Support Vector Machine

Tianhang Zheng

https://tianzheng4.github.io

Intuition

Try and find a plane that separates the classes in feature space.

If we cannot:

Not perfect separation?

Enrich and enlarge the feature space?

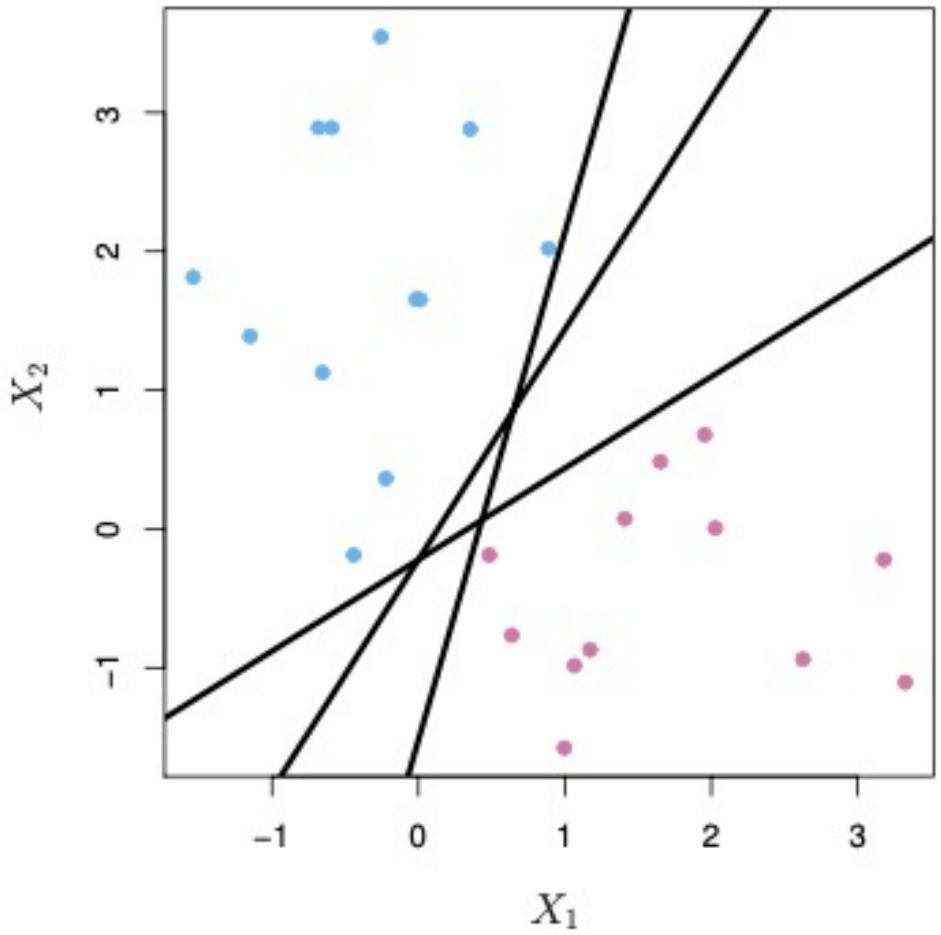
How to separate a feature space

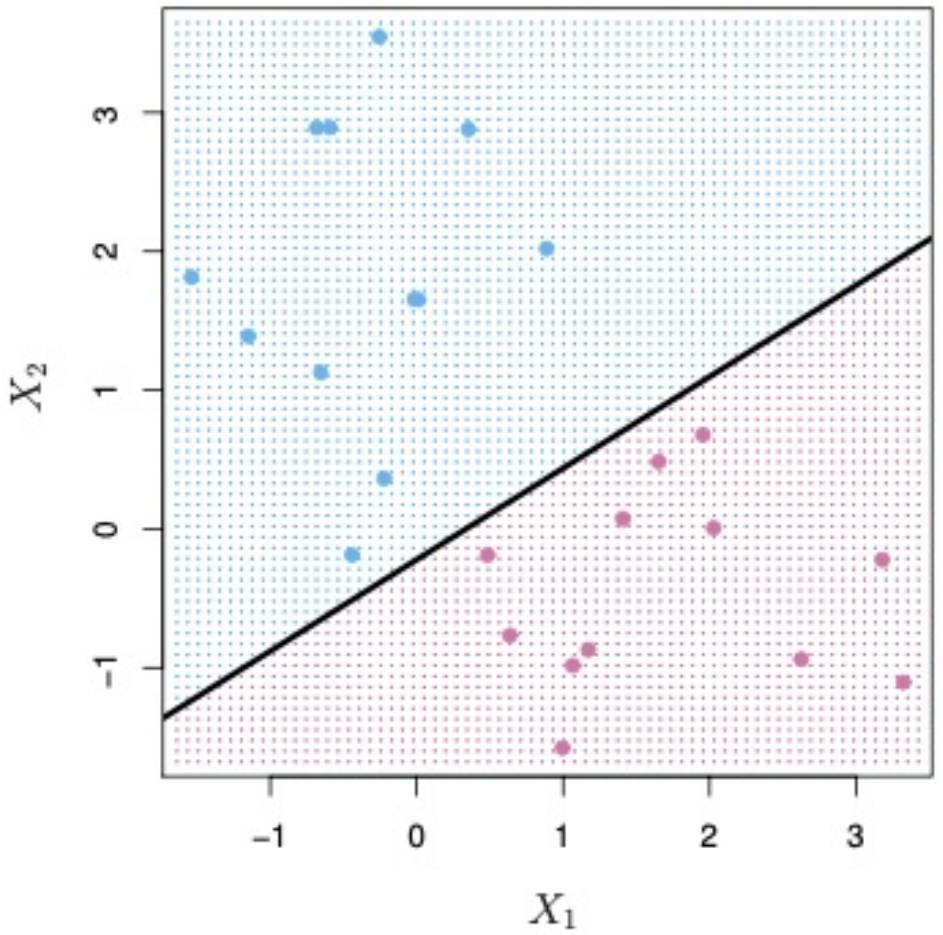
A hyperplane in p dimensions is a flat affine subspace of dimension $p - 1$.

In general the equation for a hyperplane has the form

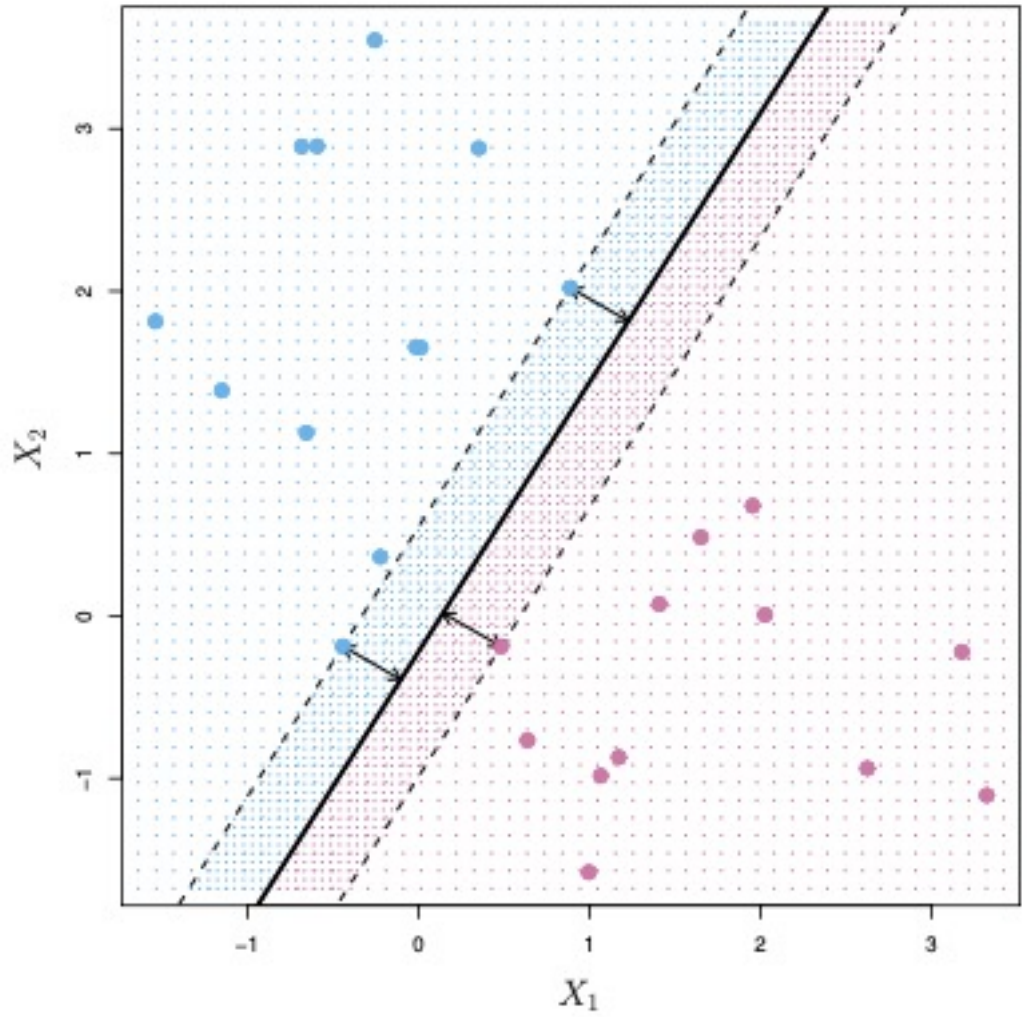
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

Separation by A Hyperplane





Best Hyperplane: Maximal Margin Classifier



Constrained optimization problem

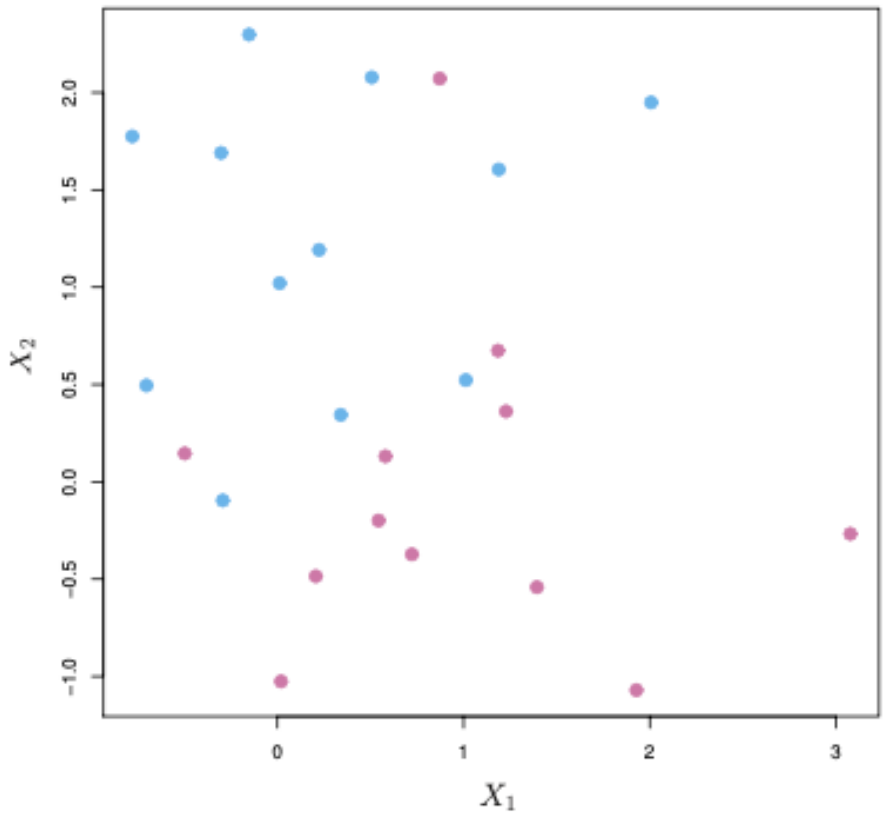
$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M$$

$$\text{for all } \; i = 1, \ldots, N.$$

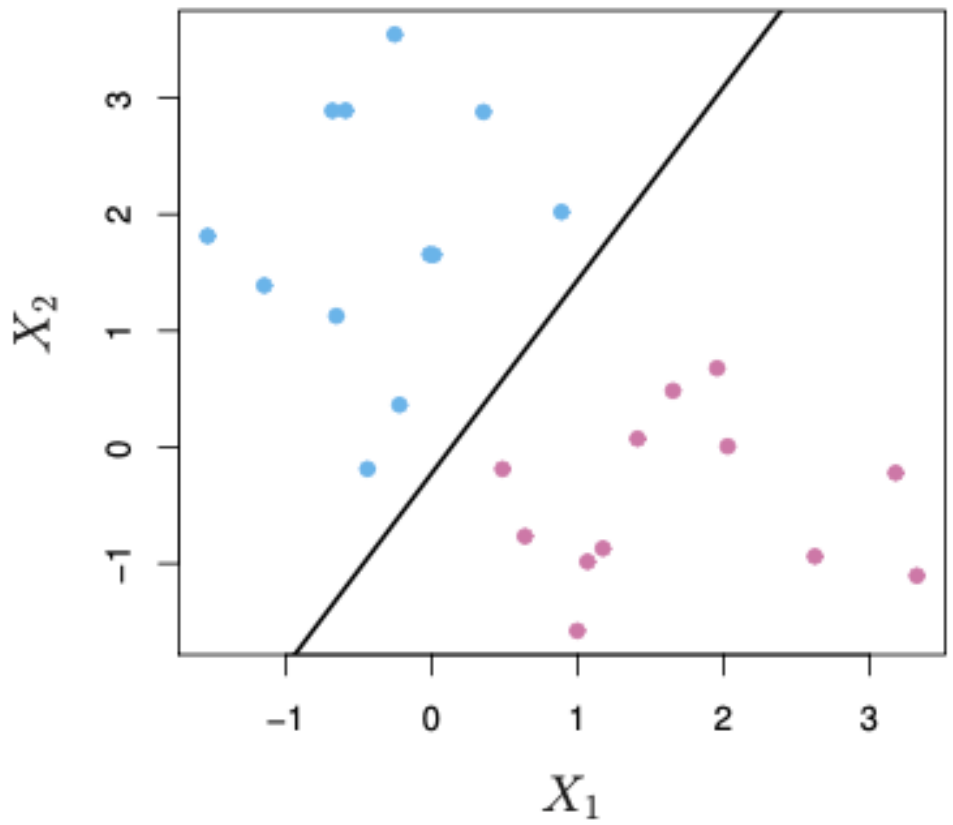
Non-separable Data

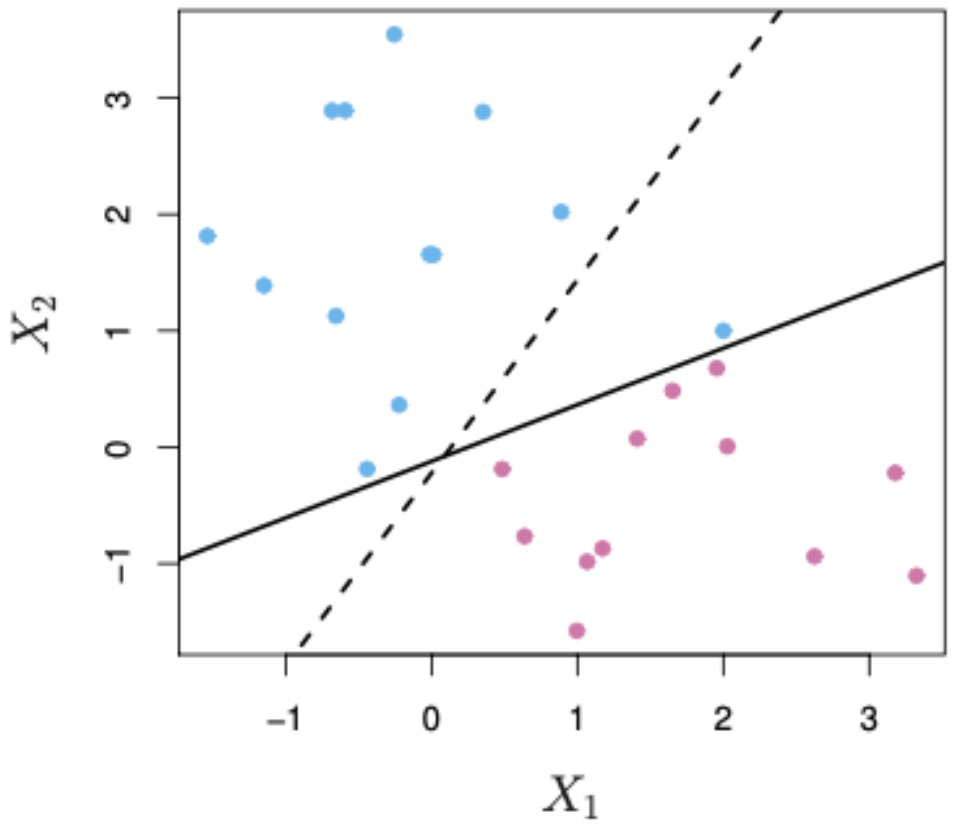


The data on the left are not separable by a linear boundary.

This is often the case, unless $N < p$.

Noisy Data





Soft Margin: Support Vector Classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

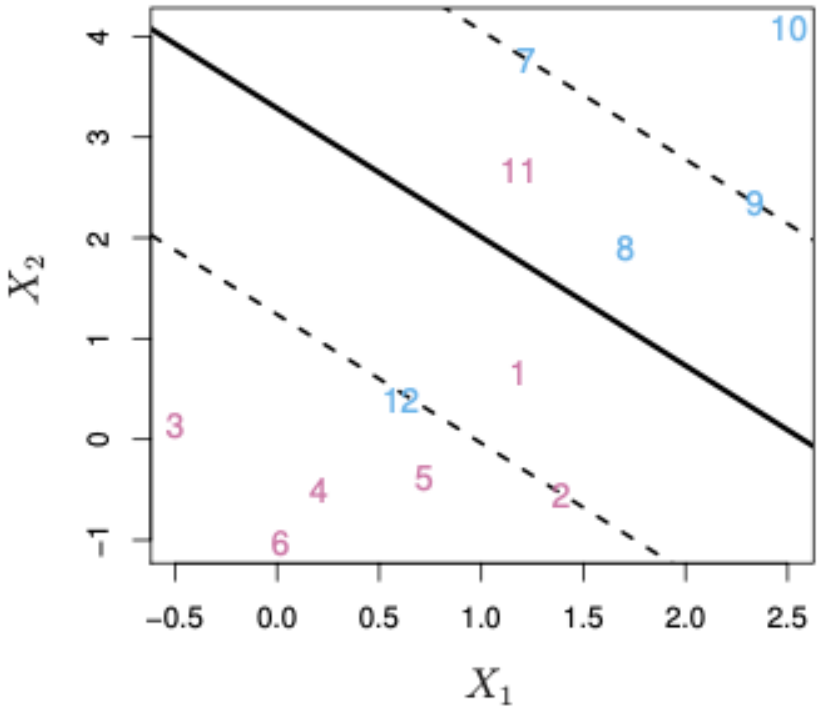
$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

Soft Margin: Support Vector Classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$



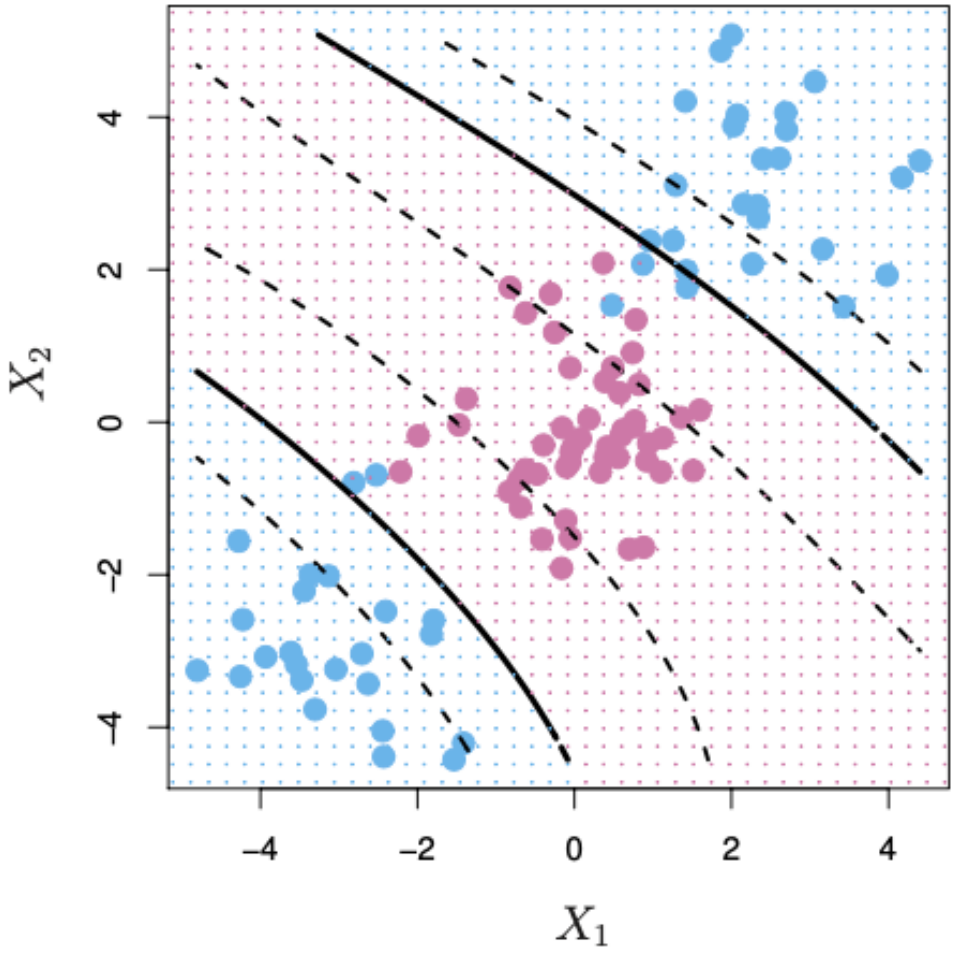
Feature Expansion

Enlarge the space of features by including transformations; e.g. X_1^2, X_1^3, X_1X_2, $X_1X_2^2$,.... Hence go from a p-dimensional space to a $M > p$ dimensional space.

Fit a support-vector classifier in the enlarged space.

This results in non-linear decision boundaries in the original space.

Cubic Polynomials



A basis expansion of cubic polynomials increases the number of variables from 2 to 9

Nonlinearities and Kernels

There is a more elegant and controlled way to introduce nonlinearities in support-vector classifiers — through the use of kernels.

Kernels: Linear Kernels; Polynomial Kernels; Gaussian Kernels

Inner products and support vectors

Inner product:

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

The linear support vector classifier can be represented as

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

Nonlinearities and Kernels

There is a more elegant and controlled way to introduce nonlinearities in support-vector classifiers — through the use of kernels.

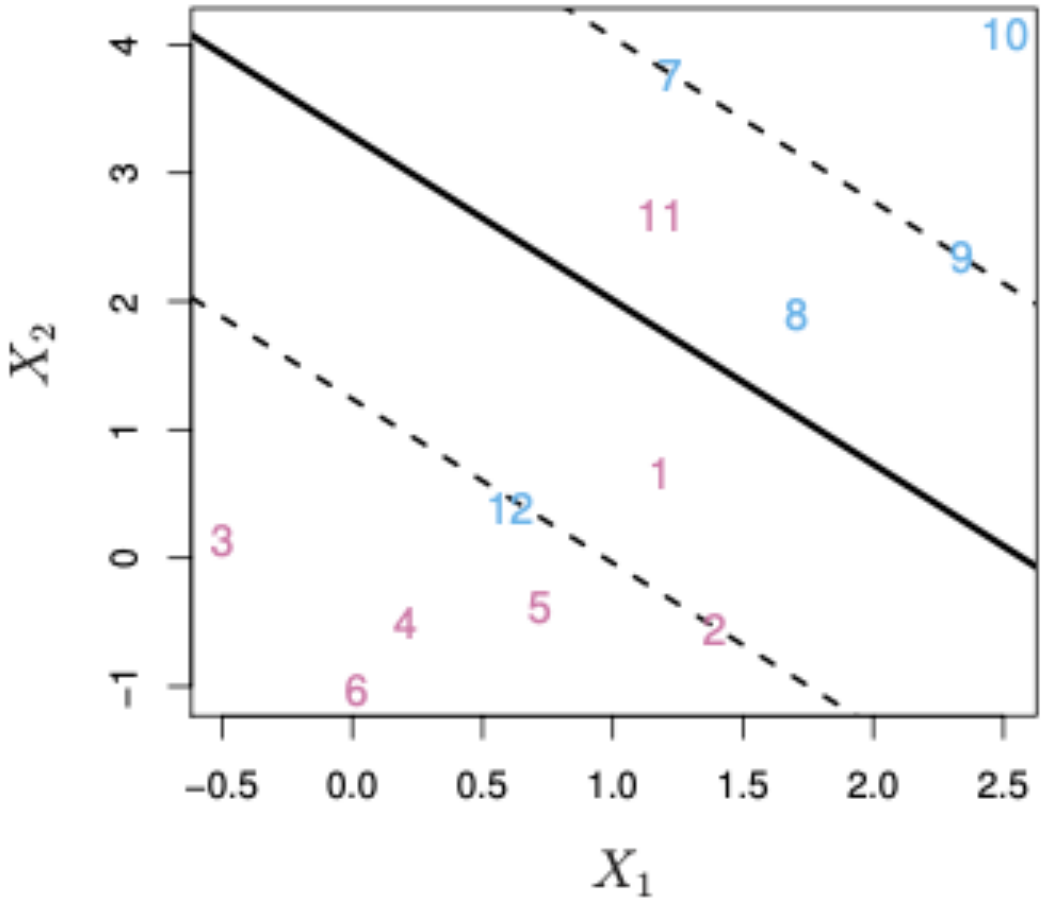
Kernels: Linear Kernels; Polynomial Kernels; Gaussian Kernels

Nonlinearities and Kernels

It turns out that most of the $\hat{\alpha}_i$ can be zero

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i \langle x, x_i \rangle$$

S is the support set of indices i such that $\hat{\alpha}_i > 0$

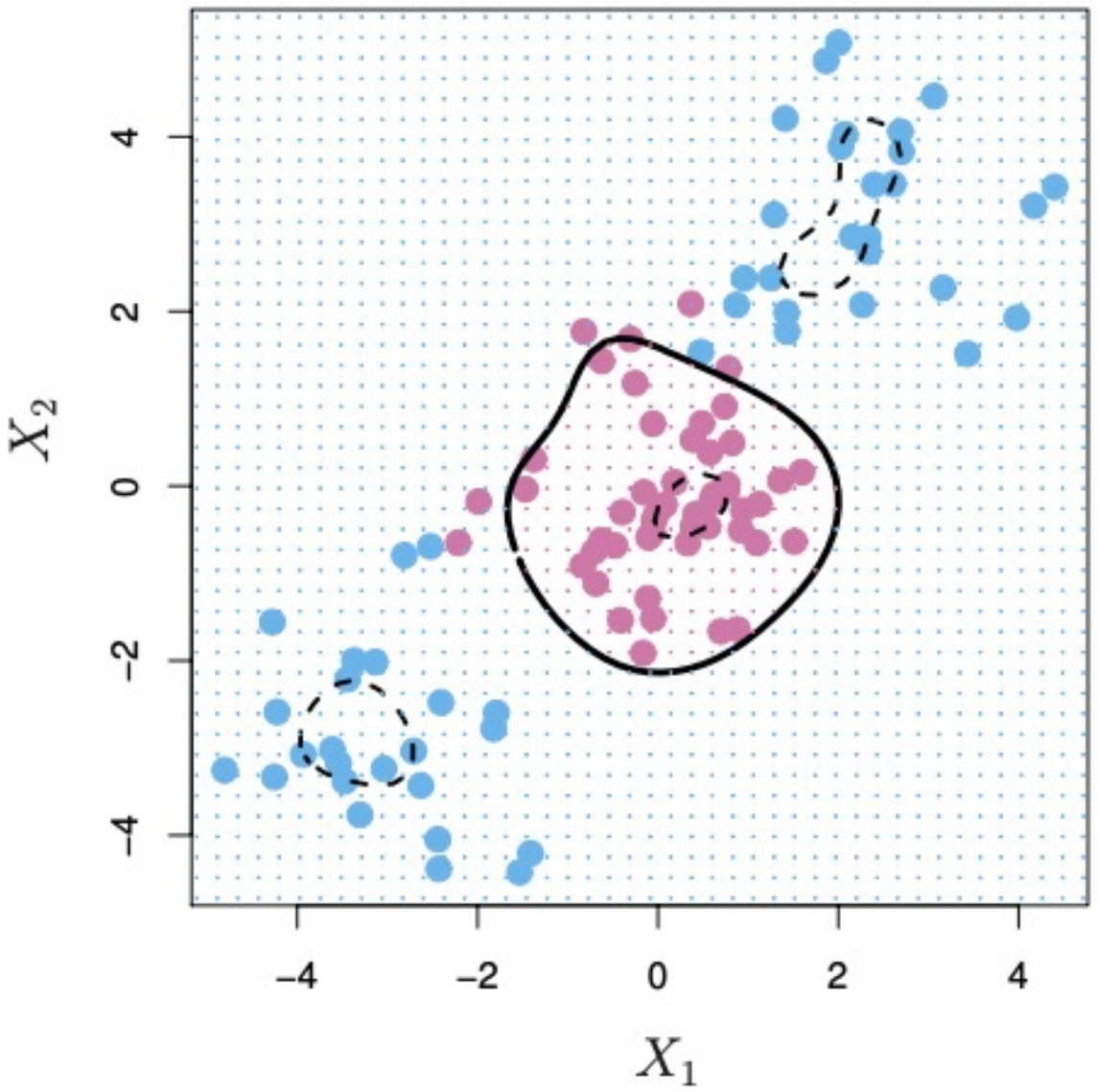


Radial Kernel

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i K(x, x_i)$$

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

The kernel is large when x is close to x_i



Multi-Class SVM

The SVM as defined works for $K = 2$ classes. What do we do if we have $K > 2$ classes?

OVA One versus All. Fit K different 2-class SVM classifiers $\hat{f}_k(x)$, $k = 1, \ldots, K$; each class versus the rest. Classify x^* to the class for which $\hat{f}_k(x^*)$ is largest.

Multi-Class SVM

The SVM as defined works for $K = 2$ classes. What do we do if we have $K > 2$ classes?

OVO One versus One. Fit all $\binom{K}{2}$ pairwise classifiers $\hat{f}_{k\ell}(x)$. Classify x^* to the class that wins the most pairwise competitions.

Support Vector vs Logistic Regression

When classes are (nearly) separable, SVM does better than LR.
So does LDA.

When not, LR (with ridge penalty) and SVM very similar.

If you wish to estimate probabilities, LR is the choice.

For nonlinear boundaries, kernel SVMs are popular. Can use kernels with LR and LDA as well, but computations are more expensive.

Q & A