TECHNO INDIA
SILIGURI INSTITUTE
OF TECHNOLOGY
A Satyam Roychowdhury Initiative

Support Vector Machines

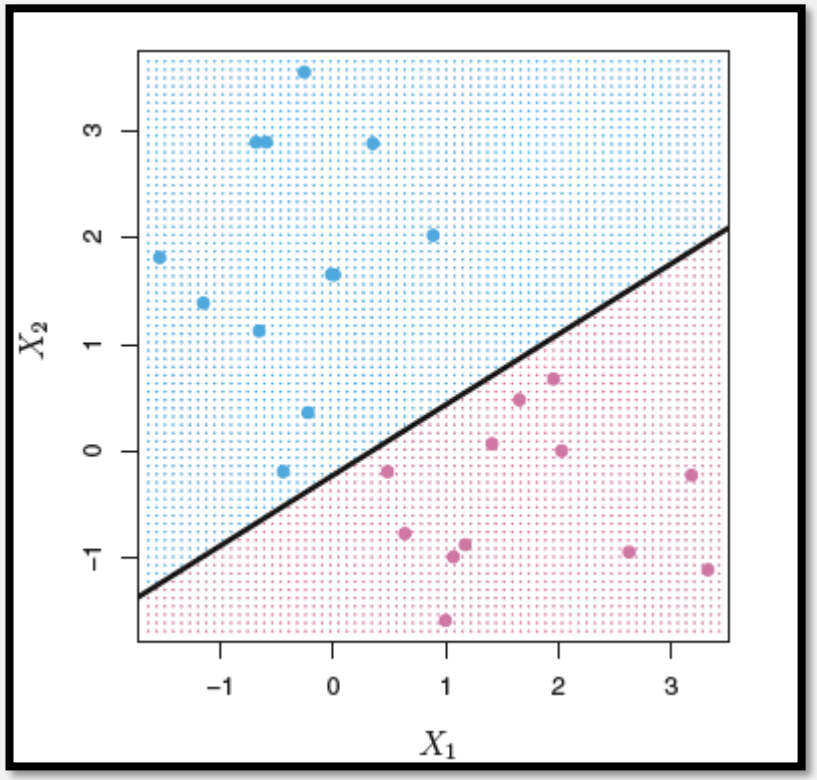
Maximum Margin Line Separators, Quadratic Programming Solution, Kernels for non linear functions

Introduction

- Support Vector Machine was developed for classification problem in 1990
- It is a generalized and simple classifier and called maximal margin classifier.
- Maximal Margin Classifier is an optimal separating hyperplane

Hyperplane

- A hyperplane is a flat affine subspace of d-1 dimension in a d-dimensional space.
- The equation of hyperplane in p-dimensional space is
- $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0$
- Any point x lies on the hyperplane satisfies the above equation. Suppose for an x $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p > 0$, then this tells that x lies on one side of the hyperplane. If $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0$,then x lies on the other side of the plane.
- Therefore, the p −dimensional space is divided into two halves. It can easily determine on which side of the hyperplane the point x lies.



Classification using a separating hyperplane

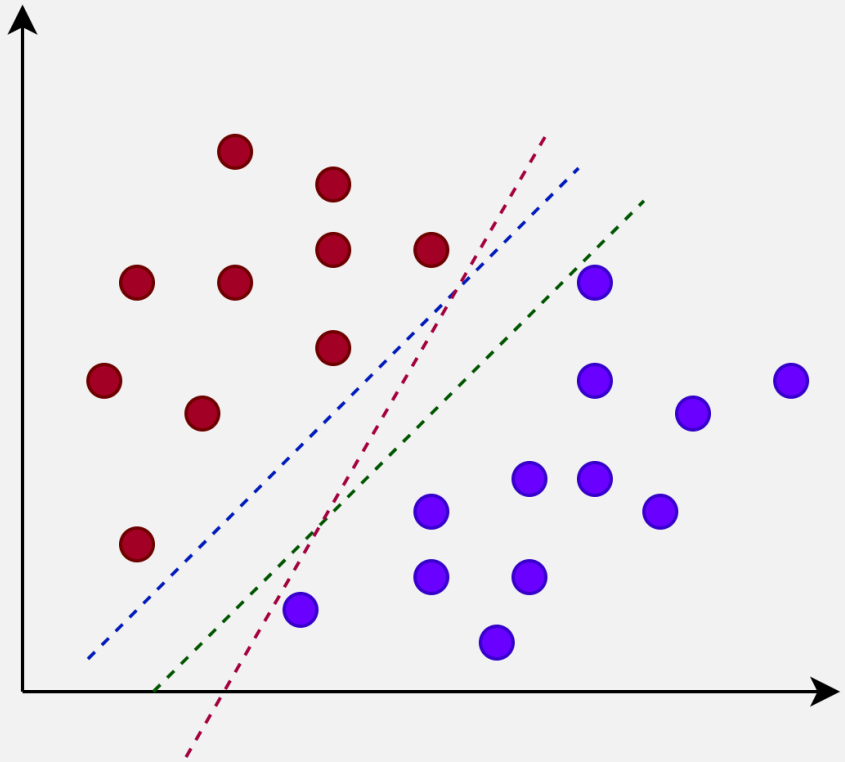
- For $n \times p$ dataset consisting n observation and p variables in p dimension space and the label of the datapoint fall into two class $\{-1, +1\}$. Where 1 represent one class and -1 represent another class.
- For the test observation x^*, , the sign of $f(x^*)$ is used to determine the class of the observation. Where $f(x^*) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
- If $f(x^*) = 0$, then the point lies on the hyperplane.
- If $f(x^*) > 0$, then the point x^* belongs to class $+1$. $f(x^*) < 0$ indicate the point lies to class -1.

Classification using a separating hyperplane

- The magnitude of $f(x^*)$ can be used to take decision about the point x^*. The point x^* will be far from the separating hyperplane if $f(x^*)$ far from the separating hyperplane.
- If $f(x^*)$ is close to zero the point x^* is close to separating plane.

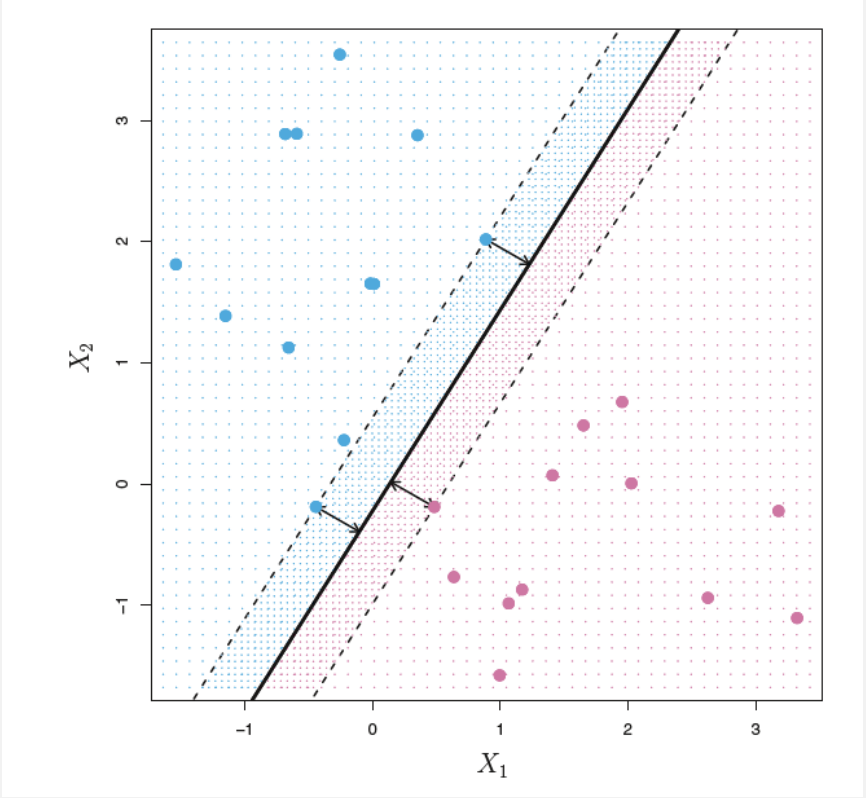
Classification using a separating hyperplane

- Infinite number of lines possible for separating the two classes. We must have a reasonable way to decide which of the infinite possible separating hyperplane to use.



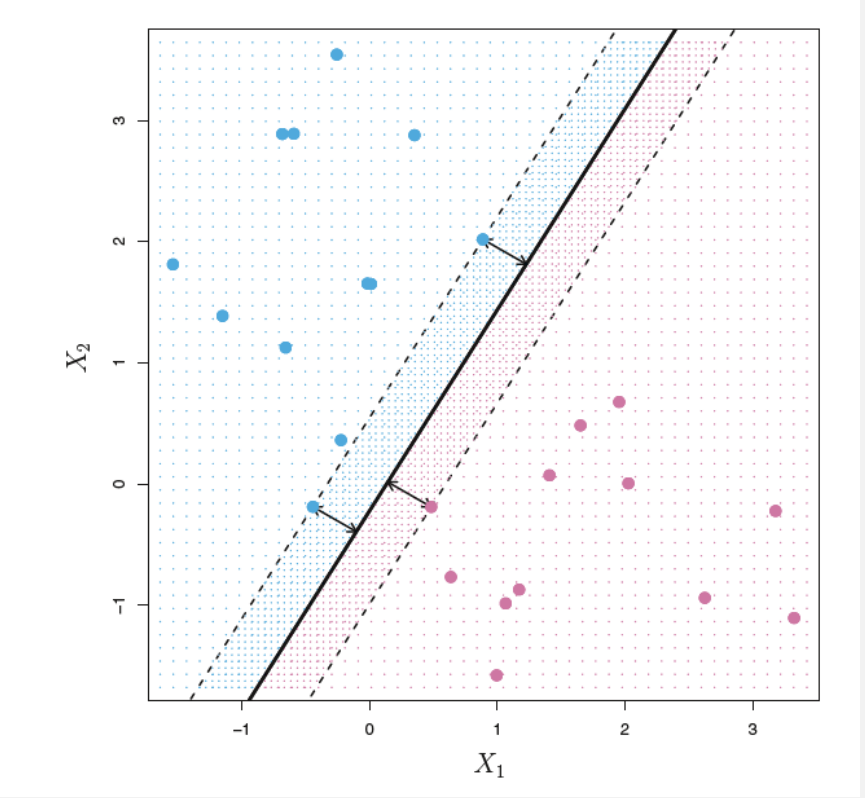
Maximum Margin Line Separator

- the separating hyperplane is farthest from the training observations.
- the perpendicular distance from each training observation to a given separating hyperplane is calculated.
- the smallest of such distance is the minimum distance from the observations to the hyperplane is known as margin.
- the maximal margin hyperplane is the separating hyperplane for which the margin is largest.



Maximum Margin Line Separator

- the maximal margin hyperplane represents the mid-line of the widest "slab" that we can insert between the two classes.
- the points whose slight move force the hyperplane to move as well are called support vectors.
- the movement of other points will not effect the separating plane.
- the maximal margin hyperplane depends directly on a small suset of observations only.



Construction of the Maximal Margin Classifier

- The optimization problem corresponding to the solution of maximal margin classifier is

$$Maximize_{\beta_0,\beta_1,\ldots,\beta_p} M$$

$$Subject\ to$$

$$\sum_{i=1}^{p} \beta_i^2 = 1$$

$$y_i\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right) \geq \mathrm{M}$$

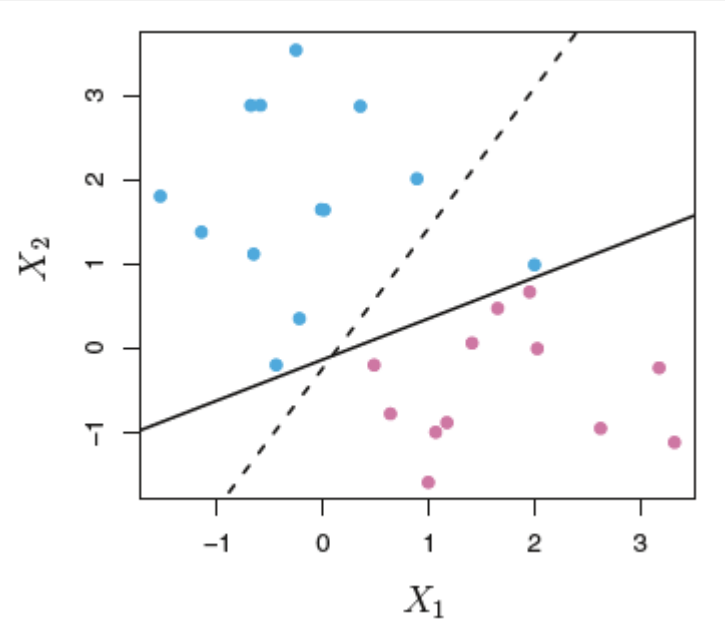
M represents the margin of our hyperplane, and the optimization problem choose $\beta_0, \beta_1, \ldots, \beta_p$ to maximize M

The non separable case

- If the observations are not linearly separable then no separating hyperplane exists, and so there is no maximal margin classifier. The optimization problem has no $M > 0$.
- The concept of separating hyperplane is extended in order to develop a hyperplane that almost separate the classes using soft margin.
- The generalization of soft margin is called support vector classifier.

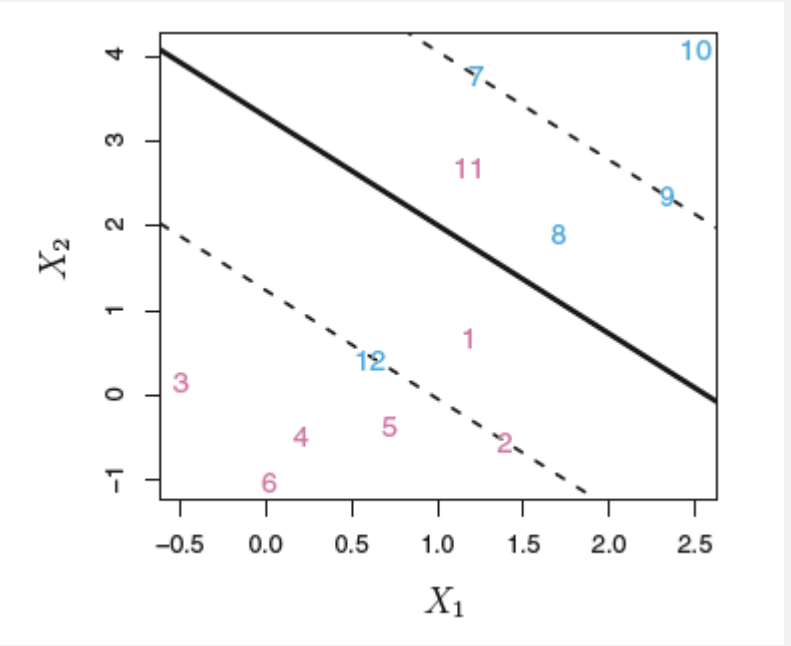
Support Vector Classifier

- Observations belong to different class may not be linearly separable.
- A separable hyperplane may exist, but the separating hyperplane may not be the desirable one.
- The maximal margin hyperplane is extremely sensitive to a change in a single observation suggest that it may overfit the training data.



Support Vector Classifier

- It could be worthwhile to misclassify a few training observations in order to do a better job in classifying observations.
- The support vector classifier sometimes called soft margin classifier.
- Rather than seeking the largest possible margin so that every observation on the correct side of the hyperplane, but also correct side of the margin and allowing some of the observation on the incorrect side.



Support Vector Classifier

- $maximize_{\beta_0,\beta_1,\ldots,\beta_p \; and \; \epsilon_1,\ldots,\epsilon_N} \; M$
- Subject to
- $\sum_{j=1}^{p} \beta_j^2 = 1$
- $y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \epsilon i)$
- $\epsilon_i \geq 0$, and $\sum \epsilon_i \leq c$
- Where $c \; is \; a \; nonnegative \; tuning \; parameter$.

Support Vector Machine

- The SVM is an extension of support vector classifier that result from enlarging the feature space in a specific way using kernel.
- We enlarge our feature space in order to accommodate a nonlinear boundary between the classes.
- The inner product of two vectors a and b is defined as $< a, b > = \sum_i a_i b_i$
- Thus the inner product of two observations x_i and x_i' is given by

$$< x_i, x_{i'} > = \sum x_{ij} xi'j$$

Support Vector Machine

- The linear support vector can be represented as
- $f(x) = x_0 + \sum_{i=1}^{n} \alpha_i < x_i, x_i' >$
- We replace the inner product <> with a generalization of the inner product of the form $K(x_i, x_i')$.
- The linear kernel of the support vector classifier can be written as

$$f(x) = x_0 + \sum_{i=1}^{n} \alpha_i K(x_i, x_i')$$

Support Vector Machine

- The polynomial kernel of degree d is
- $K(x_i, x_i') = \left(1 + \sum x_{ij} x_{i'j}\right)^d$, where d is positive integer
- Radial kernel takes the form
- $K(x_i, x_i') = exp(-\gamma \sum \left(x_{ij} - x_{i'j}\right)^2)$

SVMs with More than Two Classes

- the concept of separating hyperplanes upon which SVMs are based does not lend itself naturally to more than two classes.
- Though a number of proposals for extending SVMs to the K-class case have been made, the two most popular are the one-versus-one and one-versus-all approaches.

One-versus-one pair

- A one-versus-one or all-pairs approach constructs (KC2) SVMs, each of which compares a pair of classes.
- classify a test observation using each of the (KC2) classifiers, and tally the number of times that the test observation is assigned to each of the K classes.
- The final classification is performed by assigning the test observation to the class to which it was most frequently assigned in these (KC2) pairwise classifications.

One versus all

- fit K SVMs, each time comparing one of all the K classes to the remaining $K-1$ classes.
- Let $f_k(x)$ is the hyperplane for separating k −th class from remaining class. Let x^* denote the test observation.
- for which k, $f(x^*)$ having largest value , x^* belong to the kth class