Support Vector Machines

Contents

Basics

- In many situations, we want our machine learning algorithm to predict one of a number of (discrete) outcomes.
- **Example:**
  - An email client sorts mail into personal mail and spam mail, which has two outcomes.

Contd…

- Here, we consider a classifier that output binary values, i.e., there are only two possible outcomes.
- This machine learning task is called **binary classification**.

Contd…

- For binary classification, the set of possible values that the label/output can attain is binary, and for this module, we denote them by {+1, –1}.
- In particular, we consider classifiers of the form

$$f : \mathbb{R}^D \rightarrow \{+1, -1\}$$

- **Examples:**
  - In a cancer detection task, a patient with cancer is often labeled as +1 (or, Positive) and a patient without cancer is often labeled as –1 (or, Negative).
  - In COVID-19 testing, a patient infected with virus is labeled as +1 (or, Positive) and a patient without infection is labeled as –1 (or, Negative).

Contd…

- The approach that solves the problem of binary classification that is presented here is known as **Support Vector Machine (SVM)**.

Contd…

- We are discussing an example first, to understand the concepts and terminology behind the support vector machines to understand its theory better.

Example

- Suppose, we want to develop some criteria for determining the weather conditions under which tennis can be played.
- To simplify the matters, it has been decided to use the measures of **Temperature** and **Humidity** as the critical parameters for the investigation.
- We have some data as given in the following table.

Contd…

Temperature	Humidity	Play?
85	85	No
60	70	Yes
80	90	No
72	95	No
68	80	Yes
74	73	Yes
69	70	Yes
75	85	No
83	78	No

Contd…

- We are required to develop a criteria to know **whether one would be playing tennis on a future date** if,
  - we know the values of the temperature and humidity of that date in advance.

Contd…

1. **Two – Class Data Set**

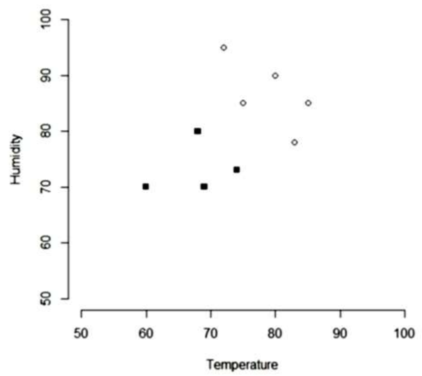
- This is our first observation.
- In the given table, the data are classified based on the values of the variable "Play".
- This variable "Play" has only two <u>values</u> or <u>labels</u>, namely "Yes" and "No".
- When there are only two class labels the data is said to be a "two-class data set".
- So the data in given table is a two-class data set.

Contd…

2. **Scatter plot of the data**

- We prepare a scatter plot.
- Temperature along the X – axis.
- Humidity along the Y – axis.
- The points that correspond to the decision "Yes" has been plotted as filled squares (■).
- The points that correspond to the decision "No" has been plotted as hollow circles (◦).

Contd…



Contd…

3. A separating line

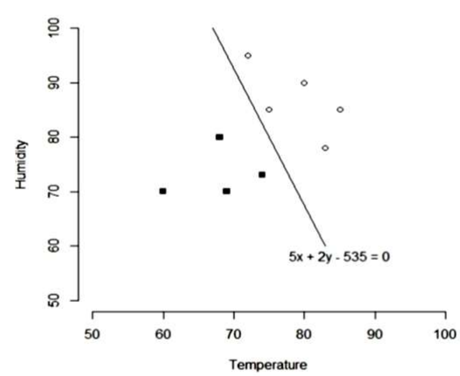
- If we examine the plot, we can see that, we can draw a straight line in the plane.
- This line can separate the two types of points in the sense that
  - all points plotted as filled squares (■) are on one side of the line and,
  - all points marked as hollow circles (∘) are on the other side of the line.
- Such a line is called a "separating line" for the data.

Contd…

- The following figure shows ***a*** **separating line** for the data in the given table.
- The equation of the separating line shown in the figure is

$5x + 2y - 535 = 0$.

Contd…



Contd…

- The **straight line** has the following property:
  - If the data point with values *(x',y')* has the value "Yes" for "Play" variable [filled squares (■)] then

    $5x' + 2y' - 535 < 0$

  - If the data point with values *(x',y')* has the value "No" for "Play" variable [hollow circles (◦)] then

    $5x' + 2y' - 535 > 0$

- **Note:** *at higher temperature and greater humidity values, don't venture out.*

Contd…

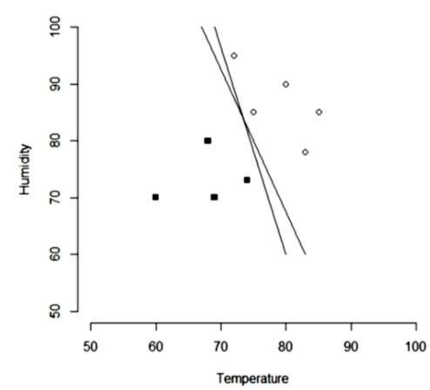
- If such a separating line exists for a given data then the data is said to be **linearly separable**.
- Thus, the data in the given table is linearly separable.
- However, not all data are linearly separable.

Contd…

4. Several separating lines

- Now, if future Temperature and Humidity values are known to us, we can determine whether to go out for playing Tennis or not.
- But **there are several separating lines** and the problem of determining **which one to choose** arises.

Contd…



SAINTGITS
LEARN.GROW.EXCEL

Contd…

5. **Margin of a separating line**

- To choose the **"best"** separating line, we introduce the concept of the **margin** of a separating line.
- We will select the best separating line as the one – **that leads to the greatest separation** between the filled squares (■) and hollow circles (◦).
  - **Greatest Advantage:** Such a line can generalize the best to the unseen future data.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

SAINTGITS
LEARN.GROW.EXCEL

Contd…

6. **Maximum margin separating line**

- As noted already, the **best separating line** is the one that gives us the **greatest separation**.
- **In other words,** the **best separating line** **is the one with the maximum margin**.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

SAINTGITS
LEARN.GROW.EXCEL

Contd…

- The separating line with the maximum margin **is called**
  - the **maximum margin line** or,
  - the **optimal separating line**.
- This line is also called the **Support Vector Machine** of the given table of data.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

SAINTGITS
LEARN.GROW.EXCEL

Contd…

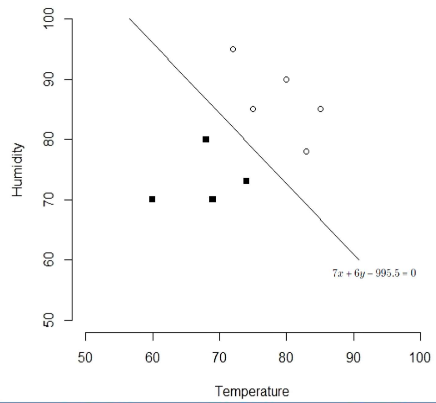
- Unfortunately, finding the equation of the maximum margin line is a non-trivial *(= important)* task.
- The maximum margin line of the given data is shown as

$$7x + 6y - 995.5 = 0.$$

- The following figure shows this maximum margin.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd...

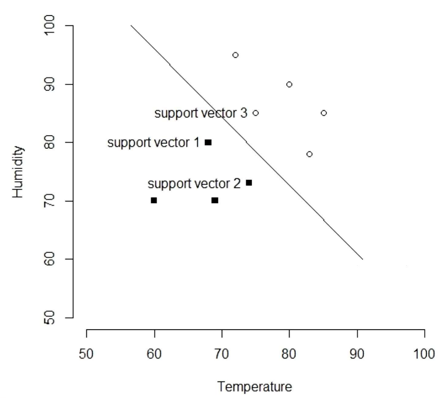


Contd...

7. Support Vectors

- The data points which are closest to the maximum margin line are called the **support vectors**.
- The support vectors for the given data is shown in the figure that follows.

Contd...



Contd...

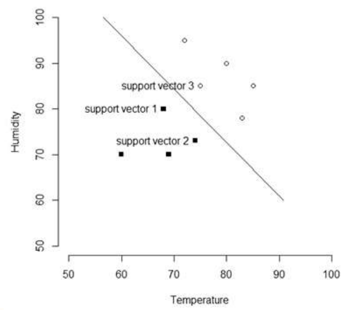
8. Required Criterion

- According to the theory of SVMs, the equation of the maximum margin line is used to devise a criterion for taking a decision on – whether to play tennis or not.
- Let **x'** and **y'** be the values of Temperature and Humidity on a given day.
- Then,
  - Decision as to play tennis on that day is "Yes", if : $7x' + 6y' - 995.5 < 0$
  - Decision as to play tennis on that day is "No", if : $7x' + 6y' - 995.5 > 0$

Contd…

9. “Street” of maximum width separating the two sets of points

- Refer the following plot.



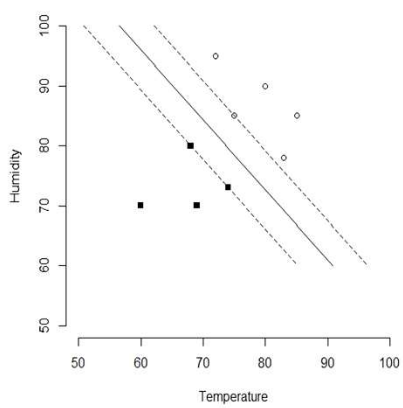
LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- We may draw a line through the support vectors 1 and 2 that is parallel to the maximum margin line.
- Similarly, we may draw a line through the support vector 3 also that is parallel to the maximum margin line.
- The two lines are shown as dashed lines in the plot that follows.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…



LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- The region between these two dashed lines can be thought of as a “street” or “road” of maximum width that separates the “Yes” data points and the “No” data points.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

10. Final Comments

- Any line given an equation of the form ax + by + c = 0 separates the coordinate plane into to halves.
- One half consists of all points for which ax+by+c>0.
- The other half contains all points for which ax+by+c<0.
- Which half contains which … is determined by the signs of the coefficients a, b and c.

Hyperplanes

- Hyperplanes **are subsets** of finite dimensional vector spaces which are similar to
  - **straight lines** ***in 2D Spaces*** and,
  - **planes** ***in 3D Spaces***.

Contd…

- **Definition:**
- Consider the n – dimensional vector space $\boldsymbol{R^n}$.
- The set of all vectors $\vec{\mathbf{x}} = (x_1, x_2, x_3, \ldots, x_n)$ in $\boldsymbol{R^n}$ whose components **satisfy** an equation of the form

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$$

where, $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n$ are scalars, is called a **hyperplane** in the vector space $\boldsymbol{R^n}$.

Contd…

- **Remark #1:**
- Let $\vec{\mathbf{x}} = (x_1, x_2, x_3, \ldots, x_n)$ and,
- $\vec{\boldsymbol{\alpha}} = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n)$
- Then, using the notation of inner product, the equation

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$$

can be written as

$$\boldsymbol{\alpha_0} + \vec{\boldsymbol{\alpha}} \cdot \vec{\mathbf{x}} = 0$$

Contd…

- **Remark #2:**
- The hyperplane defined by $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$ divides the space $\boldsymbol{R^n}$ into two disjoint halves.
- One of the halves consists of all vectors $\vec{\mathbf{x}}$ for which

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n > 0$$

- The other half consists of all vectors $\vec{\mathbf{x}}$ for which

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n < 0$$

Hyperplanes in 2D Space – *STRAIGHT LINES*

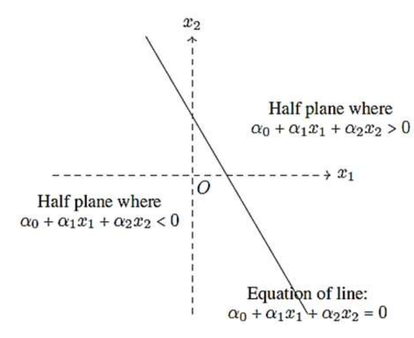
- Consider the 2-D vector space $\boldsymbol{R^2}$.
- Vectors in this space are ordered pairs of the form (x_1,x_2).
- Choosing the appropriate coordinate axes, such a vector can be represented by a point with $\vec{\mathbf{x}} = (x_1,x_2)$ in the space.
- In this space, we know that $||\vec{\mathbf{x}}||$ is the distance of the point (x_1,x_2) from the origin, and is computed as $\sqrt{x_1{}^2 + x_2{}^2}$.
- Also, the angle between the vectors $\vec{\mathbf{x}} = (x_1,x_2)$ and $\vec{\mathbf{y}} = (y_1,y_2)$ is the angle between the lines joining the origin to the points (x_1,x_2) and (y_1,y_2).

Contd…

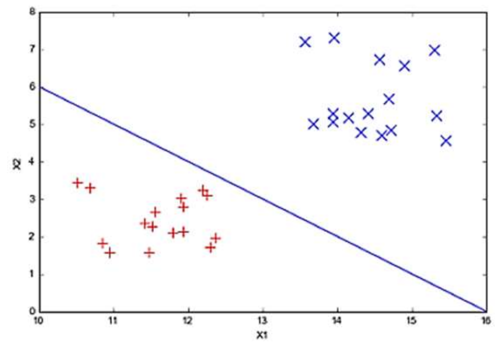
- Consider the set of all vectors $\vec{\mathbf{x}} = (x_1,x_2)$ in $\boldsymbol{R^2}$ which satisfy the equation $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 = 0$; where, $\alpha_0, \alpha_1, \alpha_2$ are scalars.
- From elementary analytical geometry, we can see that this represents **a straight line** in the 2D space.

Contd…

- This straight line $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 = 0$ divides the plane into two halves.



Contd…



Contd…

- It can be proved that one of the two halves consists of all points for which

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 > 0$$

- and, the other half consists of all points for which

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 < 0$$

Hyperplanes in 3D Space – *PLANES*

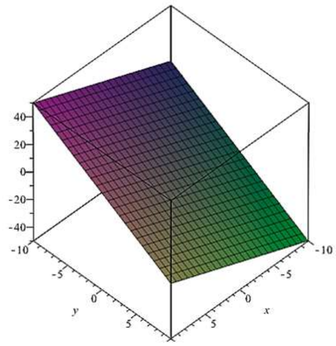
- Consider the 3-D vector space $\boldsymbol{R^3}$.
- Vectors in this space are ordered triplets of the form (x_1,x_2,x_3).
- Choosing the appropriate coordinate axes, such a vector can be represented by a point with $\vec{\mathbf{x}} = (x_1,x_2,x_3)$ in the space.
- In this space, we know that $||\vec{\mathbf{x}}||$ is the distance of the point (x_1,x_2,x_3) from the origin, and is computed as $\sqrt{{x_1}^2 + {x_2}^2 + {x_3}^2}$.
- Also, the angle between the vectors $\vec{\mathbf{x}} = (x_1,x_2,x_3)$ and $\vec{\mathbf{y}} = (y_1,y_2,y_3)$ is the angle between the lines joining the origin to the points (x_1,x_2,x_3) and (y_1,y_2,y_3).

Contd…

- Consider the set of all vectors $\vec{\mathbf{x}} = (x_1,x_2,x_3)$ in $\boldsymbol{R^3}$ which satisfy the equation $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 = 0$; where, $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ are scalars.
- From elementary analytical geometry, we can see that this represents **a plane** in the 3D space.

Contd…

- This plane $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 = 0$ divides the plane into two halves.



Contd…

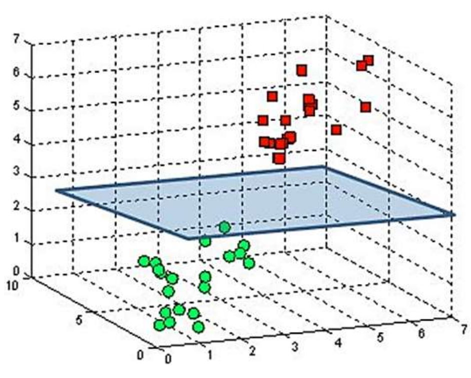
- It can be proved that one of the two halves consists of all points for which

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 > 0$$

- and, the other half consists of all points for which

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 < 0$$

Contd…

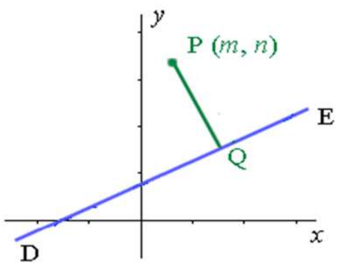


Hyperplanes *in nD Space*

- **By analogy** with the space in 2D Space (***R***^{2}) and 3D Space (***R***^{3}), we **imagine that** there is a geometrical world for any vector with *n* components.
- That is, the **points in this world** can be represented by ordered *n* tuples of the form $\vec{\mathbf{x}} = (x_1, x_2, x_3, \ldots, x_n)$.
- This world is represented as ***R***^{n}.
- The hyperplanes in ***R***^{n} are defined **by analogy** with the geometrical straight lines and planes.

Distance of a hyperplane from a point

- In 2D – Space, the perpendicular distance PQ of a point *P*(*m*,*n*) from a line $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 = 0$ is given by

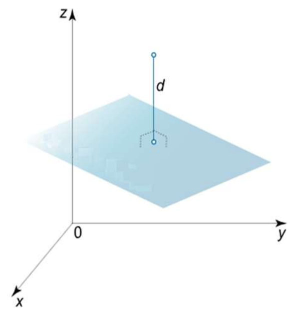


$$PQ = \frac{|\alpha_0 + \alpha_1 m + \alpha_2 n|}{\sqrt{\alpha_1^2 + \alpha_2^2}}$$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- In 3D – Space, the perpendicular distance *d* of a point *P*(*a*,*b*,*c*) from a plane $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 = 0$ is given by



$$d = \frac{|\alpha_0 + \alpha_1 a + \alpha_2 b + \alpha_3 c|}{\sqrt{\alpha_1^2 + \alpha_2^2 + \alpha_3^2}}$$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Support Vector Classifier

- In machine learning problems, the **variable being predicted** is called
  - the output variable
  - the target variable
  - the dependent variable *or*
  - the response.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- **As already noted in the beginning**, a **two-class data set** is a data set in which the target variable takes only one of two possible values.
- Such values can be of the form
  - {"yes", "no"}
  - {"TRUE","FALSE"}
  - {0,1}
  - {+1,-1} *etc.*
- **Note:** *Target variables having more than two possible values are called **multi-class dataset**.*

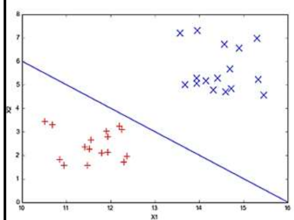
LEARN . GROW . EXCEL GOPEEKRISHNAN R

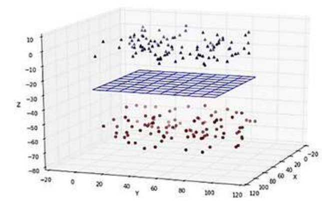
Contd…

- The methods of support vector machines were originally developed for classification problems involving two-class data sets.

Contd…

- A Support Vector Machine (SVM) can be thought of **as a surface that creates a boundary** between points of data plotted in multidimensional space that represent **examples and their feature values.**





Contd…

- This surface is often referred to as **hyperplane**.
- **Its goal is** <u>to divide the data points into somewhat fairly homogeneous partitions</u>.
- It combines aspects of both nearest neighbor methods and regression methods.

Contd…

- SVMs **can be used** for <u>both classification and numeric prediction</u>.
- Some applications of SVMs include
  - Classification of microarray gene expression to identify cancer or other genetic diseases.
  - Text categorization.
  - Document categorization.
  - Detection of events like combustion engine failure, security breaches or earthquakes.

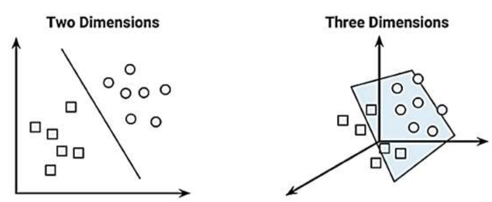
Contd…

- SVMs are easily understood, when used for binary classification (⟹ classification involving two classes).

CLASSIFICATION WITH HYPERPLANES

Classification with hyperplanes

- As noted previously, SVMs use a boundary called a hyperplane to partition data into groups of similar class values.
- The following figures show two hyperplanes in 2D-Space and 3D-Space that separate groups of circles and squares.



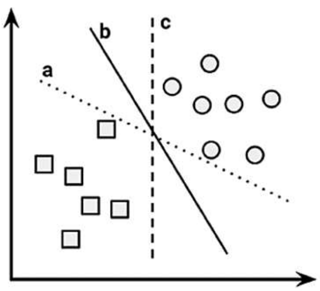
Contd…

- It is very clear that, in two-dimensional space, **the task** of SVM algorithm **is to find <u>a line</u>** (i.e., hyperplane) that separates the two classes.

SAINTGITS

Contd…

- But, as already understood, there can be **several separating straight lines** (i.e., hyperplanes) like ***a***, ***b*** and ***c*** as shown below.



SAINTGITS

Contd…

- Since there are several separating lines like ***a, b,*** or ***c*** the problem of determining <u>which one to choose</u> arises.

SAINTGITS

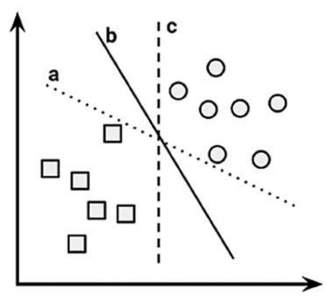
Contd…

- The answer to that question **involves** ***a search for the*** **Maximum Margin Hyperplane** (**MMH**) that creates the greatest separation between the two classes.

SAINTGITS

Contd…

- Though all the three lines ***a, b*** and ***c*** separates the circles and squares, **it is likely that** <u>the line that leads to the greatest separation will generalize the best to the future data</u>.
- We can see that it is the line ***b***.

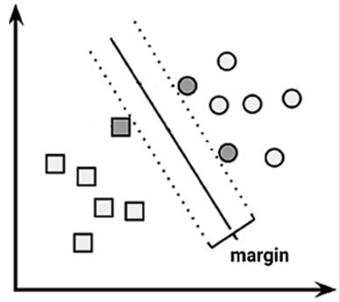


Contd…

- The maximum margin will improve the chance that, in spite of random noise, the points will remain on the correct side of the boundary.
- **Finding** the maximum margin line or the support vector machine is the important task here.

Contd…

- We can define the margin as follows…
- Consider the perpendicular distances from the training examples to the separating hyperplane H, and **consider the smallest such perpendicular distance**.
- The double of this smallest distance is called the **margin** of the hyperplane H.

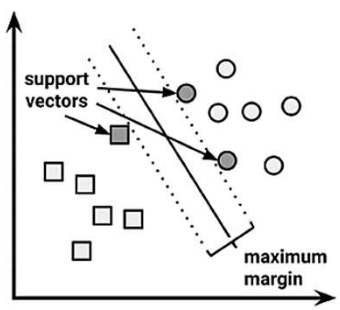


Contd…

- The hyperplane for which the margin is the largest is called the **maximum margin hyperplane** or the **support vector machine** for the given training data set.

Contd…

- Now, the data points **from each class** that are closest to the maximum margin line are called the **support vectors**.



SAINTGITS

Contd…

- **Each class** must have **at least one** support vector.
- That is, it is possible to have more than one support vector.

SAINTGITS

Contd…

- A key feature of support vector machines is that, **using the support vectors alone**, it is possible to define the MMH.

SAINTGITS

Contd…

- Classification with hyperplanes can be done when
  a) the given training data is **linearly separable**
  b) the given training data is **nonlinearly separable**

SAINTGITS

CLASSIFICATION WITH LINEARLY SEPARABLE DATA

Classification with *linearly separable data*

- **Linearly Separable Data:**
- Consider a two-class data set having n numeric features and two possible class labels +1 and -1.
- Let the vector $\vec{\mathbf{x}} = (x_1, x_2, x_3, \ldots, x_n)$ **represent** the values of the features of ***an*** example **of a data set.**
- We say that the data set is **linearly separable** if we can find a hyperplane in the n-dimensional vector space $\boldsymbol{R}^n$ say

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$$

having the two following properties:

Contd…

1. for each example $\vec{\mathbf{x}}$ with class label **+1**, we have

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n > 0$$

2. for each example $\vec{\mathbf{x}}$ with class label **-1**, we have

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n < 0$$

Contd…

- Two methods are there…**for finding maximum margins…**
  - **METHOD 1:** Using Convex Hulls
  - **METHOD 2:** Finding a set of parallel planes

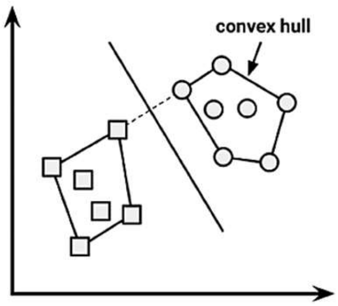
Contd…

METHOD 1: Using Convex Hulls

- It is easiest to find the maximum margin **under the assumption that** the classes are linearly separable.
- In this case, the MMH is **as far away as possible** from the outer boundaries of the two groups of data points.
- These outer boundaries are known as the **convex hull**.
- The MMH is then the **perpendicular bisector** of the shortest line between the two convex hulls.

SAINTGITS

Contd…



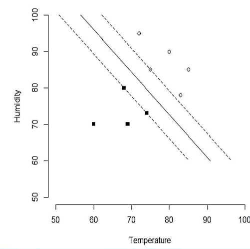
- **Perpendicular Bisector** – a line which is perpendicular to a given line segment and divides it into two equal halves is called a perpendicular bisector.

SAINTGITS

Contd…

METHOD 2: Finding a set of parallel planes

- Here we search through the space of every possible hyperplane.
- And, we **find a set of two parallel planes** that divide the points into homogeneous groups, yet, themselves are as far apart as possible.



SAINTGITS

Contd…

- Now, **how we are finding two such parallel planes?**
- Let us understand this through the equations of straight lines.
- Assume that we are given with a two-class training data set of N points of the form

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_3, y_3), \ldots, (\vec{x}_N, y_N)$$

- where, y_i's are either +1 or -1 (class labels); each $\vec{x}_i$ is an n – dimensional vector.
- We assume that the data is linearly separable.

SAINTGITS

Contd…

- Any hyperplane can be written as the set of the points $\vec{\mathbf{x}} = (x_1, x_2, x_3, \ldots, x_n)$ of the form

$$\vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{x}} + b = 0$$

Contd…

- We know our data is linearly separable.
- Thus, we can select two parallel hyperplanes (say, H_+ and H_-) that separate the two classes of data, **so that the distance between them is as large as possible.**
- The maximum margin hyperplane (***H***) is the hyperplane that lies halfway between these two.

Contd…

- The two hyperplanes can be described by equations of the following forms:

$$\vec{w} \cdot \vec{x} + b = +1$$
$$\vec{w} \cdot \vec{x} + b = -1$$

Contd…

- For any point on or above the hyperplane $\vec{w} \cdot \vec{x} + b = +1$, the class label is $+1$.
- This implies that,

$$\vec{w} \cdot \vec{x_i} + b \geq +1, \text{ if } y_i = +1$$

Contd…

- Similarly, for any point on or below the hyperplane $\vec{w} \cdot \vec{x} + b = -1$, the class label is -1.
- This implies that,

$$\vec{w} \cdot \vec{x_i} + b \leq -1, \text{ if } y_i = -1$$

Contd…

- The two conditions can now be combined into a single equation as

$$y_i(\vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{x}_i} + b) \geq +1 \; for\,all \; 1 \leq i \leq N$$

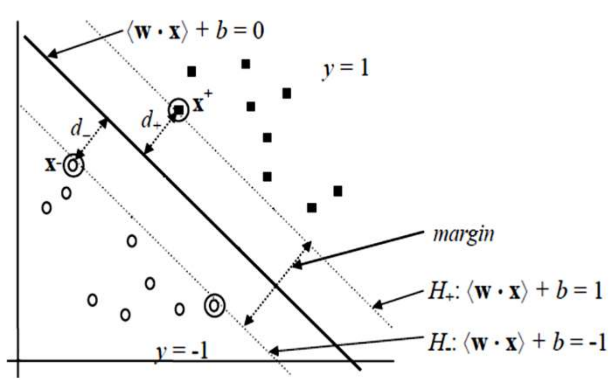
Contd…

- Now, the **distance between** the two hyperplanes ***H₊*** and ***H₋*** is calculated as

$$\frac{\mathbf{2}}{||\vec{\boldsymbol{w}}||}$$

- So, to maximize the distance between the hyperplanes $\boldsymbol{H_+}$ and $\boldsymbol{H_-}$ **we have to minimize** $||\vec{w}||$.

Contd…



Contd…

- Thus, when two parallel hyperplanes are found, the **maximum margin** can be described as $\frac{\mathbf{2}}{||\vec{\boldsymbol{w}}||}$.
- For computational simplicity, the **maximum margin** is described as

$$\frac{||\vec{\boldsymbol{w}}||^2}{\mathbf{2}}$$

SAINTGITS

Contd…

- Hence, we understood that, **there are two methods** for finding the maximum margins **when the data are linearly separable**.

SAINTGITS

Contd…

- Thus, the **support vector classifier problem** can be described as follows, when the data are linearly separable.
- Given a two-class linearly separable dataset of N points of the form

$$(\vec{\boldsymbol{x}}_1, y_1), (\vec{\boldsymbol{x}}_2, y_2), (\vec{\boldsymbol{x}}_3, y_3), \ldots, (\vec{\boldsymbol{x}}_N, y_N)$$

- where, y_i's are either +1 or -1 (class labels).
- We find a vector $\vec{\boldsymbol{w}}$ and a number b which

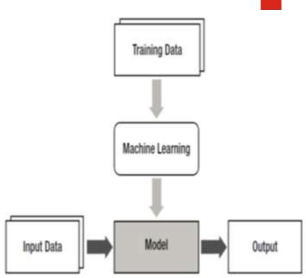
$$\textbf{minimize: } \frac{1}{2} ||\vec{w}||^2$$

$$\textbf{subject to: } y_i(\vec{\boldsymbol{w}} \cdot \vec{\boldsymbol{x}_i} + b) \geq +1 \ for\ all\ 1 \leq i \leq N$$

SAINTGITS

Contd…

- So, for solving this problem, there is an algorithm for constructing the model.
- This algorithm (*usually implemented as packages in R, Python etc.*) is not discussed here.



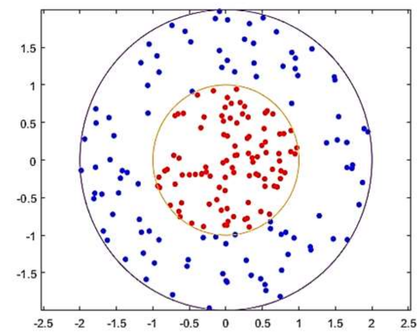
SAINTGITS

CLASSIFICATION WITH NON-LINEARLY SEPARABLE DATA

Classification with *nonlinearly separable data*

- The algorithm for finding the SVM classifier will give a solution only if the given two-class dataset is linearly separable.
- But, in real life problems, two-class datasets are only **rarely** linearly separable.
- SVMs address non-linearly separable cases by introducing two concepts:
  - **Soft Margin** and
  - **Kernel Tricks**

Contd…



Nonlinearly Separable Data

Soft Margin

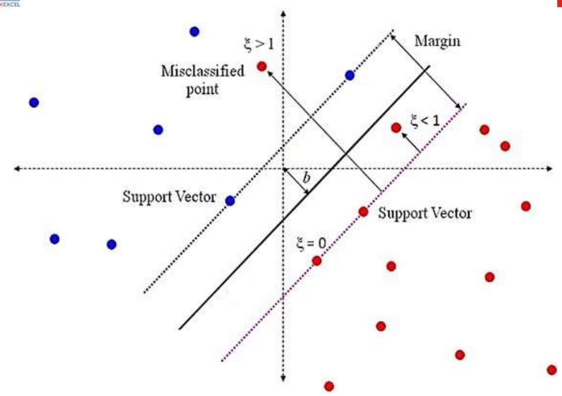
- Here, we try to find a line to separate, but tolerate one or few misclassified data points.
- For this, we introduce additional variables, ξ_i *(pronounced **Xi**)*, called **slack variables** which store **deviations from the margin.**

Contd…

- There **are two types of deviations:**
  1. An example that **lies on the correct side** of the hyperplane but **lies in the margin**, not sufficiently away from the hyperplane.
  2. An example that **lies on the wrong side** of the hyperplane and **is misclassified**.

SAINTGITS

Contd…



LEARN . GROW . EXCEL GOPEEKRISHNAN R

SAINTGITS

Contd…

- If $\xi_i = 0$, then $\vec{\mathbf{x}}$ is correctly classified.
- If $0 < \xi_i < 1$, then $\vec{\mathbf{x}}$ is correctly classified but is in the margin.
- If $\xi_i > 1$, then $\vec{\mathbf{x}}$ is misclassified.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

SAINTGITS

Contd…

- Now, the sum $\sum_{i=1}^{N} \xi_i$ is defined as the **soft error** and, this is added as a penalty to the function to be minimized.
- Also, a factor ***C*** is added to the soft error.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

SAINTGITS

Contd…

- With these modifications, the SVM problem can be reformulated as follows.
- Given a two-class linearly separable dataset of N points of the form

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_3, y_3), \ldots, (\vec{x}_N, y_N)$$

- where, y_i's are either +1 or -1 (class labels).
- We find a vectors $\vec{w}$ and $\vec{\xi}$ and a number b which

minimize: $\frac{1}{2} ||\vec{w}||^2$ + ***C***. $\sum_{i=1}^{N} \xi_i$

subject to: $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i \; for\, all \; 1 \leq i \leq N$

$\xi_i \; > 0 \; for\, all \; 1 \leq i \leq N$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- Thus, it is clear that a SVM can be trained using a slack variable even when some of the examples are misclassified.

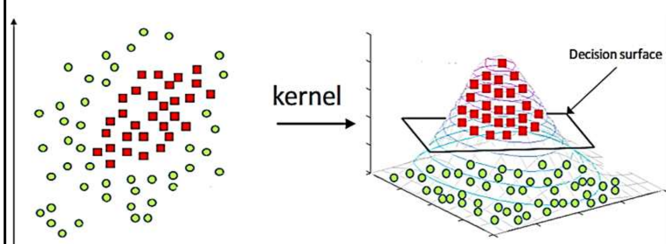
Kernel Methods

- We just have seen, a SVM can be trained using a slack variable even when some of the examples are misclassified.
- However, we have to understand that, this is not the only way to approach the problem of nonlinearity.

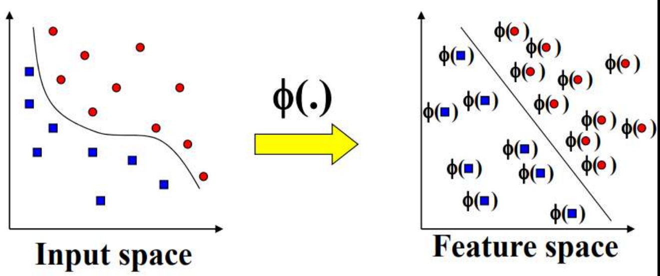
Contd…

- A **key feature of SVMs** is their ability to map the problem into a higher dimensional space using a process known as **kernel trick.**
- The idea is, mapping the non-linear separable data-set into a higher dimensional space **where we can find a hyperplane** that can separate the samples.
  - (this means… **in the original input space…data is non-linearly separable**; whereas, **in the higher dimensional feature space, data becomes linearly separable.**)

Contd…

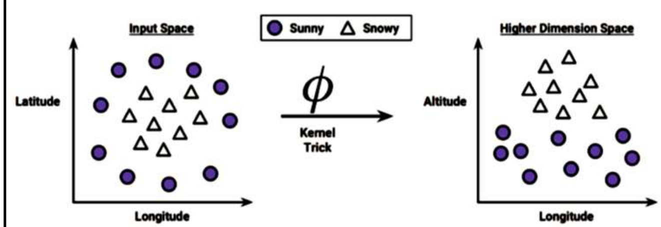


Contd…



Contd…

- **Example 1:**
- The scatterplot on the LHS depicts a non – linear relationship between a weather class and two features.
  - ❖ **Weather class** : **1.** Sunny **2.** Snowy
  - ❖ **Two features** : **1.** Latitude **2.** Longitude

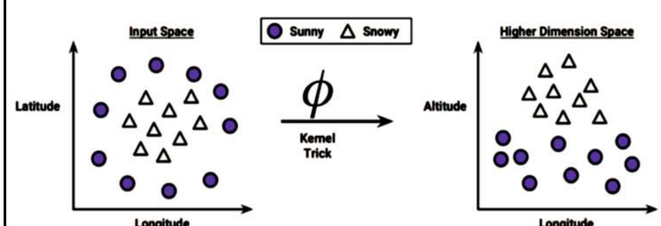


Contd…

- Data could had been generated from a set of weather reports…
- Some of the weather reports were obtained from stations **near <u>the top of a mountain</u>…**
- Some of the weather reports were obtained from stations **around <u>the base of a mountain</u>…**

Contd…

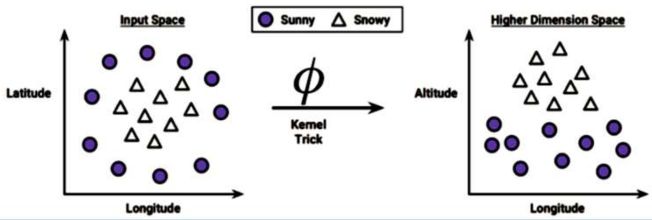
- The scatterplot on the RHS, depicts a **linear relationship** between the weather class and its features.
- This **linearity has been got** after <u>the kernel trick has been applied</u>.
- For this, data has been shown with **<u>a new dimension</u>**: **Altitude**.



SAINTGITS

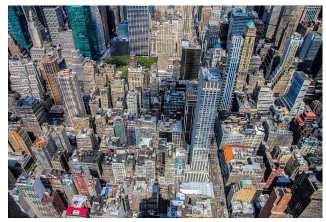
Contd…

- In the LHS plot, we are **viewing** the mountain **from a bird's eye view**.
- In the RHS plot, we are **viewing** the mountain **from a distance at the ground level**.
- Now it is clear…that…**snowy weather is found at higher altitudes…**



SAINTGITS

Contd…



Bird's Eye-view of New York City

Ground Level View of New York City

SAINTGITS

Contd…

- Thus it is clear that…when SVMs with non-linear relationships are added with additional dimensions (features), clear separations can be generated…

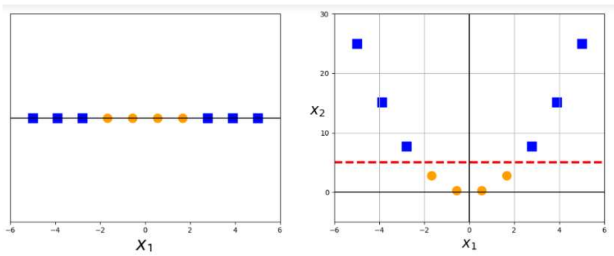
SAINTGITS

Contd…

- Thus, the **kernel trick involves the process of constructing new features** (Here, "Altitude") that express mathematical relationships between measured characteristics (Here, "Longitude" and "Latitude").

Contd…

- **Example 2:**
- A transformation from the original input feature space (1-D) into the higher dimensional feature space (2-D) is featured below.

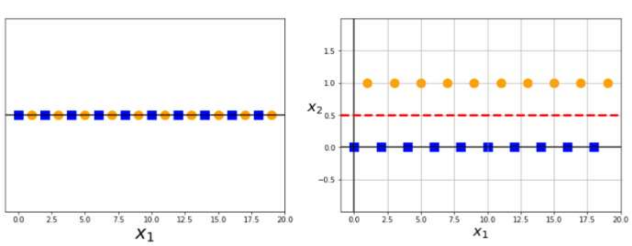


Contd…

- In 1-D, the data is not-linearly separable. But after applying the transformation $\emptyset(x) = x^2$, the data becomes linearly separable.

Contd…

- **Example 3:**
- A transformation from the original input feature space (1-D) into the higher dimensional feature space (2-D) is featured below.

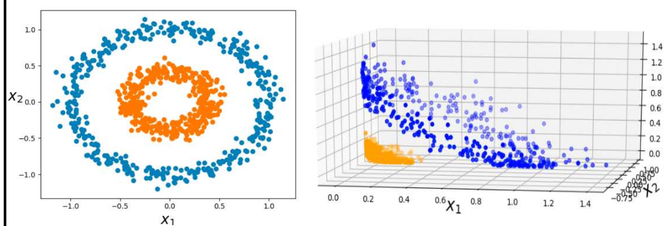


Contd…

- In 1-D, the data is not-linearly separable. But after applying the transformation $\emptyset(x) = x \bmod 2$, the data becomes linearly separable.

Contd…

- **Example 4:**
- A transformation from the original input feature space (2-D) into the higher dimensional feature space (3-D) is featured below.



Contd…

- In 2-D, the data is not-linearly separable. But after applying the transformation $\emptyset(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, the data becomes linearly separable.

Contd…

- It allows us to operate in the original input space without computing the coordinates of the data in a higher dimensional feature space.

Contd…

- In the context of SVMs, a kernel function is a function of the form $K(\vec{x}, \vec{y})$ where $\vec{x}$ and $\vec{y}$ are n-dimensional vectors.

Contd…

- **Definition:** Let $\vec{x}$ and $\vec{y}$ be arbitrary vectors in the n-dimensional vector space R^n. Let ϕ be a mapping function from R^n to some vector space. A function $K(\vec{x},\vec{y})$ is called a kernel function, if there is a function ϕ such that $K(\vec{x},\vec{y}) = \phi(\vec{x}) \cdot \phi(\vec{y})$.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- Commonly used kernel functions is listed below.

Sl.No.	Kernel Function Name	Kernel Function	Remarks
01.	Linear Kernel	$K(\vec{x},\vec{y}) = \vec{x} \cdot \vec{y}$	Linear Kernel does not transform the data at all; it is simply the dot product of the vectors $\vec{x}$ and $\vec{y}$.
02.	Homogeneous Polynomial Kernel	$K(\vec{x},\vec{y}) = (\vec{x} \cdot \vec{y})^d$	d is some positive integer.
03.	Non-homogeneous Polynomial Kernel	$K(\vec{x},\vec{y}) = (\vec{x} \cdot \vec{y} + \theta)^d$	d is some positive integer; θ is a real constant.
04.	Radial Basis Function (RBF) Kernel	$K(\vec{x},\vec{y}) = e^{-\frac{\lVert\vec{x}-\vec{y}\rVert^2}{2\sigma^2}}$	Also called ***Gaussian RBF Kernel function***.
05.	Sigmoid Kernel	$K(\vec{x},\vec{y}) = tan\,h(\alpha(\vec{x} \cdot \vec{y}) + c)$	α and c are the kernel parameters; Also called ***Hyperbolic Tangent Kernel function.***

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- **Which Kernel function to choose?**
- No specific rule…for the selection of the function…
- The fit depends on
  - *the concept to be learned…*
  - *the amount of training data…*
  - *the relationships among the features…*

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd…

- **Example:** Let $\vec{x} = (x_1, x_2) \in R^2$ and $\vec{y} = (y_1, y_2) \in R^2$. Also, let $K(\vec{x},\vec{y}) = (\vec{x} \cdot \vec{y})^2$. We show that $K(\vec{x},\vec{y})$ is a kernel function.

- **Given,**
  - $\vec{x} = (x_1, x_2) \in R^2$
  - $\vec{y} = (y_1, y_2) \in R^2$
  - $K(\vec{x},\vec{y}) = (\vec{x} \cdot \vec{y})^2$

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd...

- $\vec{x} \cdot \vec{y}$ = $(x_1, x_2) \cdot (y_1, y_2)$ = $(x_1y_1 + x_2y_2)$
- $(\vec{x} \cdot \vec{y})^2$ = $(x_1y_1 + x_2y_2)^2$ = $(x_1{}^2y_1{}^2 + 2x_1y_1x_2y_2 + x_2{}^2y_2{}^2)$
- i.e., $\boldsymbol{K}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$ = $(x_1{}^2y_1{}^2 + 2x_1y_1x_2y_2 + x_2{}^2y_2{}^2)$ *...... (1)*

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd...

- Let us define
  - $\boldsymbol{\phi}(\vec{\boldsymbol{x}})$ = $(x_1{}^2, \sqrt{2}\,x_1x_2, x_2{}^2) \in \boldsymbol{R^3}$
  - $\boldsymbol{\phi}(\vec{\boldsymbol{y}})$ = $(y_1{}^2, \sqrt{2}\,y_1y_2, y_2{}^2) \in \boldsymbol{R^3}$
- Now,
  - $\boldsymbol{\phi}(\vec{\boldsymbol{x}}) \cdot \boldsymbol{\phi}(\vec{\boldsymbol{y}})$ = $(x_1{}^2, \sqrt{2}\,x_1x_2, x_2{}^2) \cdot (y_1{}^2, \sqrt{2}\,y_1y_2, y_2{}^2)$
    = $(x_1{}^2y_1{}^2 + 2\,x_1x_2y_1y_2 + x_2{}^2y_2{}^2)$ *...... (2)*

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd...

- Equating *(1)* and *(2)*, it is clear that $\boldsymbol{K}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$ is a kernel function.

LEARN . GROW . EXCEL GOPEEKRISHNAN R

Contd...

- **Strengths and Weaknesses of Non-Linear Kernels**

STRENGTHS	WEAKNESSES
Can be used for both classification and numeric prediction.	Finding the best SVM model requires testing of various combinations of kernel functions and model parameters.
Not overly influenced by noisy data.	Can be slow to train, particularly when the input dataset has a larger number of features or examples.
Not very prone to overfitting.	Results in a complex black box model that is difficult to interpret.
There are a number of Standard SVM algorithms in existence. Therefore, SVM methods are more easier and popular than Neural network modeling.	
High Accuracy is there in results. Thus, gaining more popularity.	

LEARN . GROW . EXCEL GOPEEKRISHNAN R

References

1. Machine Learning with R, Second Edition, Brett Lantz, PACKT Publishing.