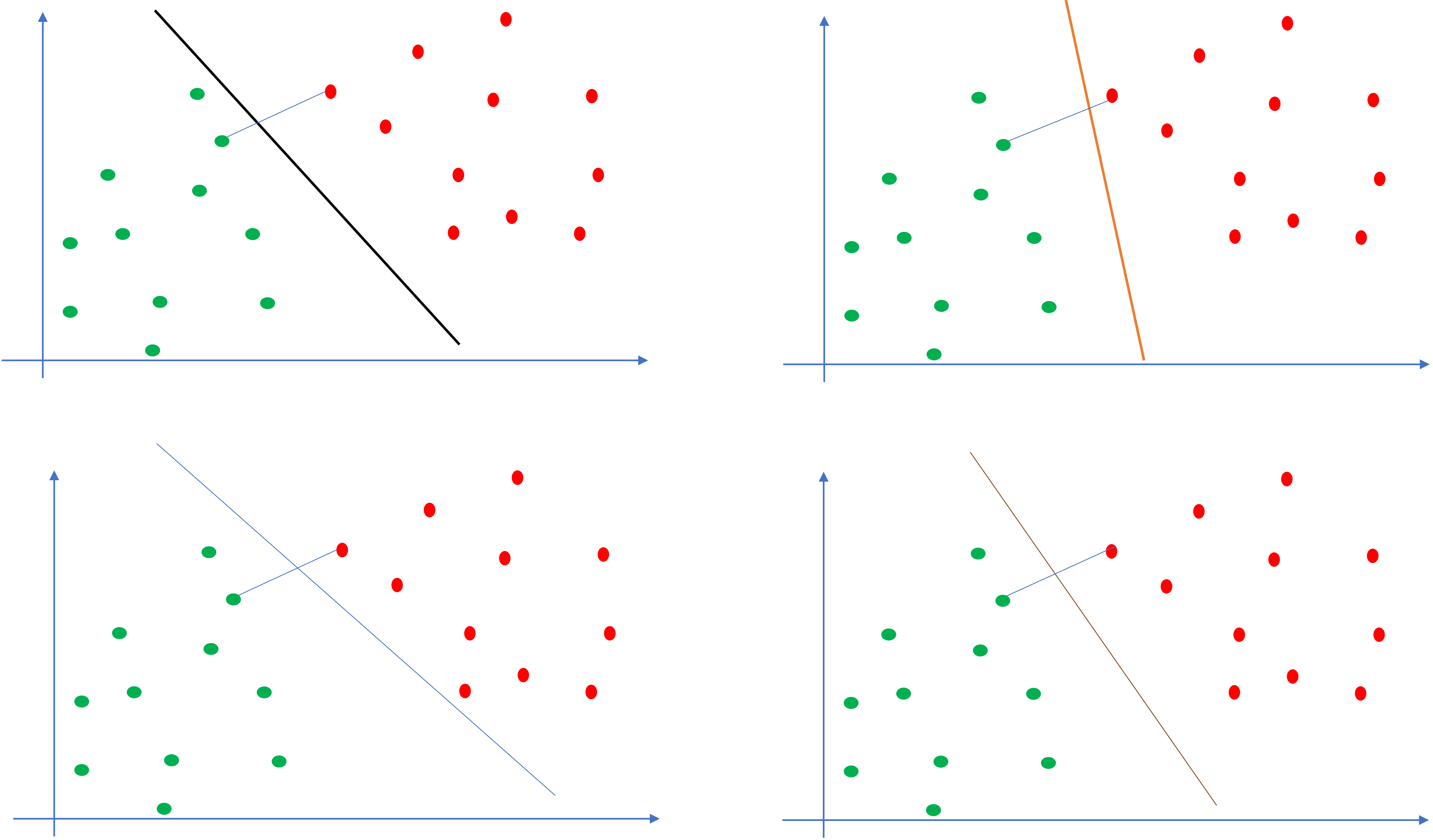
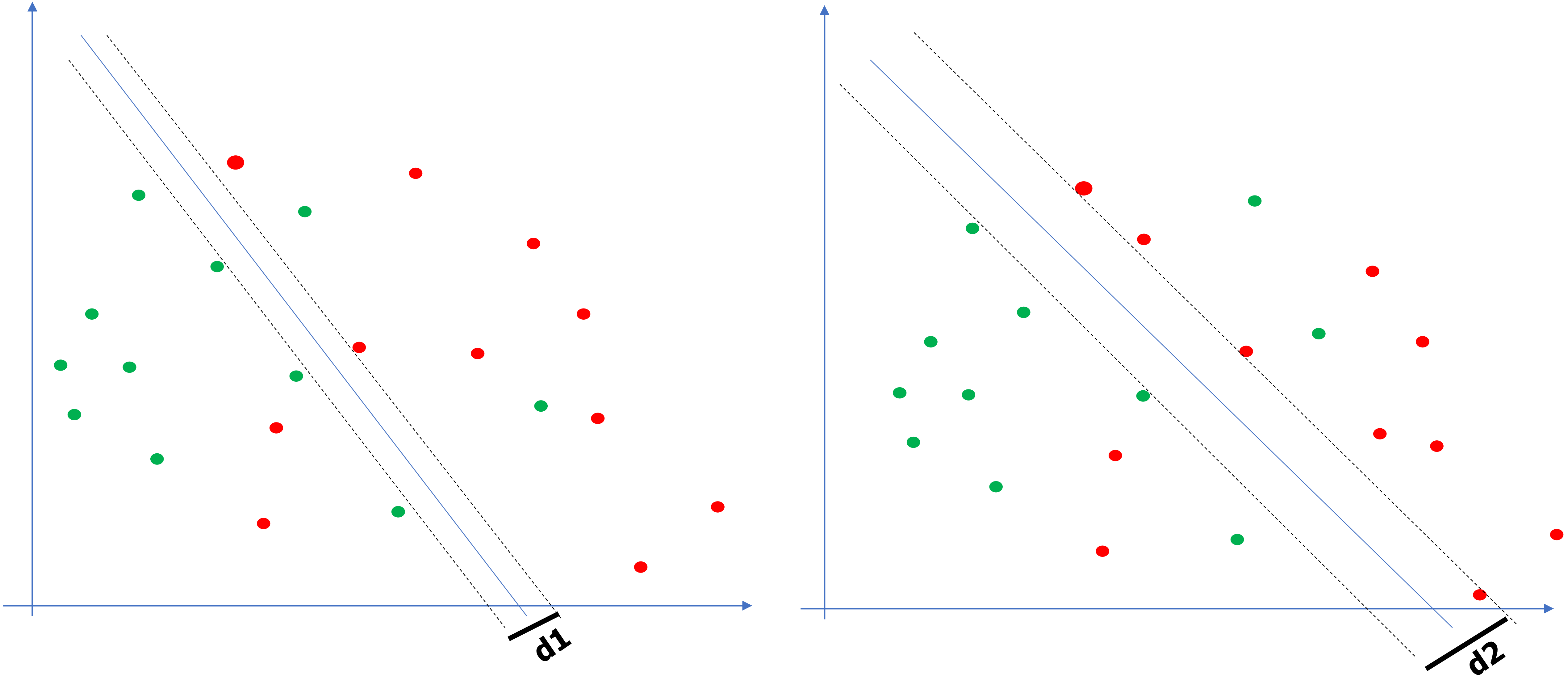
SVM
Support Vector Machines

SVM – Support Vector Machine

- Supervised machine learning algorithm
- Classification technique for **linear and non-linear** separable classes
- Alternate to Logistic regression ?
- Classification based on finding a **hyperplane** that maximises the margin between two classes
  - ➢ Mathematically speaking, SVM's are co-ordinates of the data/observations
- Mainly used in binary classification
- SVM algorithm has a feature to ignore outliers
- Complex algorithm, computing resources high, but SVM performs very well

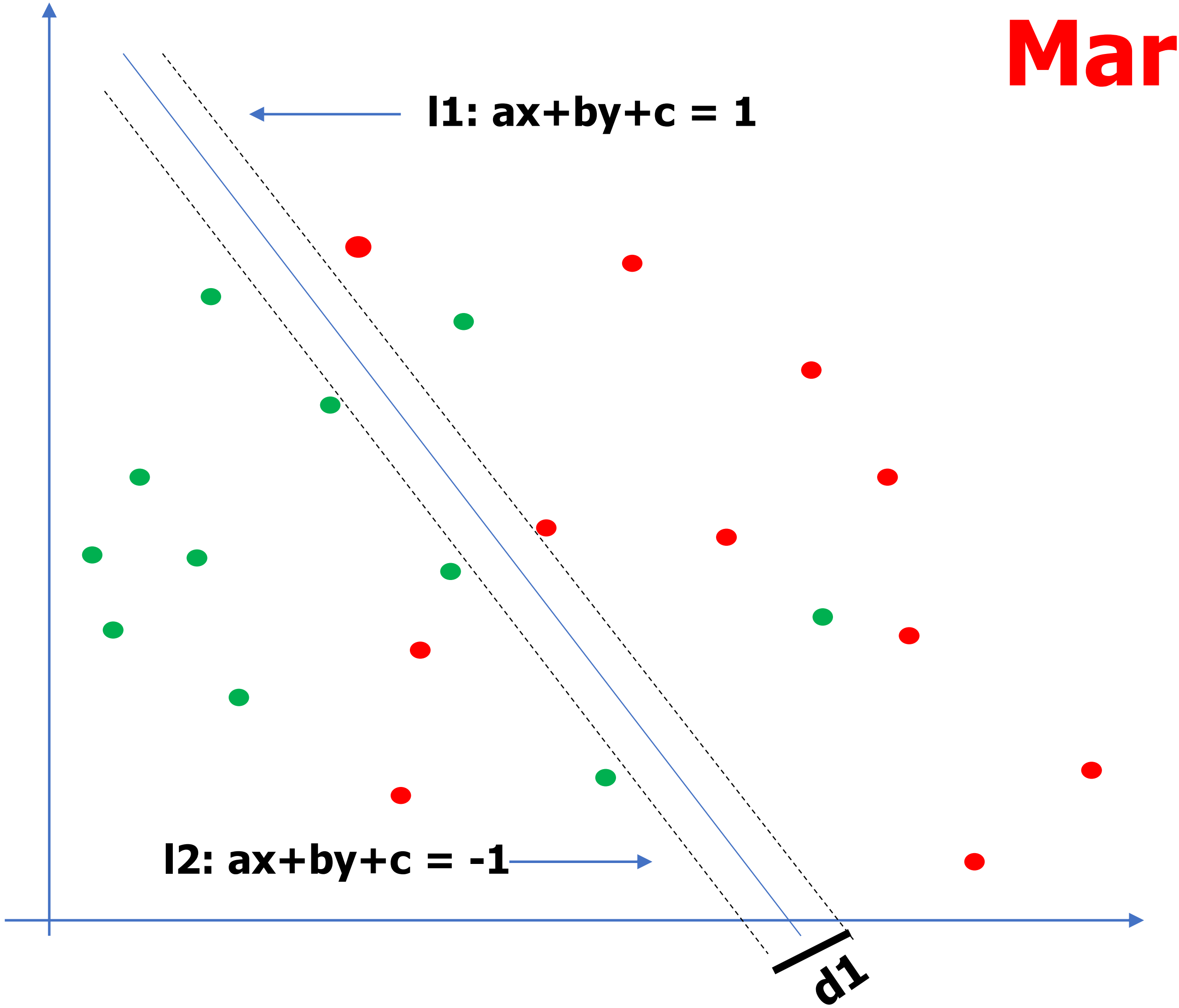




Which line is better ?

Margin Error defines the best line
Larger distance (d) means less margin error; Smaller distance (d) means larger margin error

Margin Error



Let the equations of the lines be
l1: ax+by+c = 1
l2: ax+by+c = -1

Margin (d1) = Perpendicular distance joining each line

Formula (d1) = 2 / $\sqrt{a^2 + b^2}$

Margin Error: $a^2 + b^2$

Goal of SVM is to minimise this error

Which line is better depends on
whether we need too many classification errors or
too many margin errors

Cost Parameter

- C controls the cost of misclassification on the training data
- (C * Classification Error) + (Margin Error)
- Value of C
  - ✓ **Small C**
    - Cost of misclassification low (“too strict”)
    - Large Margin Error
  - ✓ **Large C**
    - Cost of misclassification high and potentially overfit (“too loose”)
    - Low Margin Error
- The goal is to find the balance between "not too strict" and "not too loose“
- Cross-validation and resampling are good ways to finding the best C

- The goal of SVM is to find a hyperplane that would leave the widest possible "cushion" between input points from two classes. There is a trade-off between "narrow cushion, little / no mistakes" and "wide cushion, quite a few mistakes".

Find the best Kernel and other parameters

Cost

- Known as the Penalty parameter **(C)**
- Controls the cost of misclassification on the training data
- High C → more data points chosen as support vectors
  - ➢ High variance : Low Bias → Overfit
- Low C → less data points chosen as support vectors
  - ➢ Low variance : High Bias → Underfit

Gamma

- Influence of data points on the decision boundary
- Shape of the decision boundary line depends on gamma
  - ➢ High Gamma → decision boundary depends on data points near the decision boundary
  - ➢ Low Gamma → decision boundary depends on far away points

- The goal of SVM is to find a hyperplane that would leave the widest possible "cushion" between input points from two classes.

- There is a trade-off between "narrow cushion, little / no mistakes" and "wide cushion, quite a few mistakes".

Find the best Kernel and other parameters

Kernel

- Kernels are mathematical functions
- Measures the similarity between 2 data points
- Sometimes, it is difficult to draw decision boundary

- This kernel technique is black-box

Kernel types

- RBF (Radial Basis Kernel Function) (observations > features)
- Linear Kernel (features > observations)

Non-Linear classification / Kernel Trick

- Uses “kernel” technique to convert non-linear classes to linear classes to fit multi-classes
  - Quite efficient in multi-class prediction

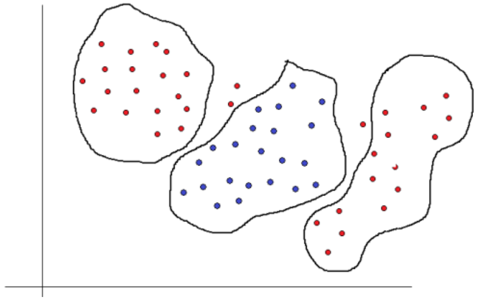
- Uses higher dimension feature space for calculation (i.e. converting non-linear separable classes to separable classes)

- SVM is popular as it works efficiently in large datasets having multi-classes

- Algorithm to arrive at an optimum hyperplane can be computationally expensive and time consuming

- More features and more observations complicate the algorithm

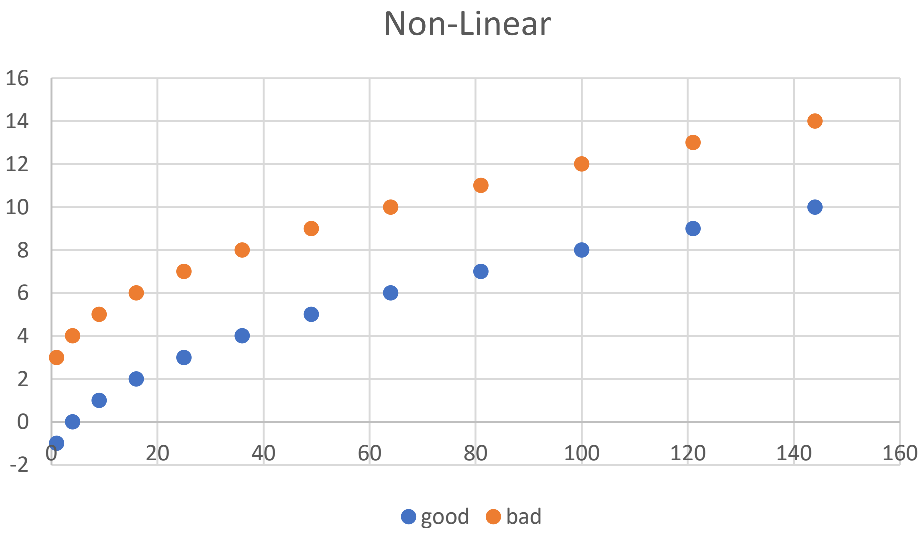
- Choice of Kernel for non-linear datasets
  - A big challenge
  - Black-box performance
  - Uses complex data transformation techniques

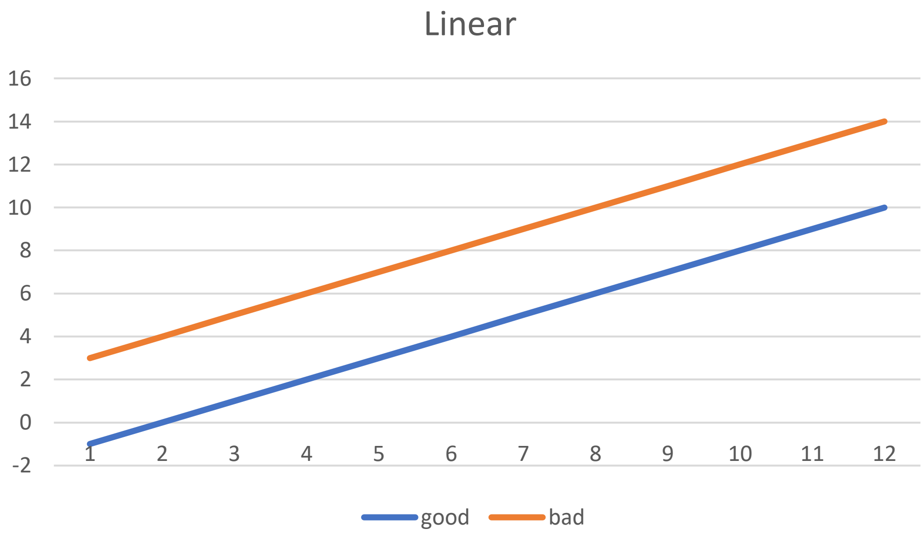


x	good	bad
1	-1	3
4	0	4
9	1	5
16	2	6
25	3	7
36	4	8
49	5	9
64	6	10
81	7	11
100	8	12
121	9	13
144	10	14

Square root of x

x	good	bad
1	-1	3
2	0	4
3	1	5
4	2	6
5	3	7
6	4	8
7	5	9
8	6	10
9	7	11
10	8	12
11	9	13
12	10	14


Non-Linear
● good
● bad


Linear
good
bad