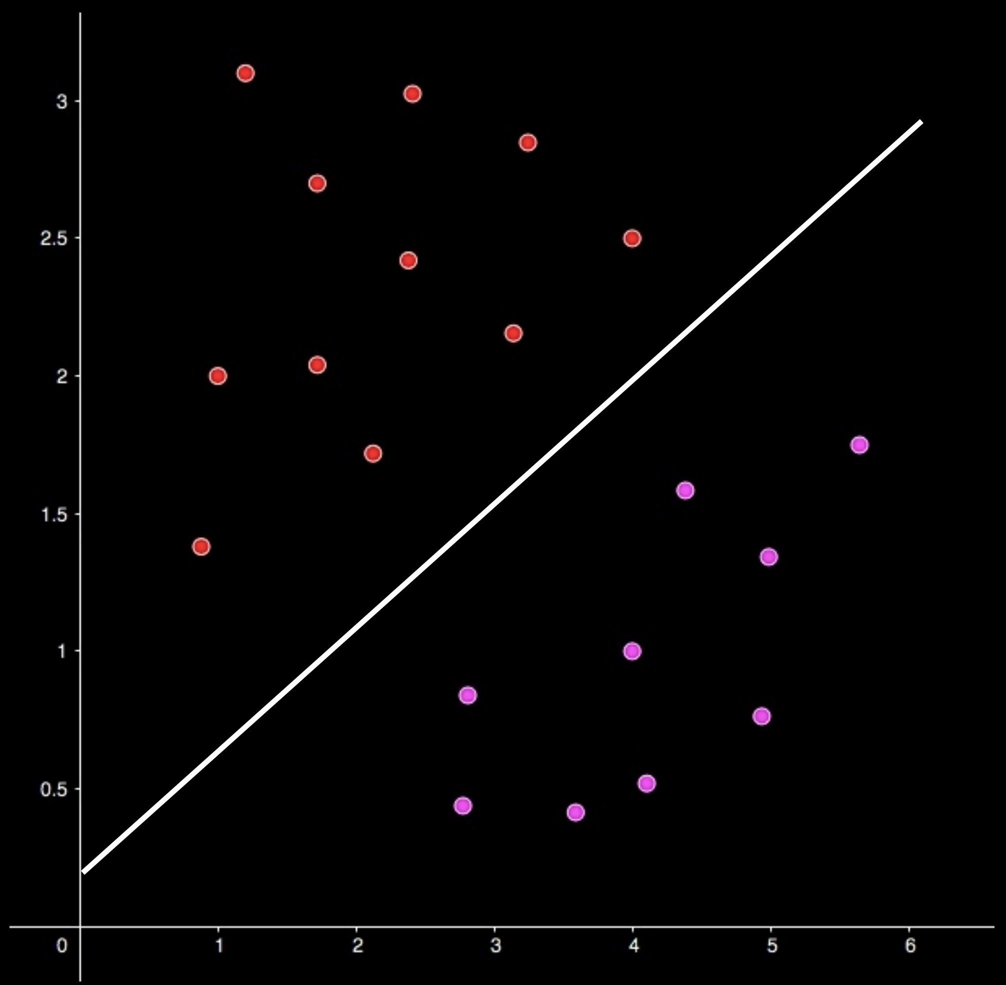
ML@LSE - Bootcamp 3: Support Vector classification methods

Ref. for today: Chapter 9, Hastie et al.

Introduction

Last time we studied a supervised learning method that could be applied to regression and classification.
Today we're going to talk about another supervised learning method used for classification: Support Vector Classification.

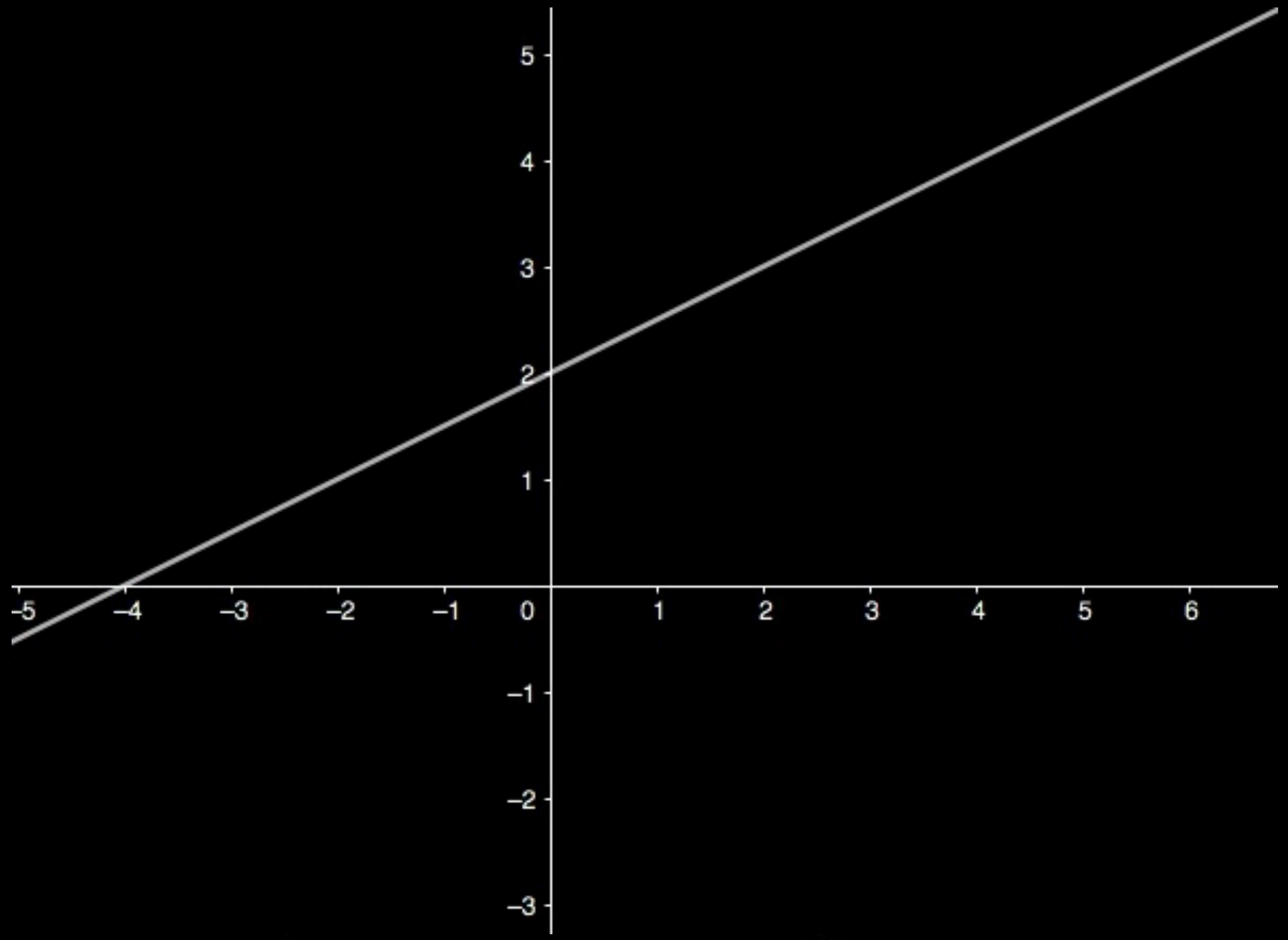
What is this about?



What is a hyperplane?

Abstract definition: A flat affine subspace that is one dimension less than its ambiant space.

What does that mean??



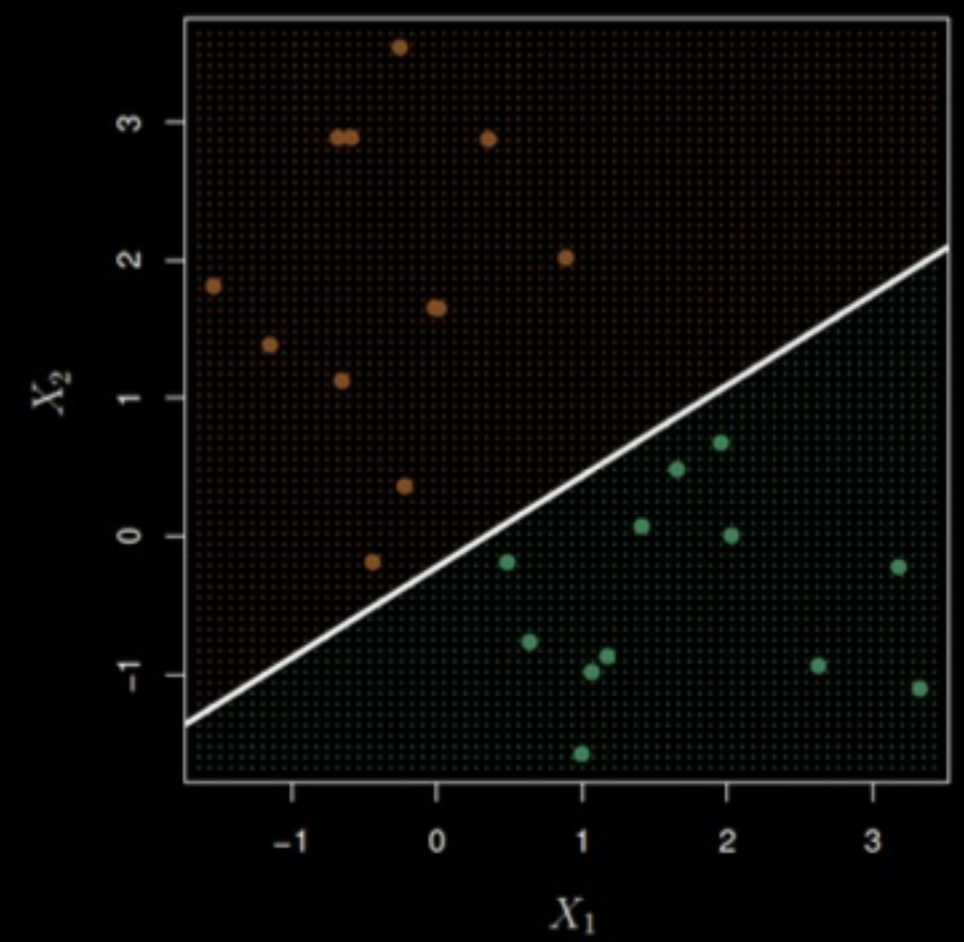
Hyperplanes and maximal margin

Mathematically, a hyperplane in p-dimension can be described by the equation:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0$$



Let Y_i=1 if brown and Y_i=-1 if green.

Let $f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

Then given an observation $\mathbf{x}_0$ we can classify $\mathbf{x}_0$ according to the sign of $f(\mathbf{x}_0)$.

Maximum Margin Hyperplane

How can we find this separating hyperplane?

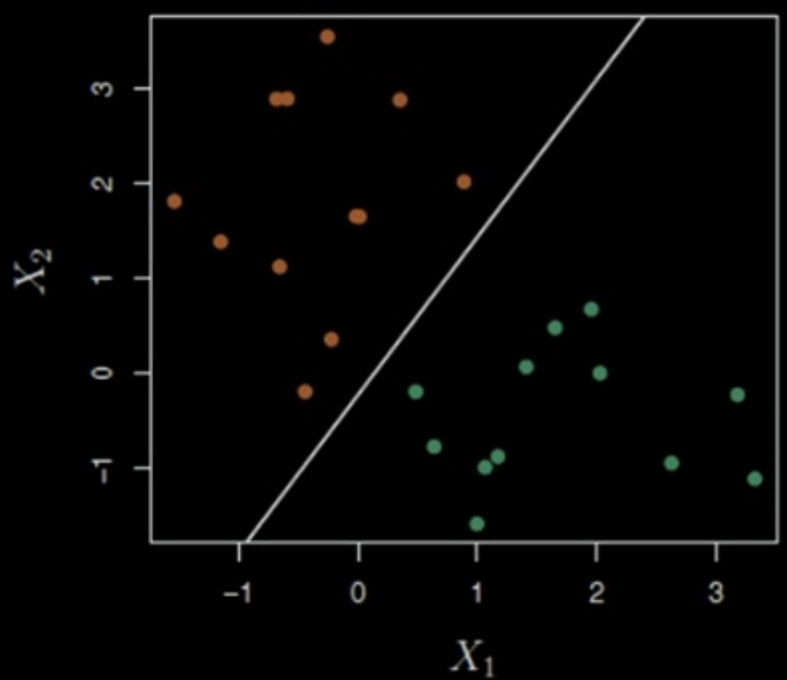
Let's try to find one manually

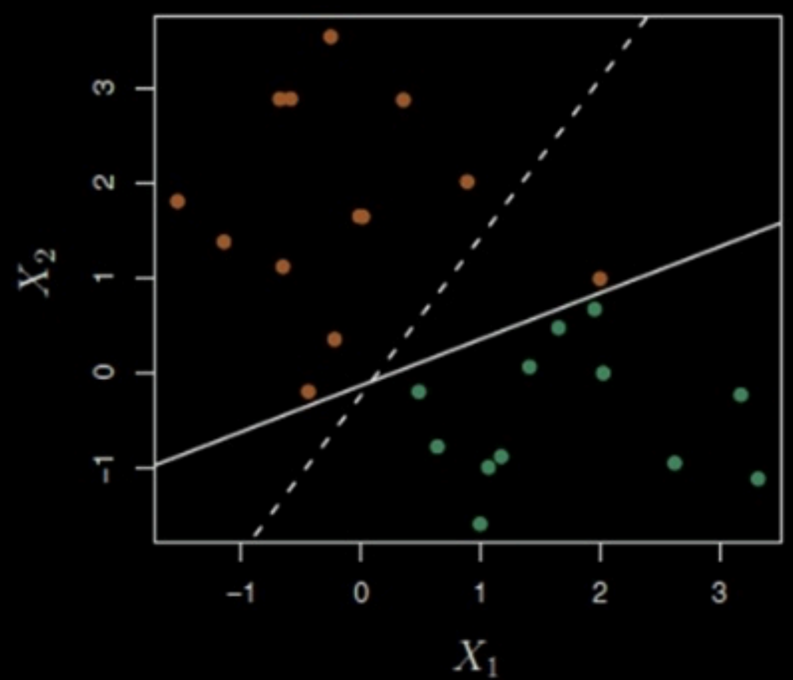
Maximum Margin Hyperplane

How can we find this separating hyperplane?

Let's try to find one manually

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \quad M \quad s.t. \quad \sum_{j=1}^{p}\beta_j^2 = 1, \;\; y_i(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}) \geq M$$

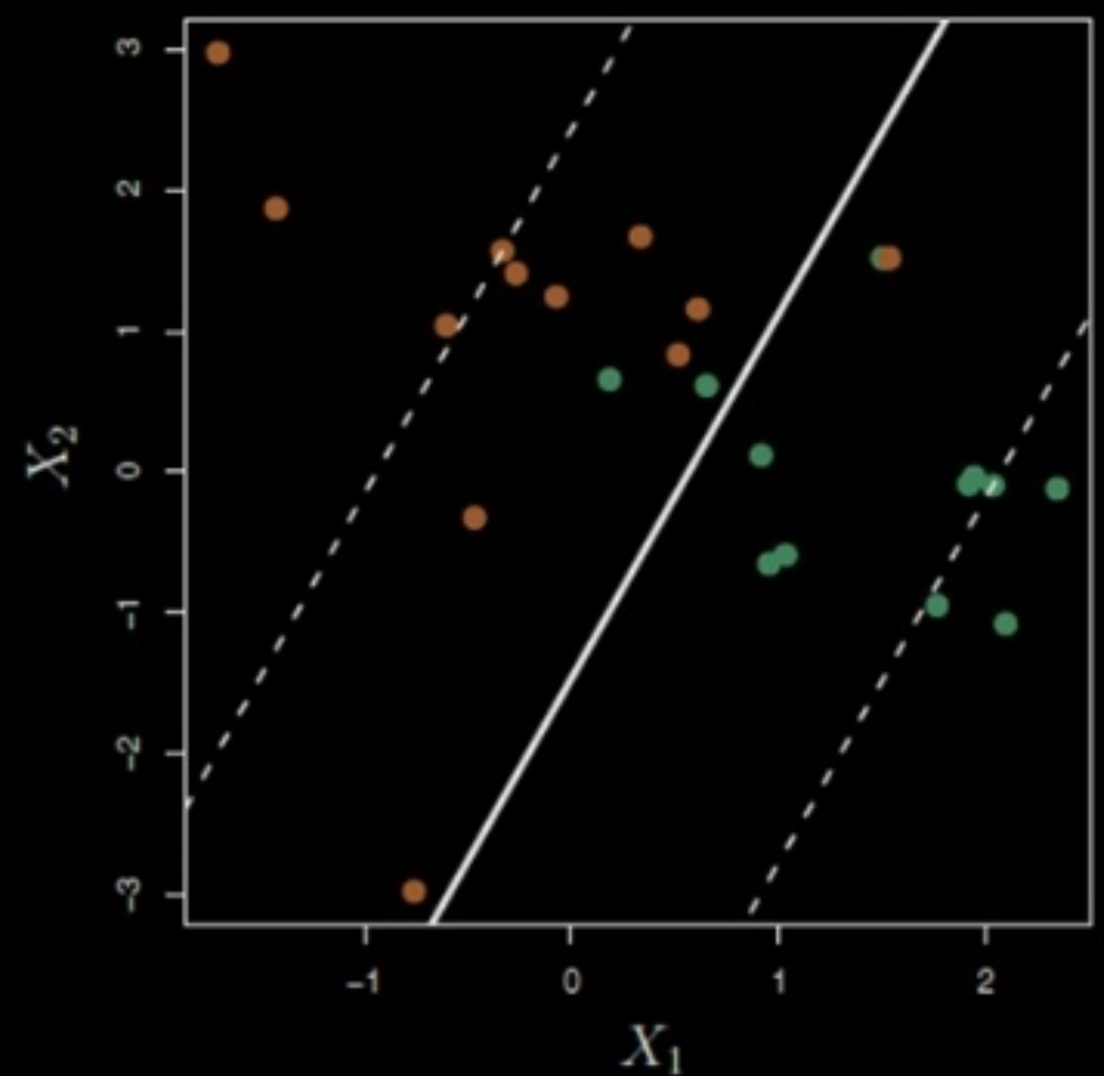




Soft Margin Classifier

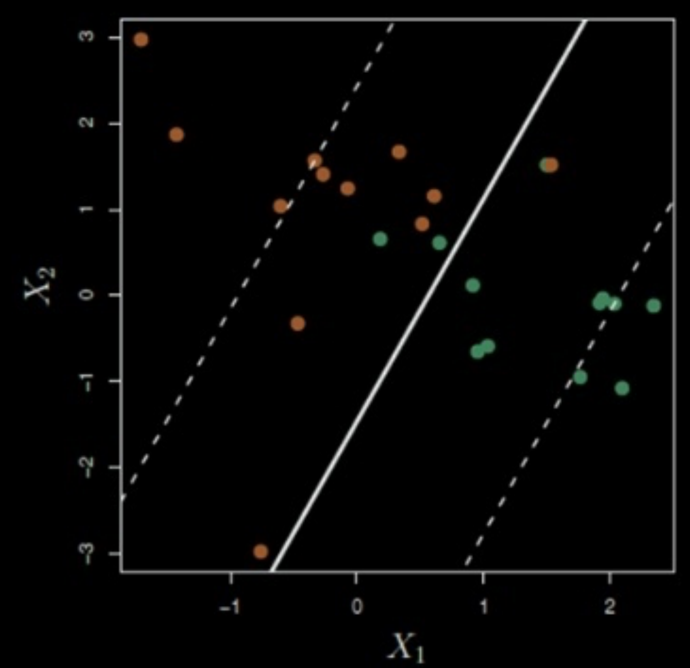
A more robust approach: Soft Margin Classifier (or support vector classifier):

We allow for *some* observations to be within the margin, or even to be on the other side of the hyperplane.



Soft Margin Classifier: Optimization problem

The optimization problem:

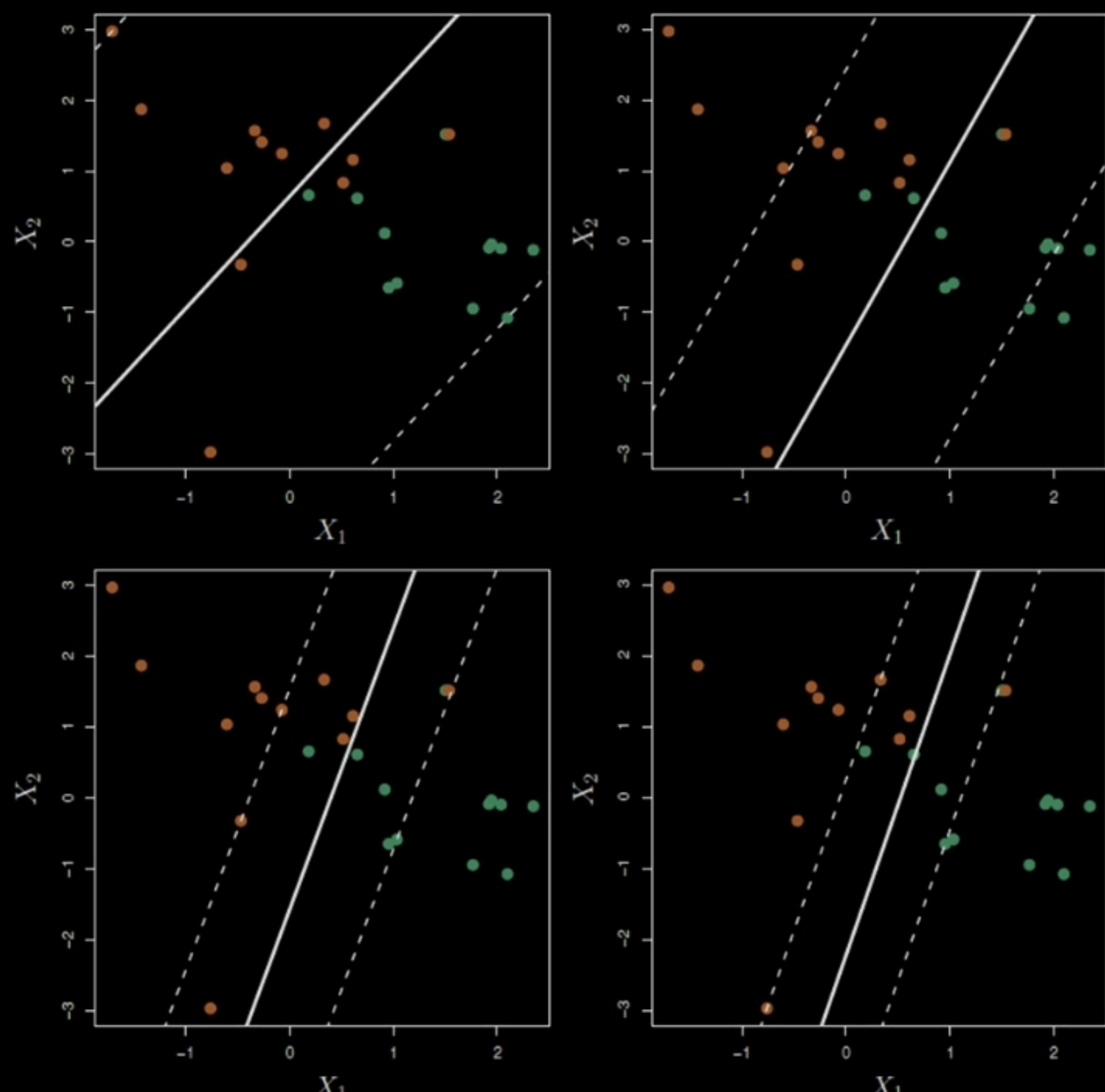


$$\max_{\beta_0,\beta_1,\ldots,\beta_p,s_1,\ldots s_n} M \quad s.t. \begin{cases} \sum_{j=1}^{p} \beta_j^2 = 1 \\ y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \geq M(1 - s_i)) \\ s_i \geq 0, \sum_{i=1}^{p} s_i \leqslant C \end{cases}$$

- if $s_i = 0$, the i-th observation is on the correct side of the margin.
- if $0 < s_i < 1$, the i-th observation is on the wrong side of the margin but on the right side of the hyperplane.
- if $s_i < 0$, the i-th observation is a the wrong side of the hyperplane.

Soft Margin Classifier: Optimization problem

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,s_1,\ldots s_n} M \quad s.t. \quad \begin{cases} \sum_{j=1}^{p} \beta_j^2 = 1 \\ y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \geq M(1 - s_i)) \\ s_i \geq 0, \sum_{i=1}^{p} s_i \leqslant C \end{cases}$$



C is a regularization parameter

Non-linear Decision Boundaries

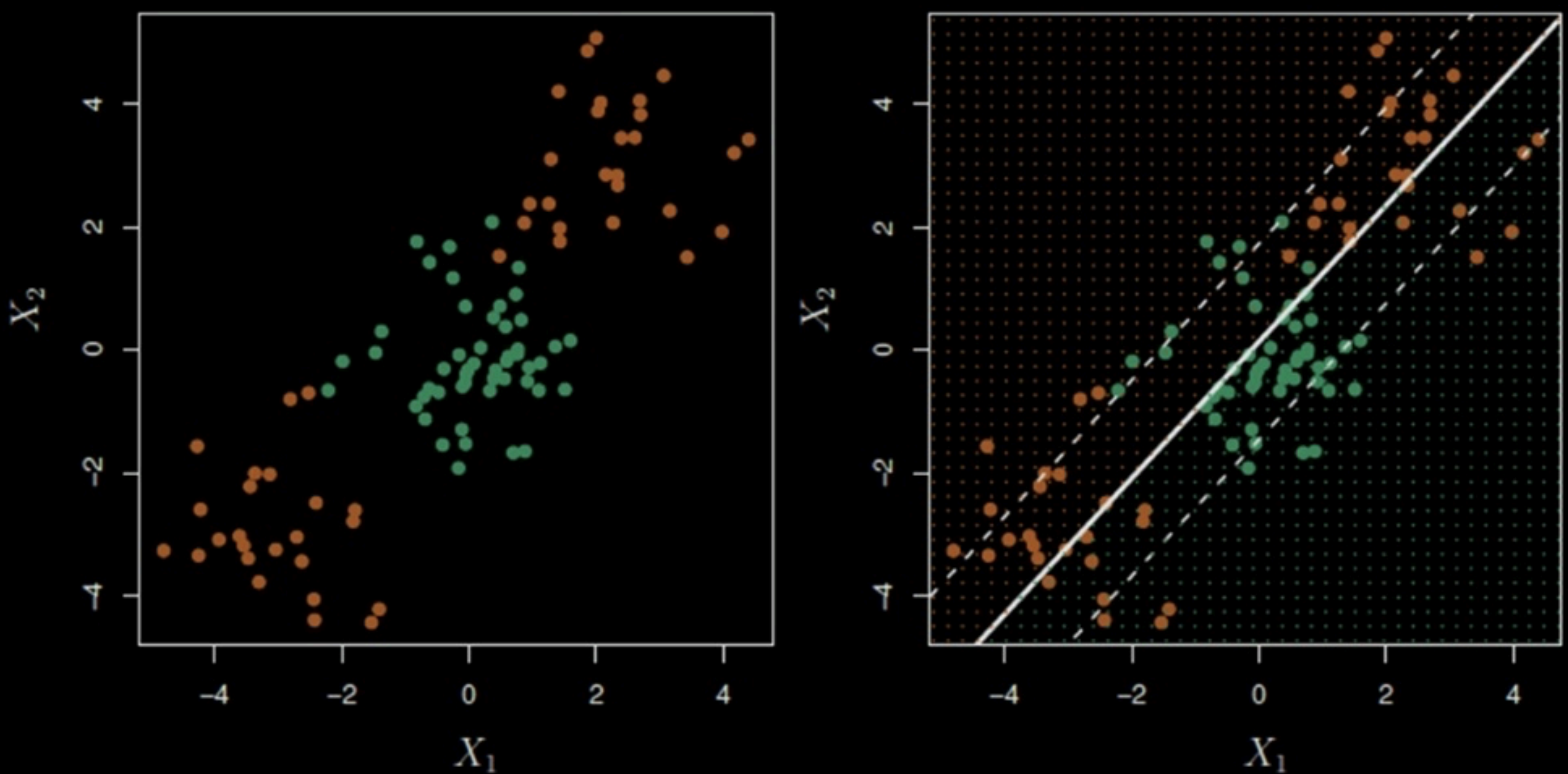
A- Linear Classification

1. **Maximal Margin Classifier**
   a. Hyperplanes
   b. Maximal Margin Hyperplane
2. **Soft Margin Classifier**
   a. The Optimization Problem of SMC
   b. The Bias-Variance tradeoff

B- Support Vector Machines

1. **Non-linear Decision Boundaries**
   a. **What if we enlarged the Feature Space?**
   b. The New Optimization Problem
2. **Kernel Tricks**
   a. Inner Product and the Kernel Idea
   b. Support Vector Machines

What can we do when the data is not linearly separable?



Enlarge the feature space by adding powers of the features:

$$X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2$$

Non-linear Decision Boundaries

A- Linear Classification

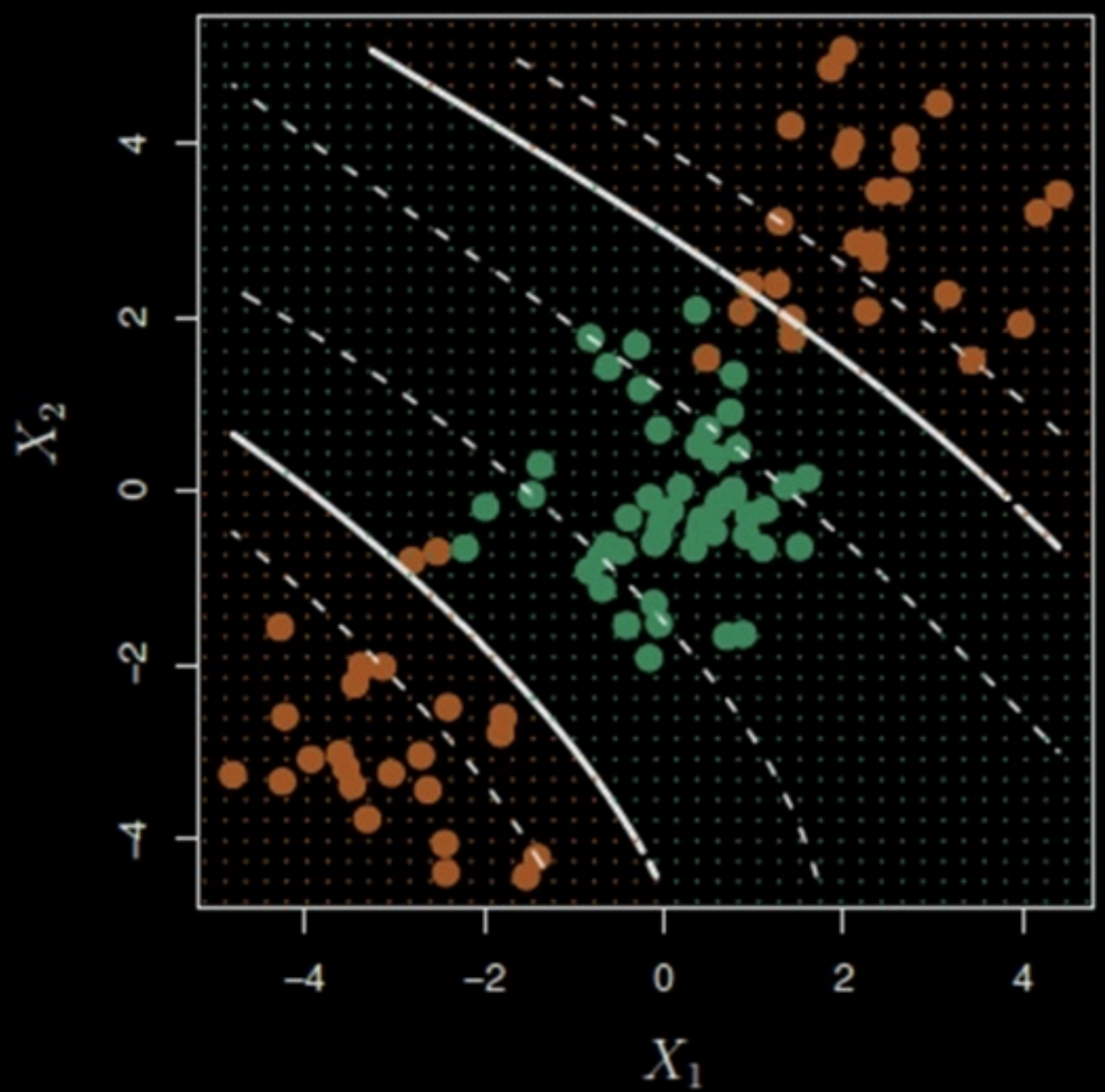
1. **Maximal Margin Classifier**
   a. Hyperplanes
   b. Maximal Margin Hyperplane
2. **Soft Margin Classifier**
   a. The Optimization Problem of SMC
   b. The Bias-Variance tradeoff

B- Support Vector Machines

1. **Non-linear Decision Boundaries**
   a. **What if we enlarged the Feature Space?**
   b. The New Optimization Problem
2. **Kernel Tricks**
   a. Inner Product and the Kernel Idea
   b. Support Vector Machines

Enlarge the feature space by adding powers of the features:

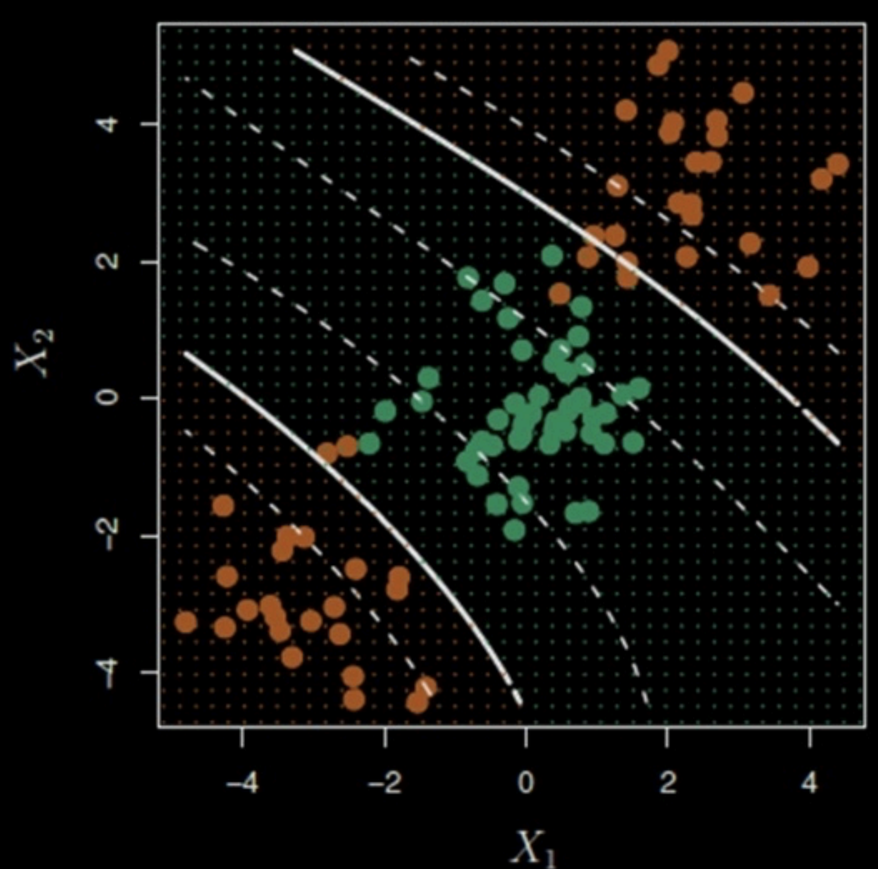
$X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2$



Optimization problem with polynomials

Here is the new optimization problem:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,s_1,\ldots s_n} M \quad s.t. \begin{cases} \sum_{j=1}^{p}\beta_j^2 = 1 \\ y_i(\beta_0 + \beta_{11}x_{i1} + \ldots + \beta_{p1}x_{ip} + \beta_{12}x_{i1}^2 + \beta_{p1}x_{ip}^2) \geq M(1-s_i)) \\ s_i \geq 0, \sum_{i=1}^{p} s_i \leqslant C \end{cases}$$



Kernels

! *More advanced*

We can represent the Support Vector Classification function with the classical inner product (dot product):

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \alpha_i < \mathbf{x}, \mathbf{x}_i >$$

We can generalize this by replacing the dot product above by a Kernel. A Kernel is the dot product between the imagine of two vectors $\mathbf{x}$ and $\mathbf{y}$ by φ, where $\varphi : \mathbb{R}^n \to \mathbb{R}^m$. $Ker(\mathbf{x}, \mathbf{y}) = < \varphi(\mathbf{x}), \varphi(\mathbf{y}) >$.

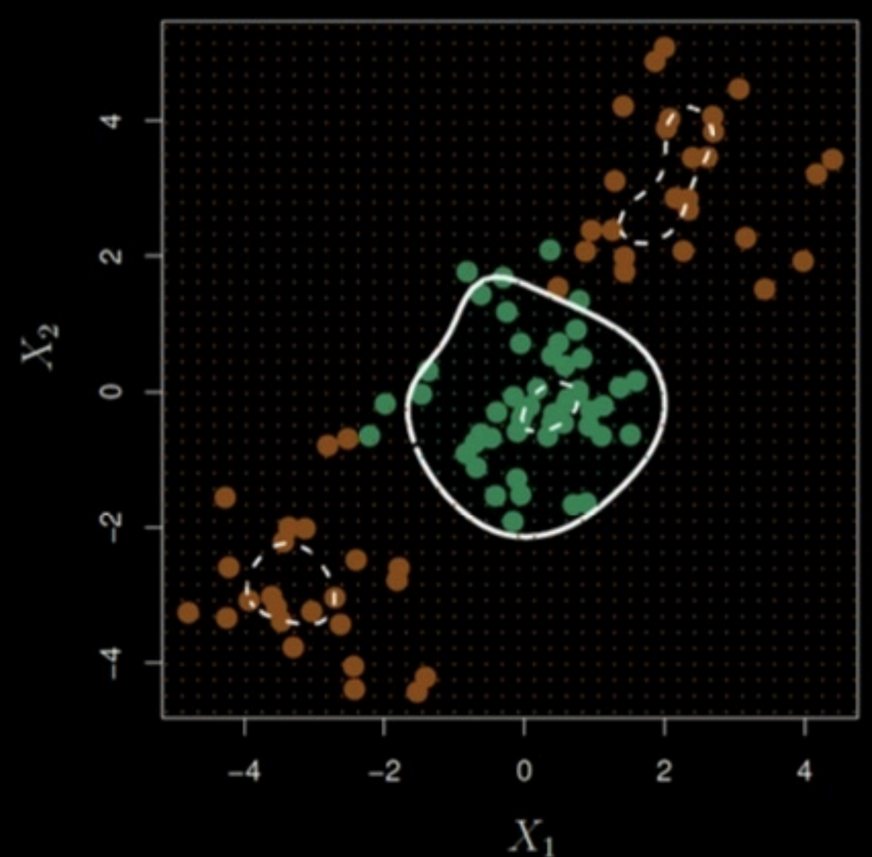
$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \alpha_i Ker(\mathbf{x}, \mathbf{x}_i).$$

Kernels

! *More advanced*

We can generalize this by replacing the dot product above by a Kernel. A Kernel is the dot product between the imagine of two vectors $\mathbf{x}$ and $\mathbf{y}$ by φ, where $\varphi : \mathbb{R}^n \to \mathbb{R}^m$. $Ker(\mathbf{x}, \mathbf{y}) = <\varphi(\mathbf{x}), \varphi(\mathbf{y})>$.

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \alpha_i Ker(\mathbf{x}, \mathbf{x}_i).$$



Example of Kernel: Radial Kernel

$$Ker(\mathbf{x}, \mathbf{x}_i) = exp\Big\{ -\gamma \sum_{j=1}^{p} (x_j - x_{ij})^2 \Big\}$$

Thanks for coming!