# Support Vector Machines

Often considered one of the best "out of the box" classifiers.

Flow

1. **Maximal Margin Classifier**
   Separable data
2. **Support Vector Classifiers**
   Non-separable data
3. **Support Vector Machines**
   Non Linear class boundaries

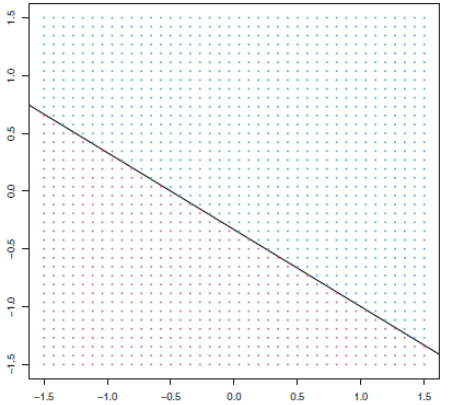
# Maximal Margin Classifier

Divides P dimensional space into two parts

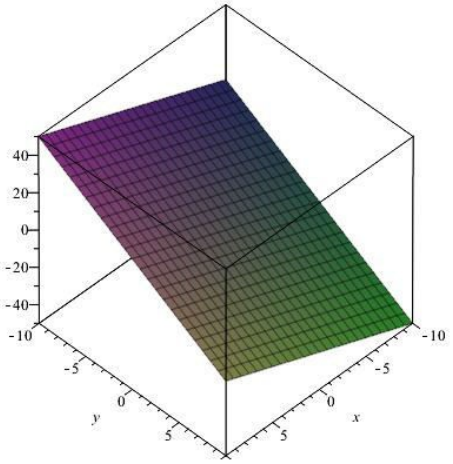
## Hyperplane

1. One Dimensional space

2. Two Dimensional space




3. Three Dimensional space
   Will be a 2 Dimensional plane

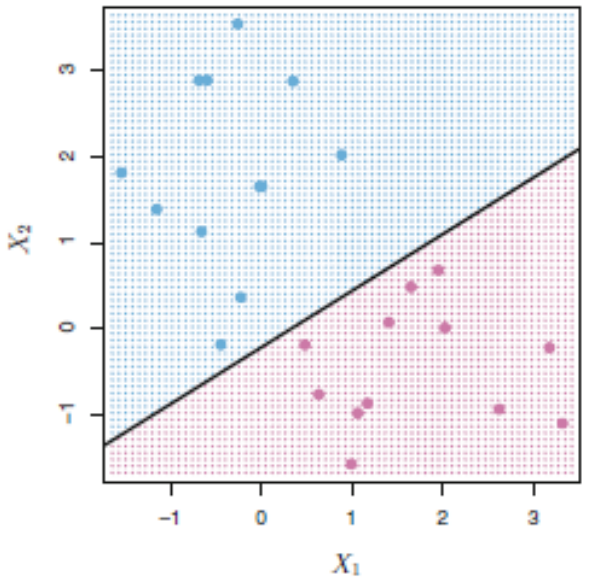


# Maximal Margin Classifier

## Hyperplane

| X1 | X2 | Category |
|---|---|---|
| 60 | 82 | Pass |
| 20 | 42 | Fail |
| … | … | … |
| 91 | 72 | Pass |

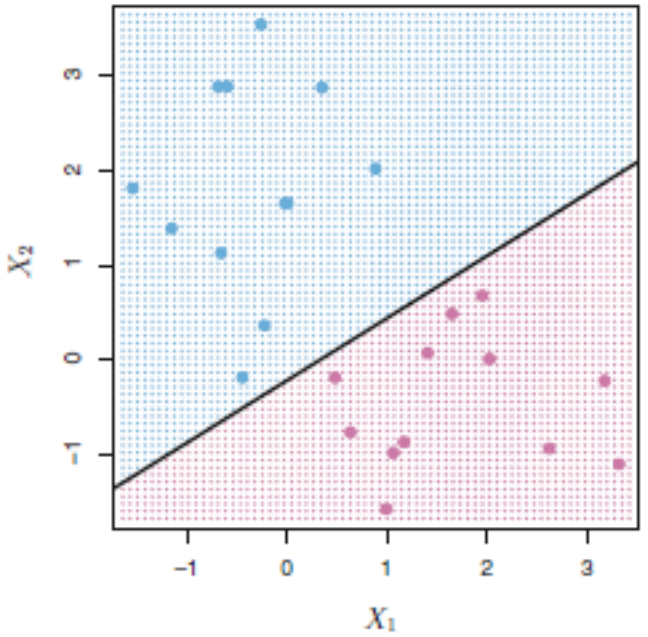
- Two predictor variables ->
  2D predictor space
- We want to find 1D (Line) hyperplane which separates this space into 2 parts

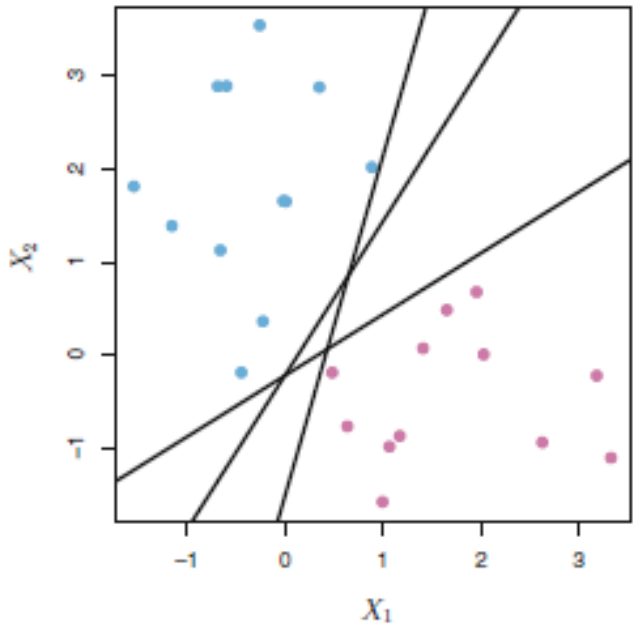


# Maximal Margin Classifier

## Infinite hyper planes

If data is perfectly separable



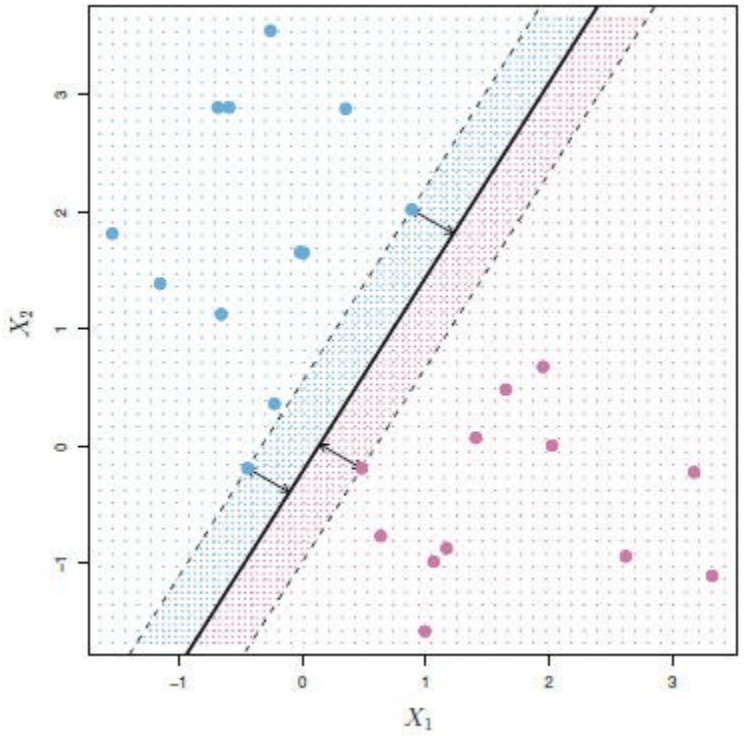




There are infinite hyper plane

# Maximal Margin Classifier

## Steps

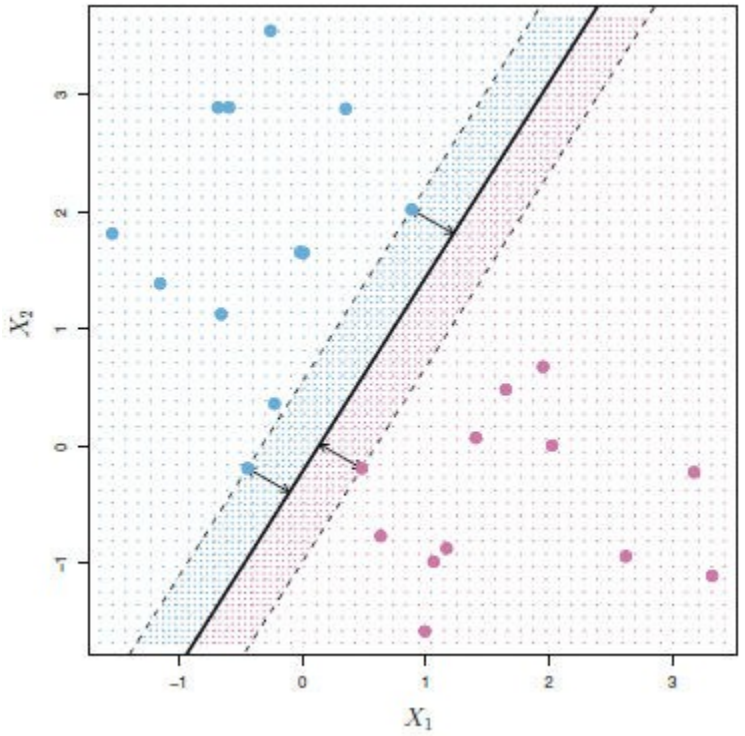
1. Calculate the perpendicular distance of observations from Hyperplane
2. Minimum value of distance is called margin
3. Choose the Hyperplane with maximum value of Margin



# Maximal Margin Classifier

## Support Vectors

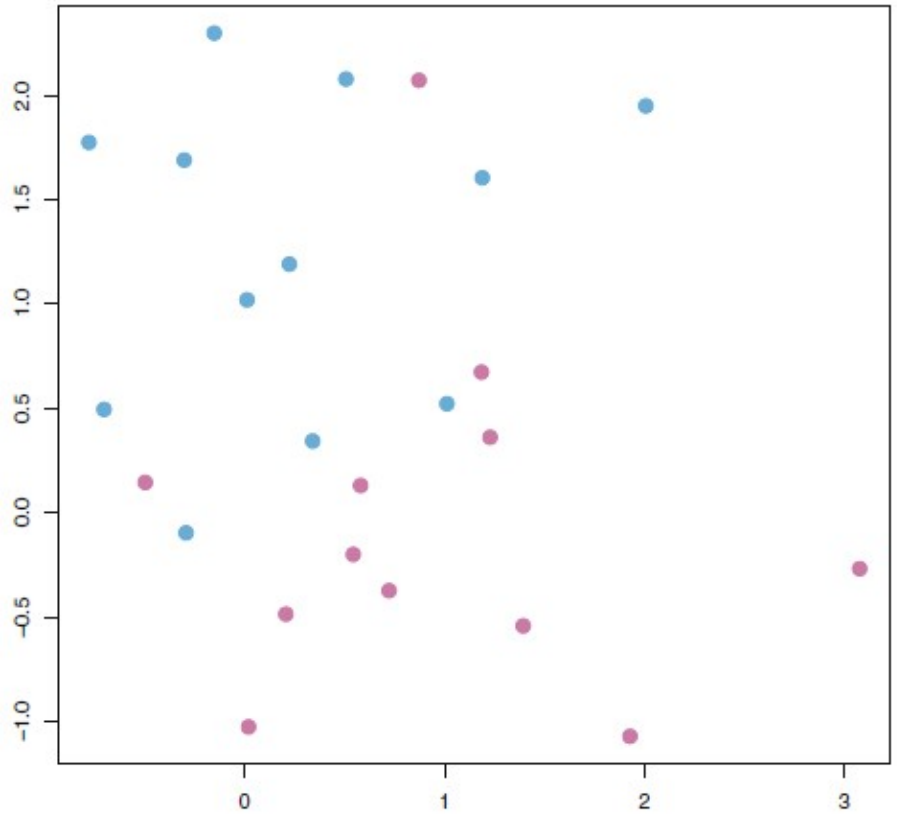
1. The observations which fall on margin are known as Support Vectors
2. These classifiers depend on support vectors only
3. That is why this technique is different from conventional ML techniques



# Maximal Margin Classifier

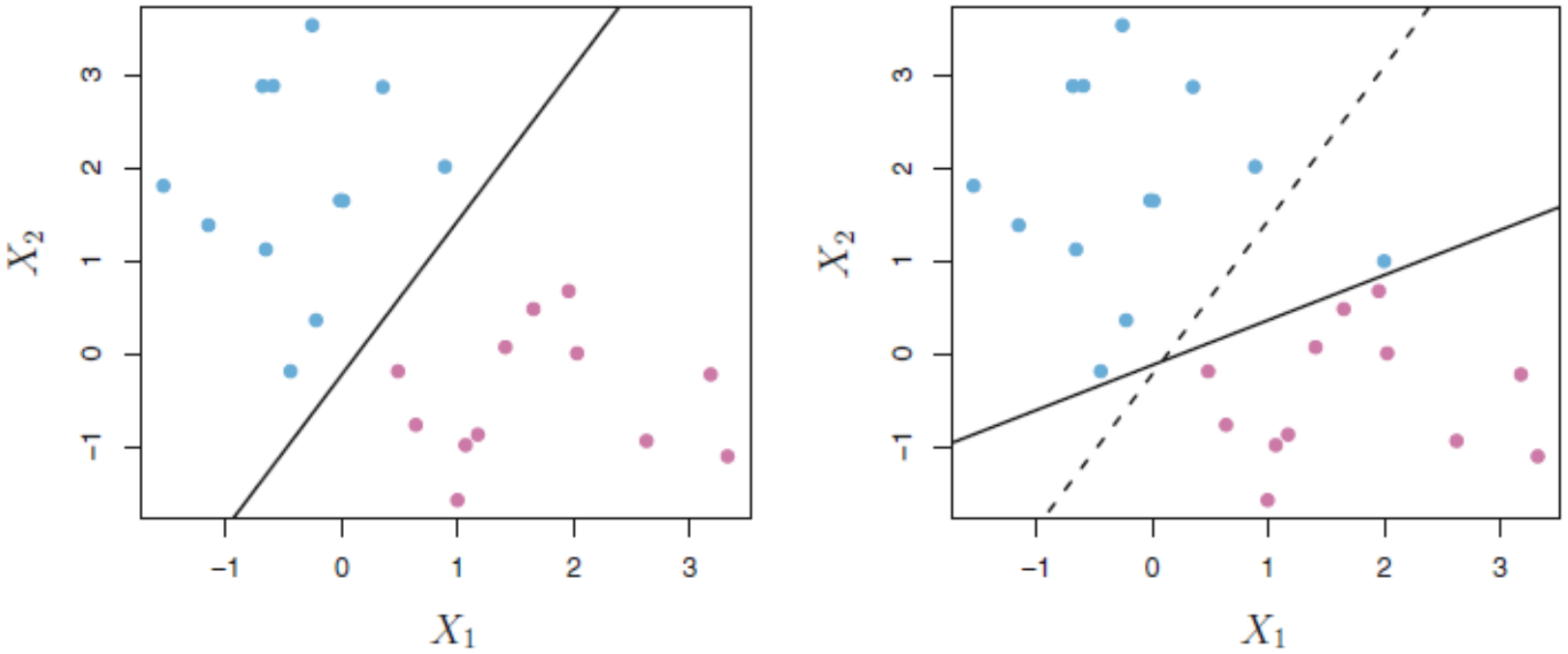
## Limitation 1

Maximal margin classifier cannot be used if the two classes are not separable by a hyperplane



# Maximal Margin Classifier

## Limitation 2




Maximal margin classifier is very sensitive to support vectors, an additional observation can lead to a dramatic shift in the maximal margin hyperplane

# Support Vector Classifier

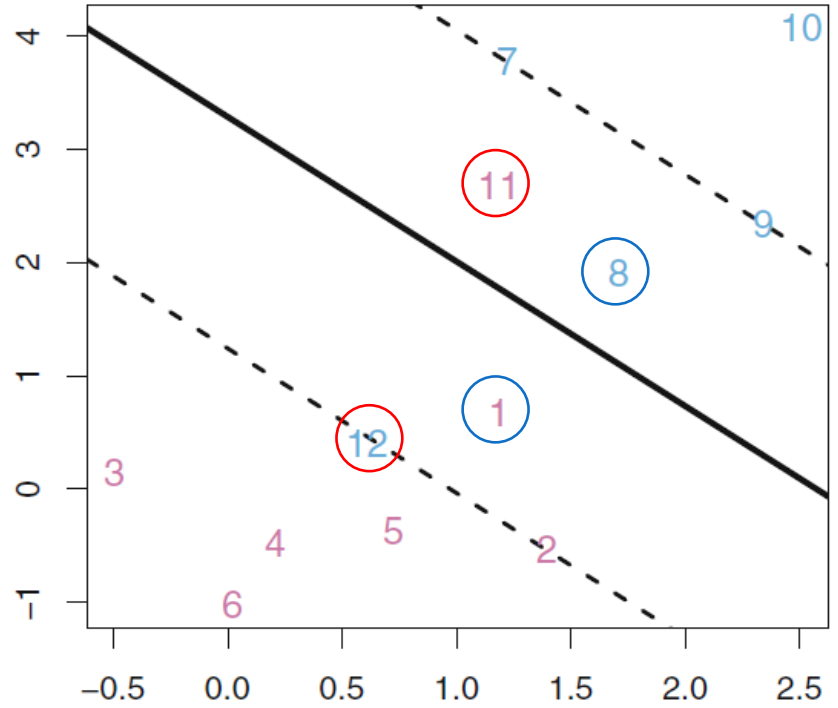
## Why

1. To handle non perfectly separable scenario
2. Greater robustness to individual observations

# Support Vector Classifier

## What

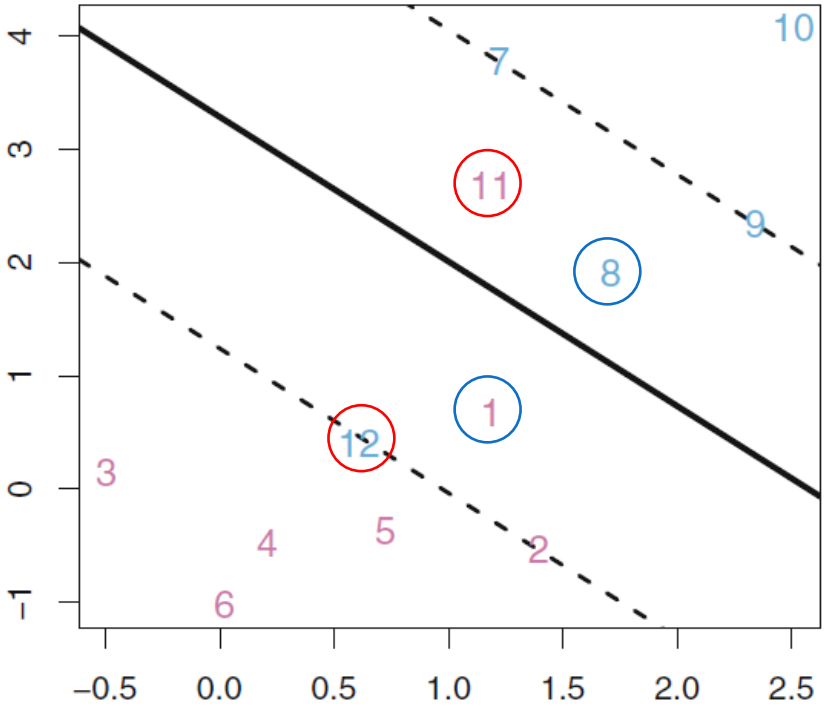
1. Support vector classifier is a soft margin classifier
2. We will allow some observations to be incorrectly classify or to be on the wrong side of the margin



# Support Vector Classifier

## How

1. We create a misclassification budget ( B )
2. We limit sum of distances of the points on the wrong side of the margin
   $(x1 + x2 + x3 + x4) < B$
3. We try to maximize margin while trying to stay within budget
4. Usually in our software packages we use C (Cost - multiplier of the error term) which is inversely related to B



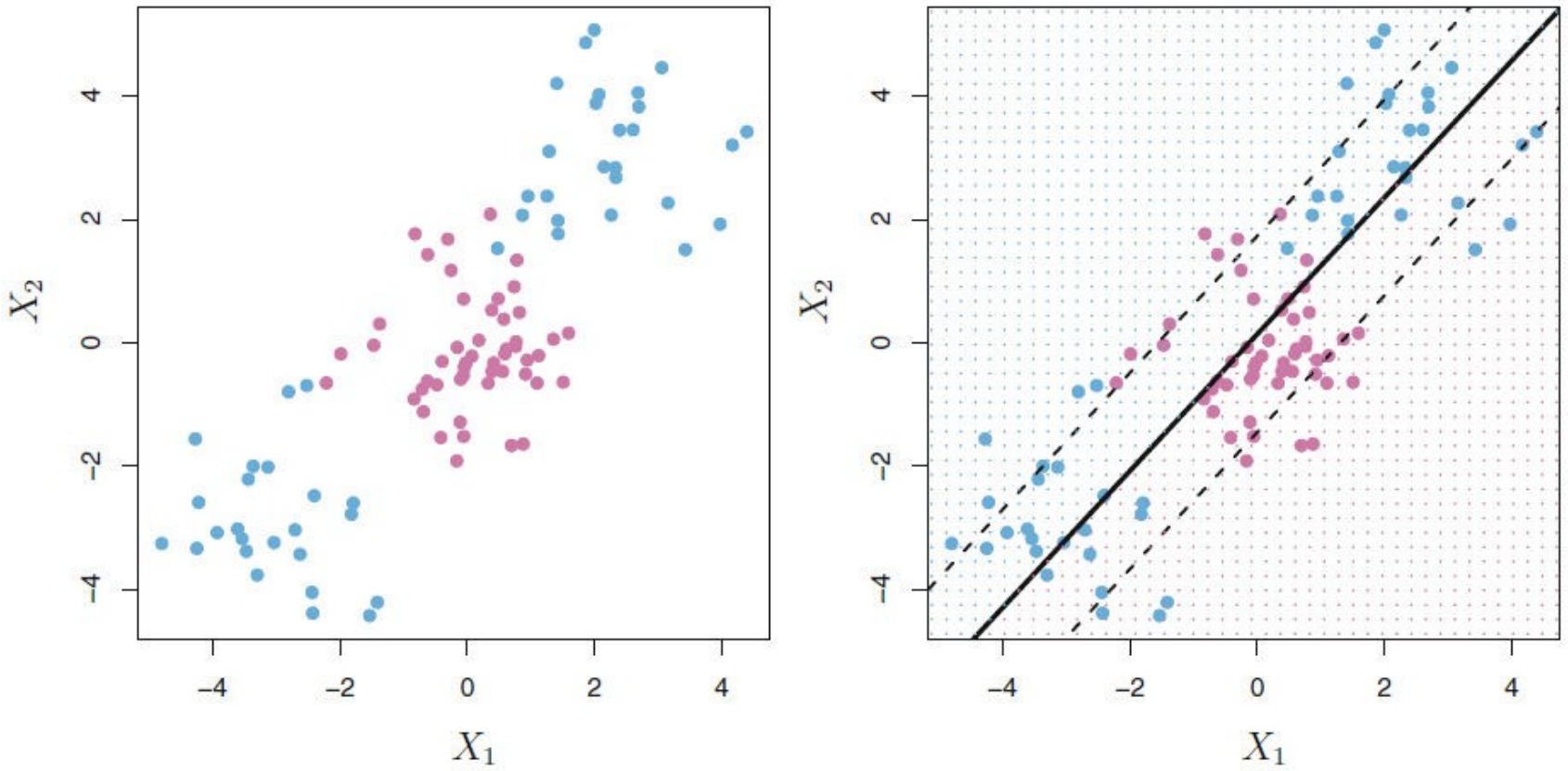
# Support Vector Classifier

## Impact of C

1. When C is small, margins will be wide and there will be many support vectors and many misclassified observations
2. When C is large , margins will be narrow and there will be fewer support vectors and fewer misclassified values
3. However, low cost value prevents overfitting and may give better test set performance
4. We try to find optimal value of C at which we get best test performance
5. https://cs.stanford.edu/~karpathy/svmjs/demo/

# Support Vector Classifier

## Limitation 1




Support Vector classifier is a linear classifier, it cannot classify non linear separable data

# Support Vector Machines

## What

Support vector machine (SVM) is an extension of the support vector classifier which uses Kernels to create non linear boundaries

Kernels

Some functional relationship between two observations.
Some popular kernels

1. Linear
2. Polynomial
3. Radial

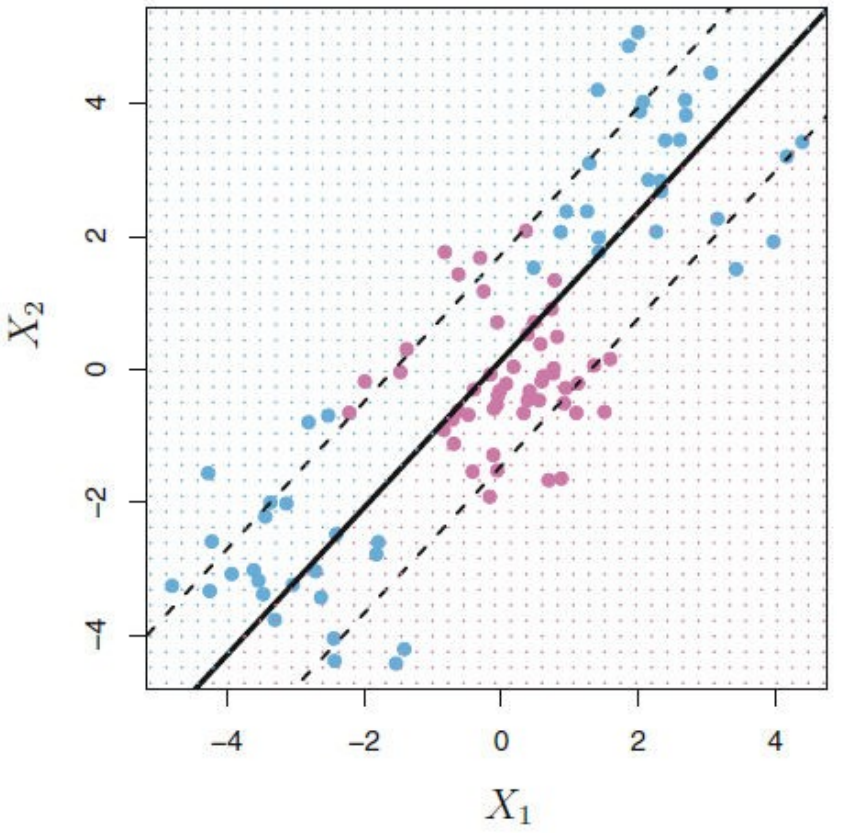
# Support Vector Machines

## Linear Kernel

Linear kernel takes inner product of two observations

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

This kernel effectively is a support vector classifier

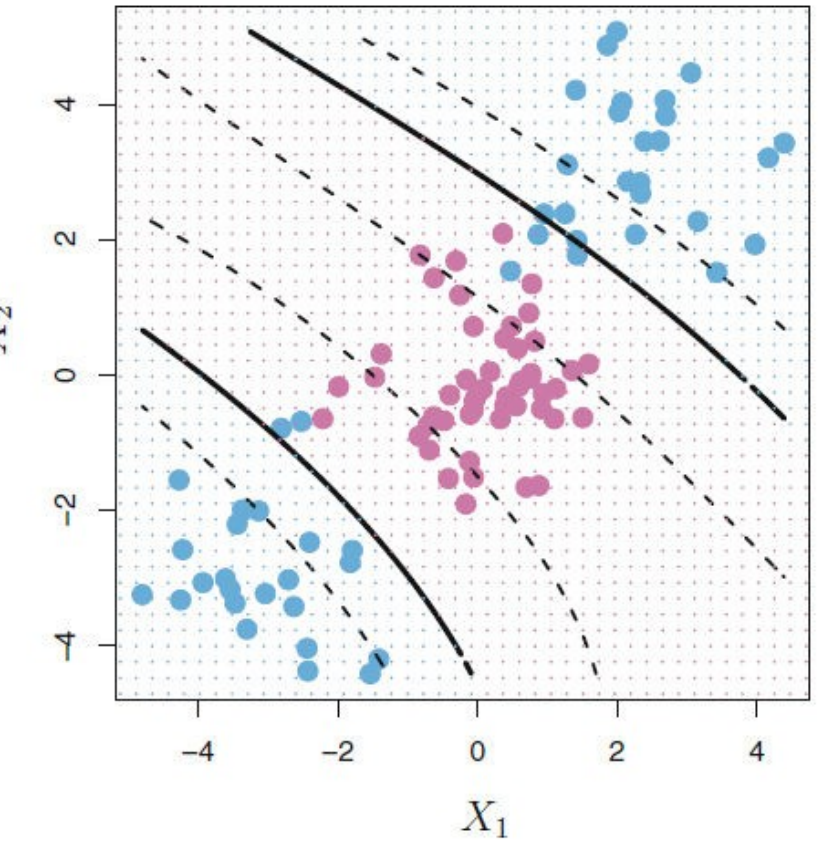


# Support Vector Machines

## Polynomial Kernel

Polynomial kernel uses power function to create non linear boundaries

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$$



# Support Vector Machines

## Radial Kernel

Radial kernel uses radial function to create radial boundaries

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

$\gamma$ *is a positive constant*

Gamma defines how much influence a single training example has. The larger gamma is, the closer other examples must be to be affected.

https://cs.stanford.edu/~karpathy/svmjs/demo/