Support Vector Machines

SVM

Support Vector Machines

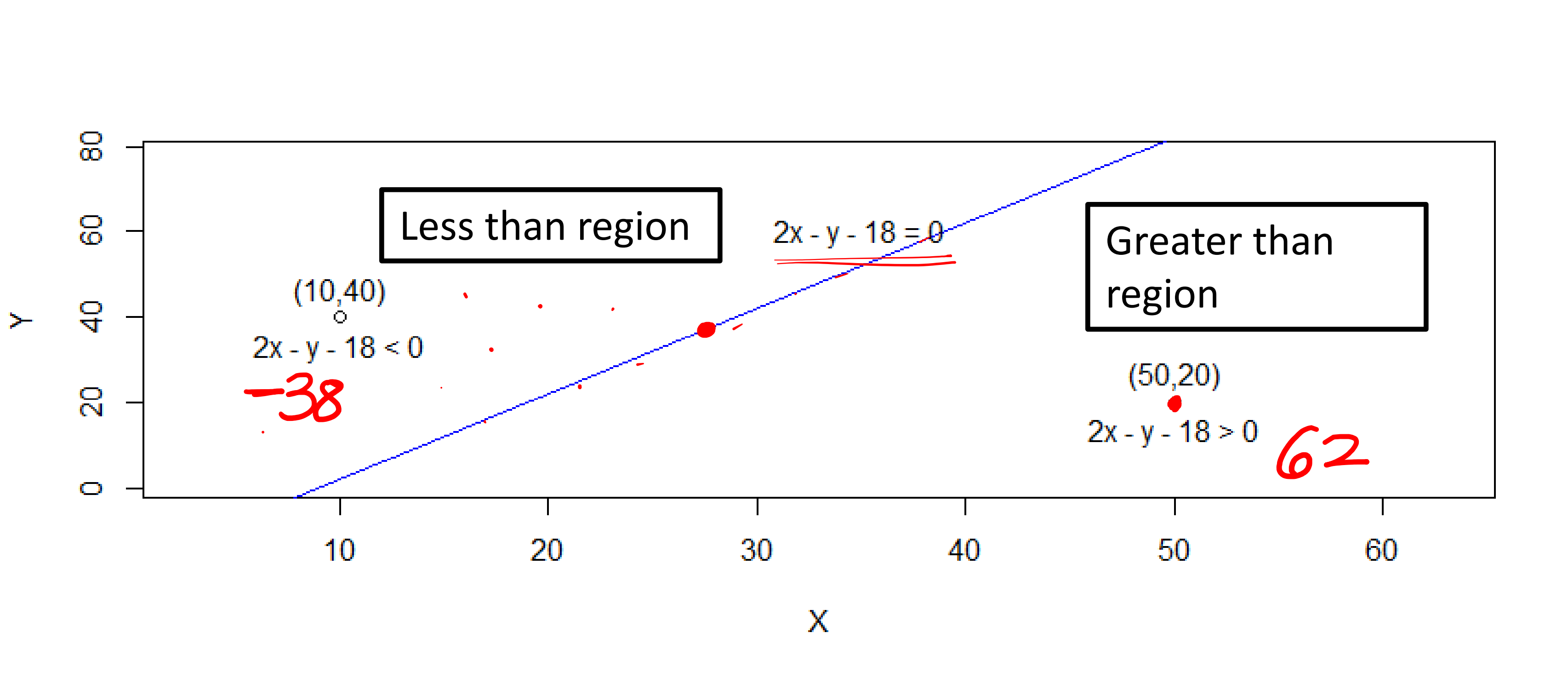
- Support Vector Machines can be used for classification as well as regression
- Usage of SVM is popular for classification than for regression
- We will be covering SVM for classification.

Understanding SVM

- SVM is a generalization of a simple classifier *maximum margin classifier*.
- The concept of *maximum margin classifier* can be extended to that of ***support vector classifier*** and **support vector machines**.

Straight Line Fundamentals (Revision)

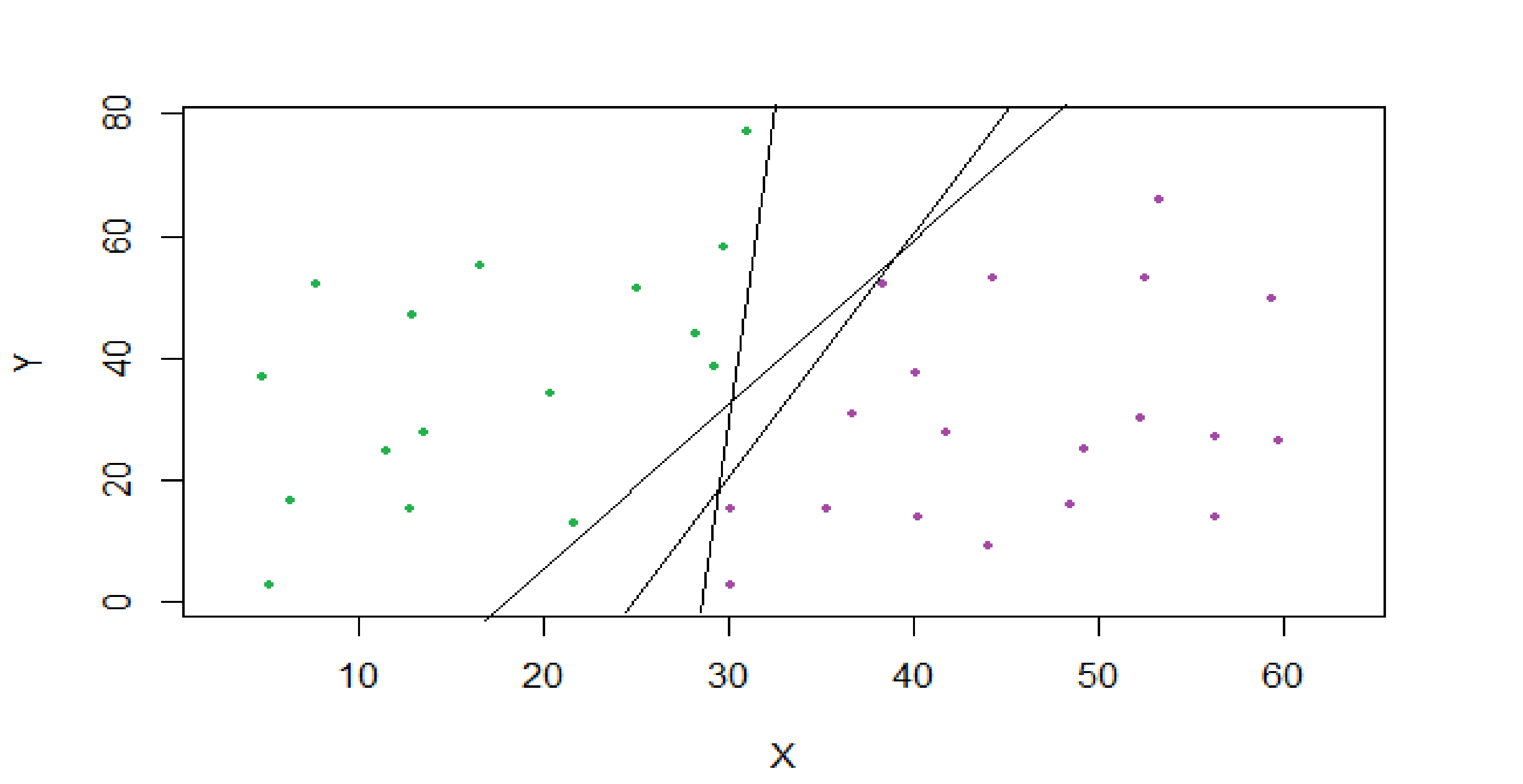
- Consider a line with equation, $ax + by + c = 0$
- Any point $(x1,y1)$ which is lying on the line satisfies the equation of the line i.e. we can write $ax1 + by1 + c = 0$.



Separating Hyperplanes

- Let us understand the concept on 2-dimensional plane which can be further extended to multi-dimensional hyperplane
- Suppose that, it is possible to have three hyperplanes for a data

Separating Hyperplanes

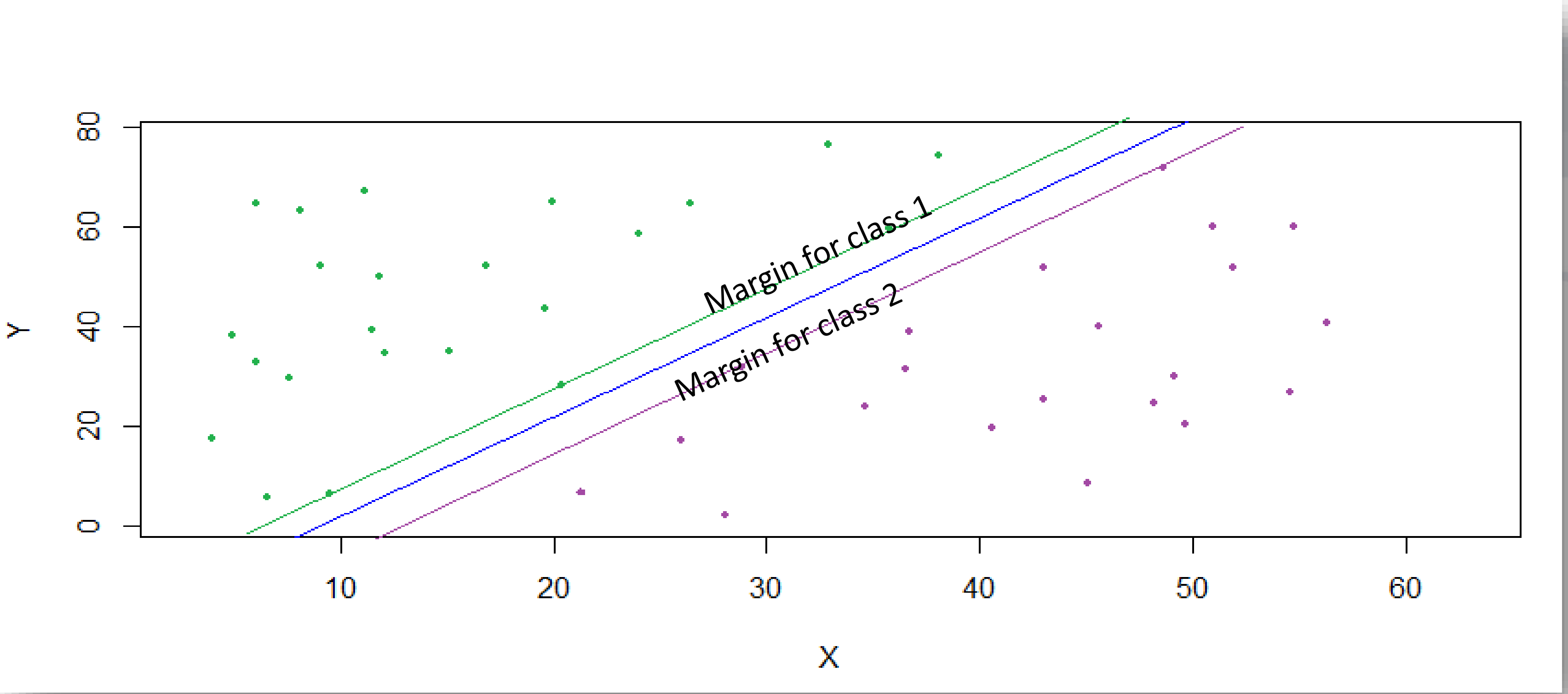


- We observe here that, three hyperplanes have separated the data. Any point which lies in the region of green points can be classified as category of green and similarly with purple.

Maximum Margin Classifier

- If our data can be perfectly separated using a hyperplane, then there will in fact exist an infinite number of such hyperplanes which will separate different categories in our response variable
- This can be made possible with a given separating hyperplane shifted a tiny bit up or down, or rotated, without coming into contact with any of the observations
- Hence we can imagine a separating hyperplane which has maximum distance from any nearest point in the data. This is called *maximum margin classifier*.

Maximum Margin Classifier



- We observe that 5 observations are equidistant from maximal margin hyperplane. These points are called *support vectors*.
- These points are called “support” in the sense that if these points were moved slightly then the maximal margin hyperplane would move as well.

Non-Separable Case

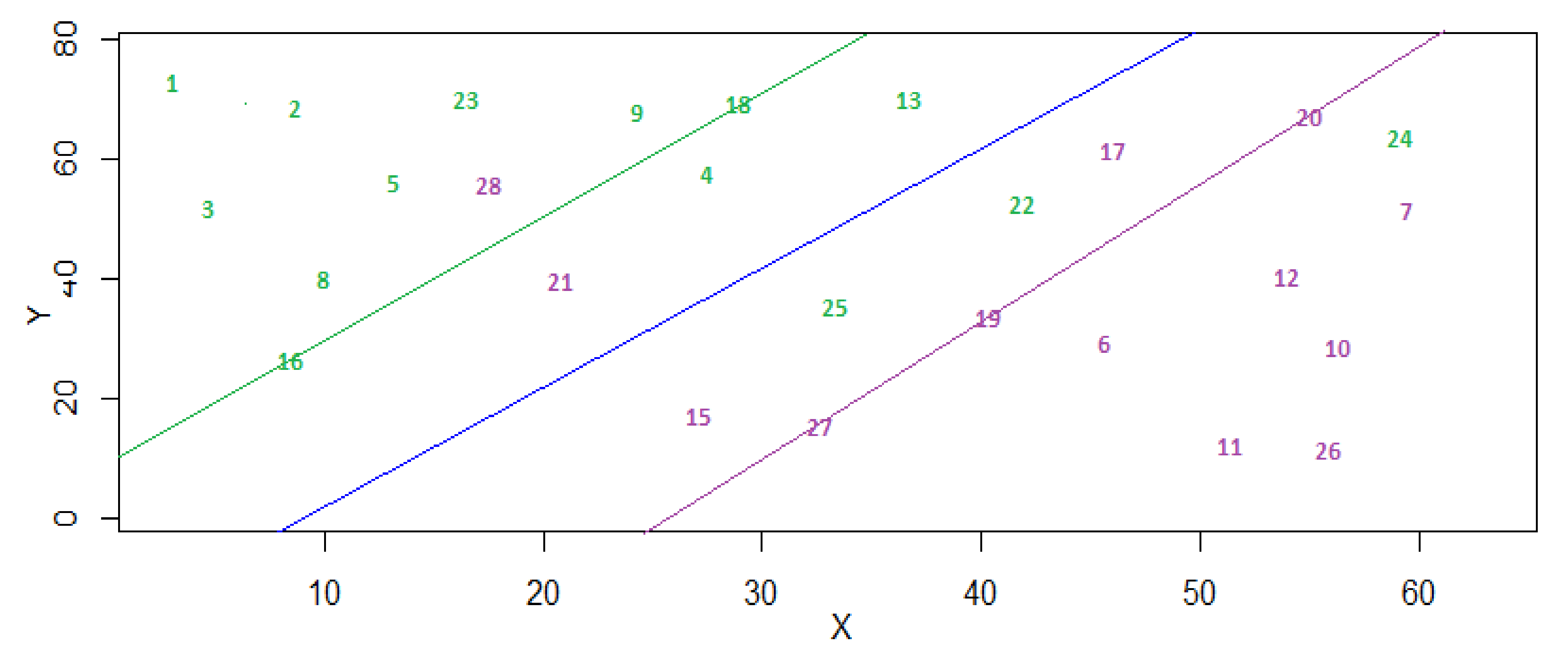
- In case, if a separating hyperplane is not available then we cannot exactly separate two classes
- Instead, we can find a hyperplane that almost separates the two classes
- A generalization of the maximal margin classifier to the non-separable case is called as the *support vector classifier*

Sane's STATS Academy of Statistics

Support Vector Classifier

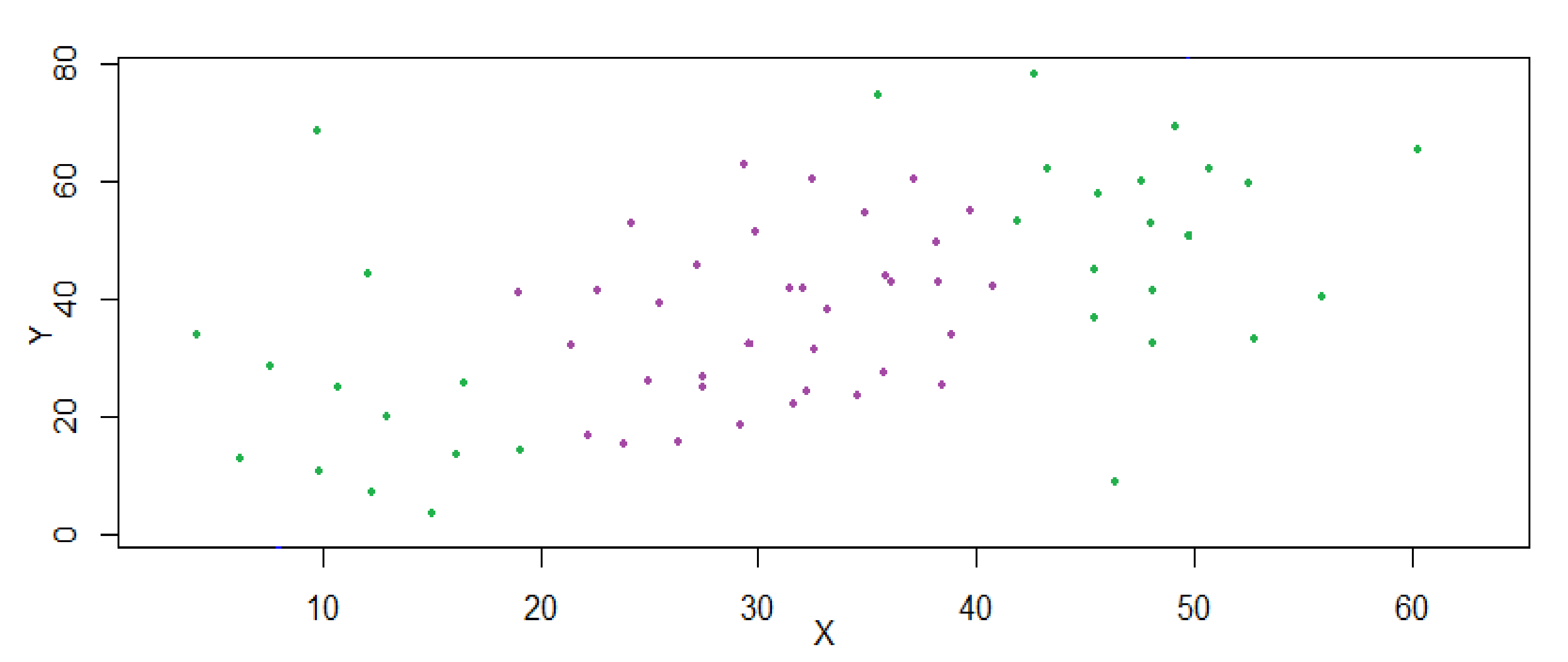
- In case, if a separating hyperplane is not available then a classifier can be considered which exactly does not separate the two classes but classifies most of the training set observations correctly
- In this case, some observations can be allowed to be on the incorrect side of the margin or also incorrect side of separating hyperplane
- This separating hyperplane can also be called as soft margin classifier as it can allow some violations

Illustration : SV Classifier



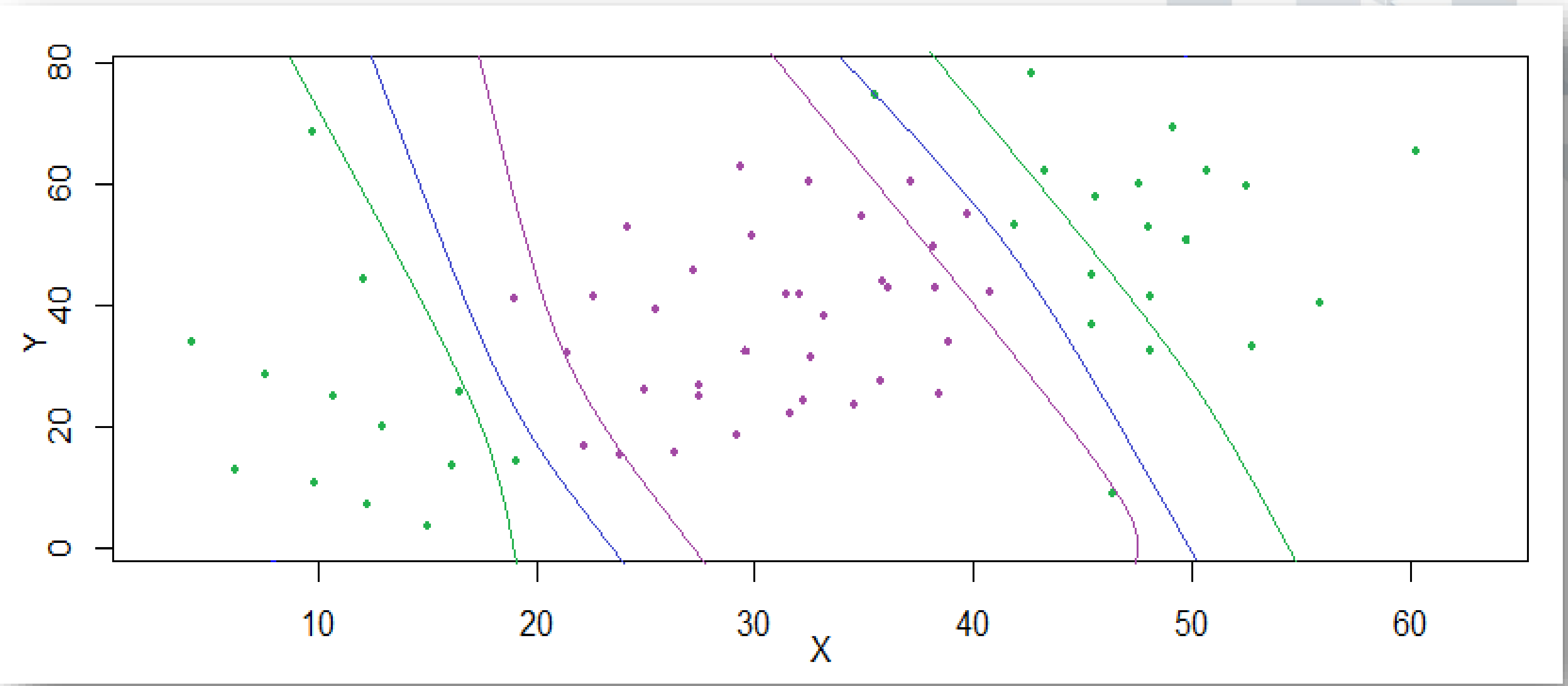
- Consider that the above diagram represents a support vector classifier fitted to a small dataset with 27 observations
- Observations 16, 18, 20, 19, 27 are on the margin
- Observations 4, 13, 15, 17 are on the wrong side of their respective margins
- Observations 21, 28, 25, 22, 24 are not only on the wrong side of their respective margins but also on the wrong side of the separating hyperplane

Classification with Non-Linear Decision Boundaries



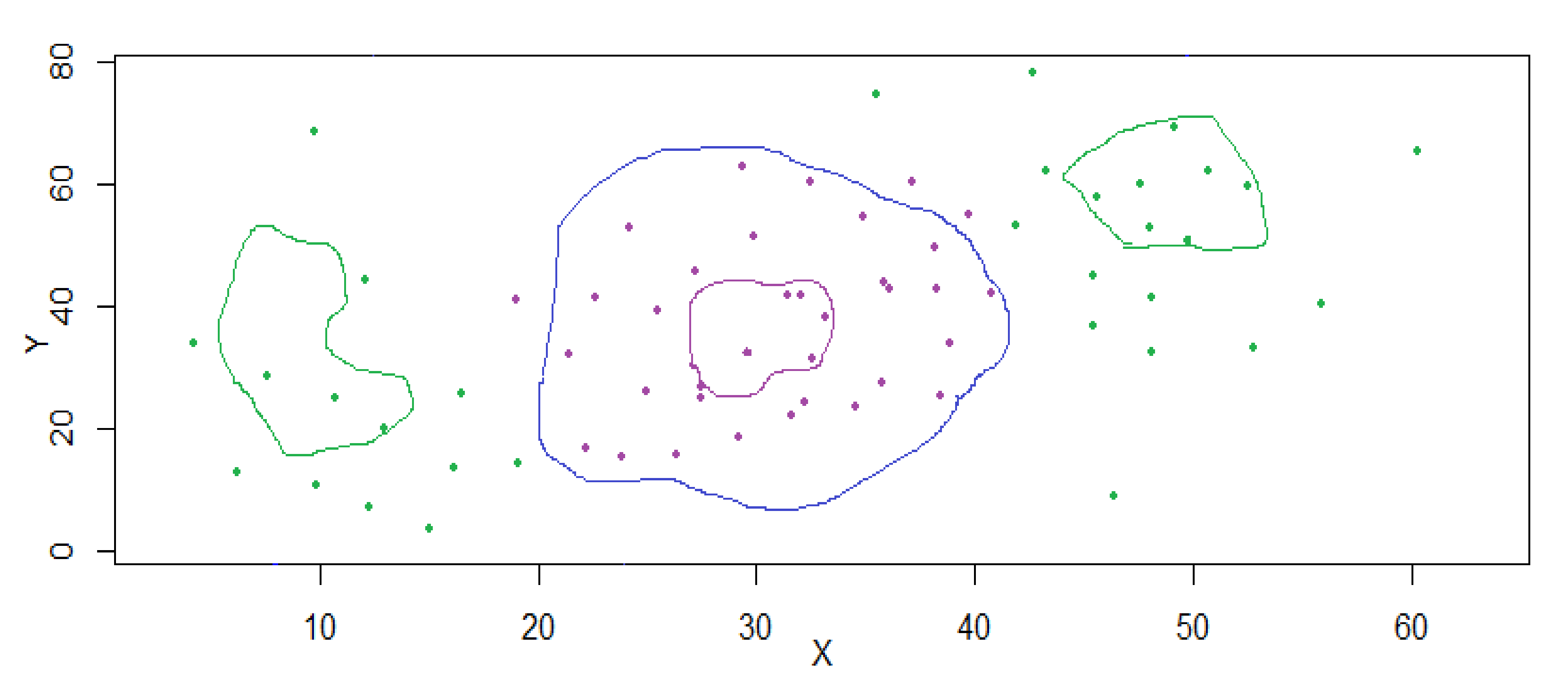
- When the class boundaries are non-linear, then the feature space (predictors) is enlarged with non-linear components in it.
- **Support Vector Machine** is an extension of support vector classifier that is constructed from enlarging feature space in a specific way using kernel functions

Possible Solutions



Polynomial Kernel

Possible Solutions



Radial Kernel

SVM – More than Two Classes

- There are two approaches most popular approaches for SVM with more than two classes:
  - One – Versus – One Classification
  - One – Versus – All Classification

One – Versus – One Classification

- Suppose there are K (K>2) classes for a SVM problem
- This approach considers SVMs comparing a pair of classes with each combination $\binom{K}{2}$
- A test observation is classified by tallying the assignments to each of the K classes
- The final classification is decided by the majority assignments to a particular class

One – Versus – All Classification

- K (K>2) SVMs are fitted each time comparing one of the K classes to the remaining K-1 classes
- A test observation is assigned to that class out of K classes for which function of the estimated parameters is highest.

SVM in Python

- Classification has been implemented in Python using function **SVC**, NuSVC and LinearSVC from package sklearn.svm
- There are three different implementations of Support Vector Regression: **SVR**, NuSVR and LinearSVR from package sklearn.svm
- We will cover SVC and SVR

Syntax :

sklearn.svm.SVC(C=1.0, kernel='rbf', degree=3, gamma='auto',..)

sklearn.svm.SVR(C=1.0, kernel='rbf', degree=3, gamma='auto',epsilon=0.1,...)

Example: Riding Mowers

- A riding-mower manufacturer **MOW-EASE** took part in a Industrial Exhibition in which it got an opportunity to show a demo of its product to 180 different audience.
- The land owned by each of the audience and their approximate income have been recorded in the file RidingMowers.csv

Example: Riding Mowers

- The Data contains two predictors Area Owned (Lot_Size) and Income with response variable as “Bought” and “Not Bought” values

	Income	Lot_Size	Response
1	34	26	Not Bought
2	34	40	Not Bought
3	34	46	Not Bought
4	34	48	Not Bought
5	34	53	Not Bought
6	34	58	Not Bought
7	34	59	Not Bought
8	34	63	Not Bought
9	34	64	Not Bought
10	34	66	Bought
11	35	41	Not Bought

Cost Parameter

- The C parameter tells the SVM optimization how much you want to avoid misclassifying each training record.
- For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly.
- Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.
- For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable.

Tuning Non-Linear

The *kernel function* can be any of the following:

- linear: $\langle x, x' \rangle$.
- polynomial: $(\gamma \langle x, x' \rangle + r)^d$. d is specified by keyword `degree`, r by `coef0`.
- rbf: $\exp(-\gamma \|x - x'\|^2)$. γ is specified by keyword `gamma`, must be greater than 0.
- sigmoid $(\tanh(\gamma \langle x, x' \rangle + r))$, where r is specified by `coef0`.

- For kernel = "polynomial", **`degree`** argument can be tried for various values, as degree being actually degree of polynomial, **`coef0`** is the intercept
- For kernel = "rbf", **`gamma`** argument can be tried for various values
- For kernel = "sigmoid", **`coef0`** can be tried