$

Non-linear SVM $\quad \phi(x_i) \cdot \phi(x_j)$

Kernel function $\quad k(x_i \cdot x_j)$
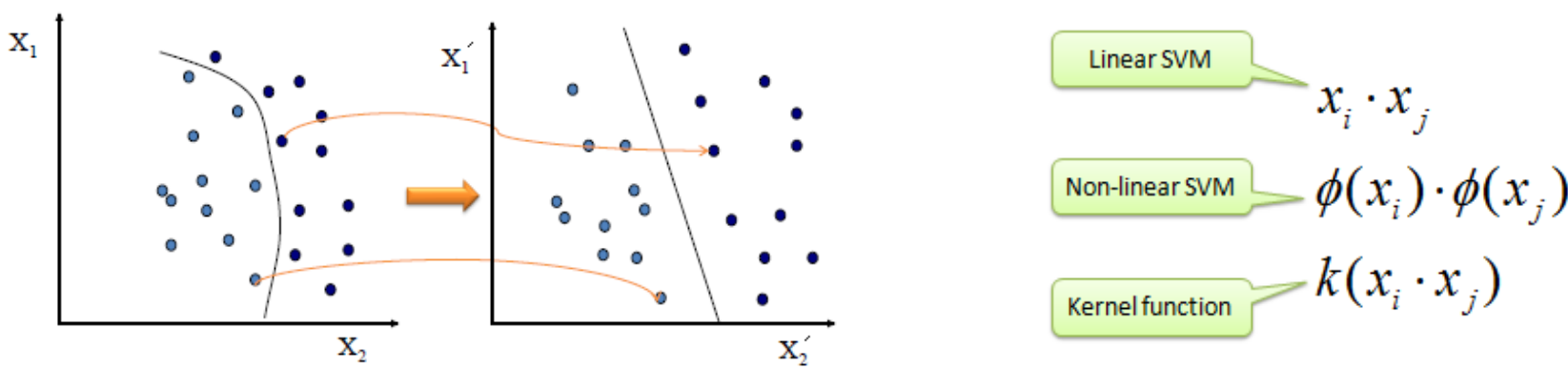
Kernel functions map data into new space, then take the inner product of the new vectors. The image of the inner product of the data is the inner product of the images of the data. Two kernel functions are shown below:

Polynomial

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$