Find S, SVM, LR

Find S- Features

- Supervised Learning Algorithm
- Finds the most specific hypothesis that fits all the positive examples.
- Starts with the most specific hypothesis and generalizes this hypothesis each time it fails to classify an observed data
- It moves from the most specific hypothesis to the most general hypothesis.

Terminologies

- **?** indicates that any value is acceptable for the attribute.
- **Φ** indicates that no value is acceptable.
- The most **general hypothesis** is represented by: **{?, ?, ?, ?, ?, ?}**
- The most **specific hypothesis** is represented by: **{φ, φ, φ, φ, φ, φ}**

Find S Algorithm

1. Initilize h to the most specific hypothesis in H
2. For each positive training instance x
    For each attribute contraint a_i in h
        If the contraint a_i is satisfied by x
            then do nothing
        Else
            replace a_i in h by the next more general constraint that is satisfied by x
3. Output the hypothesis h

Example 1

- Apply Find S to predict if Mohan will be able to enjoy the sport in the given climatic conditions **<Sky-Sunny, AirTemp-Cold, Humidity-Normal, Wind-Strong, Water-Warm, Forecast-Same>**

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

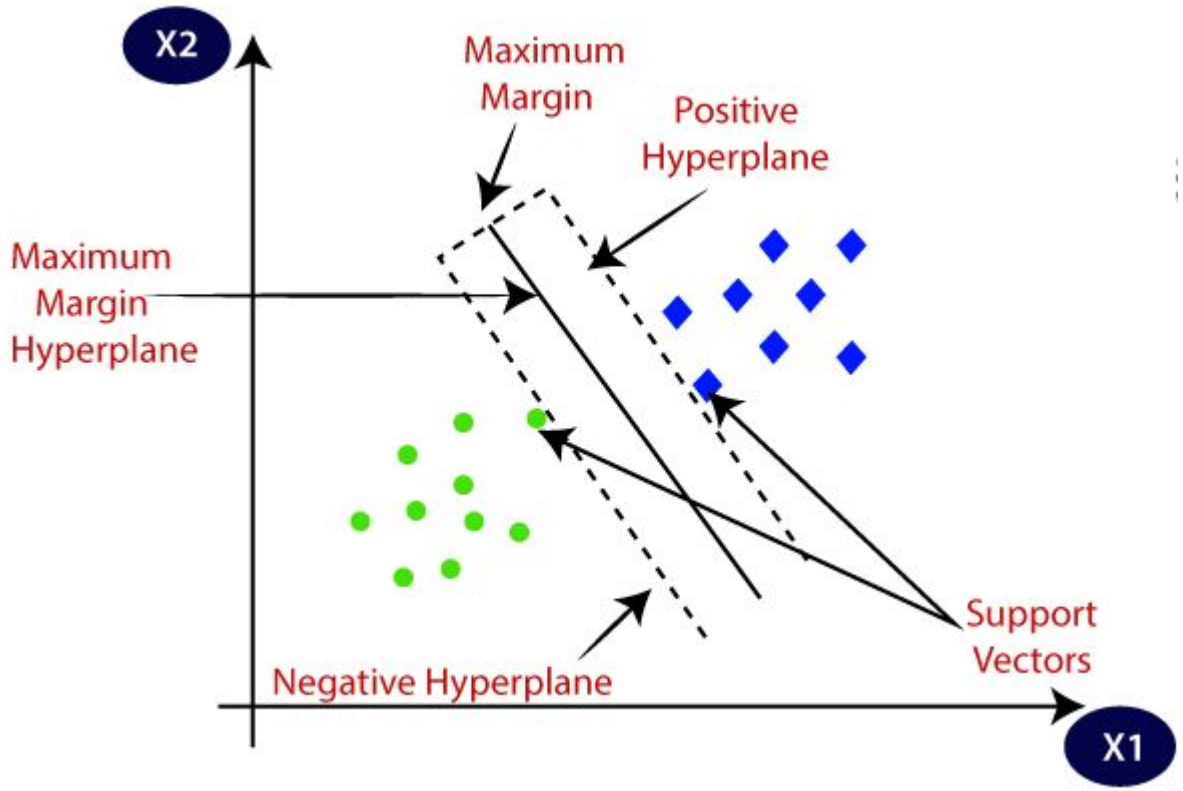
Example 2

- Find the most specific hypothesis which fit all the training examples

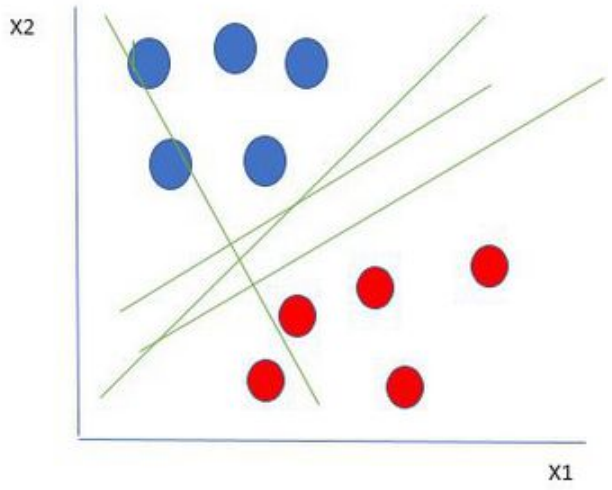
EXAMPLE	COLOR	TOUGHNESS	FUNGUS	APPEARANCE	POISONOUS
1.	GREEN	HARD	NO	WRINKELD	YES
2.	GREEN	HARD	YES	SMOOTH	NO
3.	BROWN	SOFT	NO	WRINKLED	NO
4.	ORANGE	HARD	NO	WRINKLED	YES
5.	GREEN	SOFT	YES	SMOOTH	YES
6.	GREEN	HARD	YES	WRINKLED	YES
7.	ORANGE	HARD	NO	WRINKLED	YES

Support Vector Machine (SVM)

- SVM is a supervised machine learning algorithm used for both classification and regression.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a **hyperplane**.
- The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible.
- The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane.


X2
Maximum
Margin
Positive
Hyperplane
Maximum
Margin
Hyperplane
Support
Vectors
Negative Hyperplane
X1

- Let’s consider two independent variables x_1, x_2 and one dependent variable which is either a blue circle or a red circle.

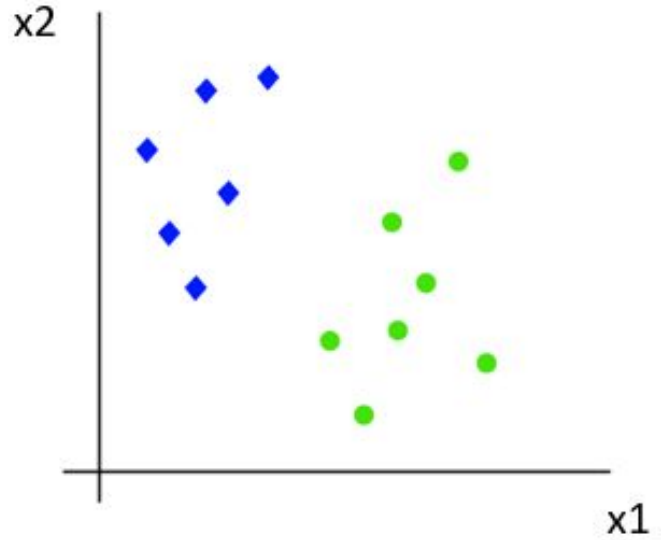


Linearly Separable Data points

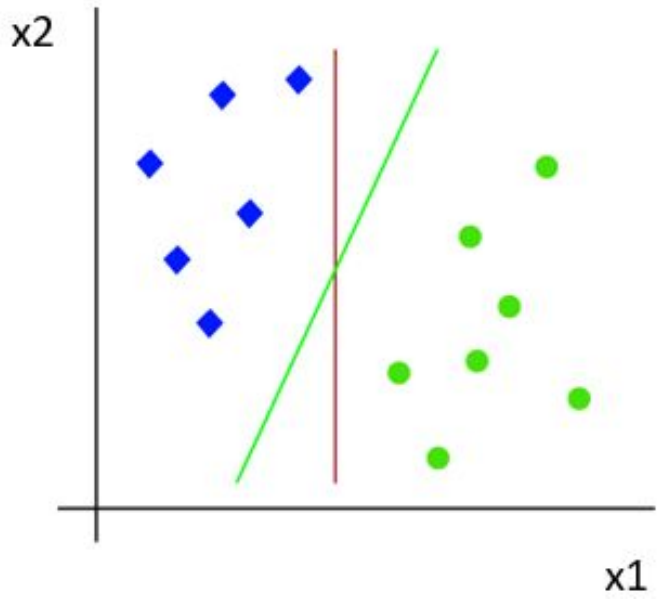
- It’s very clear that there are multiple lines that segregate our data points between red and blue circles.
- How to choose the best line or the best hyperplane that segregates our data points?

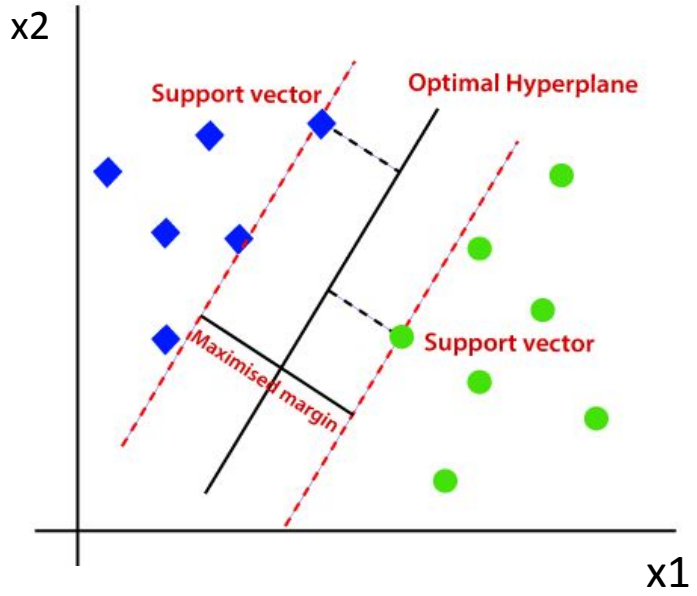
Working of SVM

- Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:

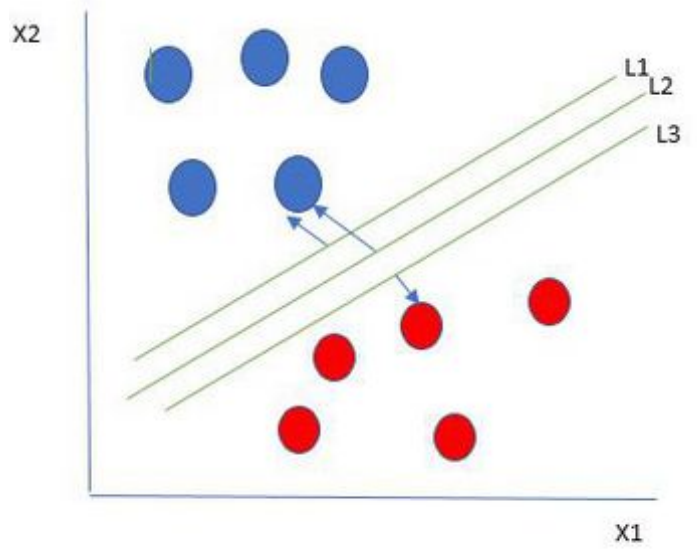


- There can be multiple lines that can separate these classes. Consider the next image:





- SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors.
- The distance between the vectors and the hyperplane is called as **margin** and the goal of SVM is to maximize this margin.
- The **hyperplane** with maximum margin is called the **optimal hyperplane**.
- In the above image, Green line is the best hyperplane as it has the maximum margin



Multiple hyperplanes separate the data from two classes

- The best hyperplane is the one that represents the largest separation or margin between the two classes.
- Choose the hyperplane whose distance from it to the nearest data point on each side is maximized.
- If such a hyperplane exists it is known as the **maximum-margin hyperplane/hard margin**. So from the below figure, we choose L2

- Let’s consider a scenario like shown in Fig. 1

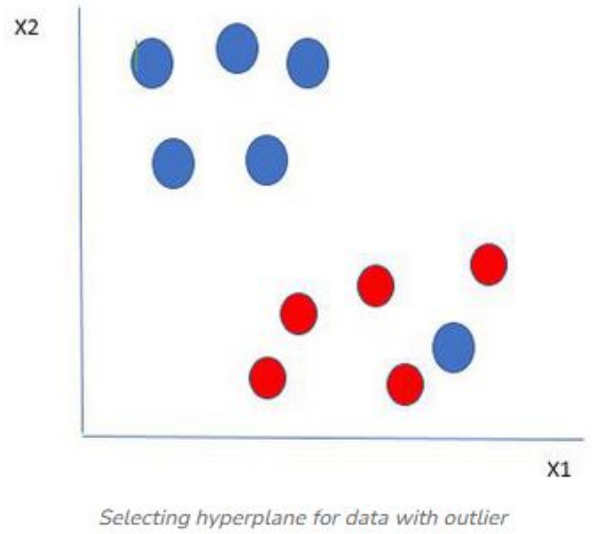


Fig. 1

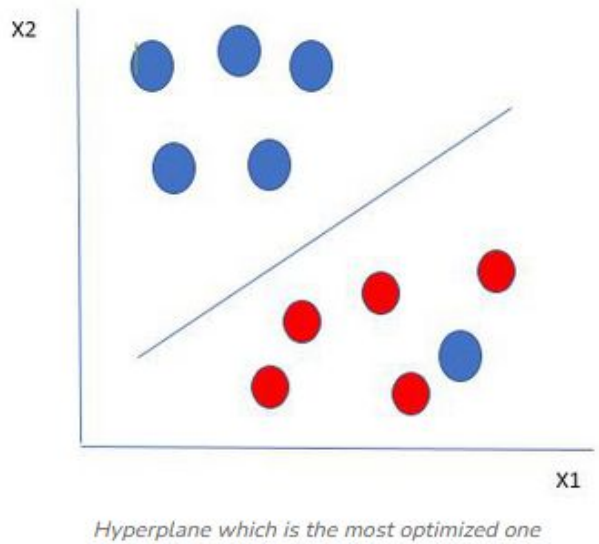
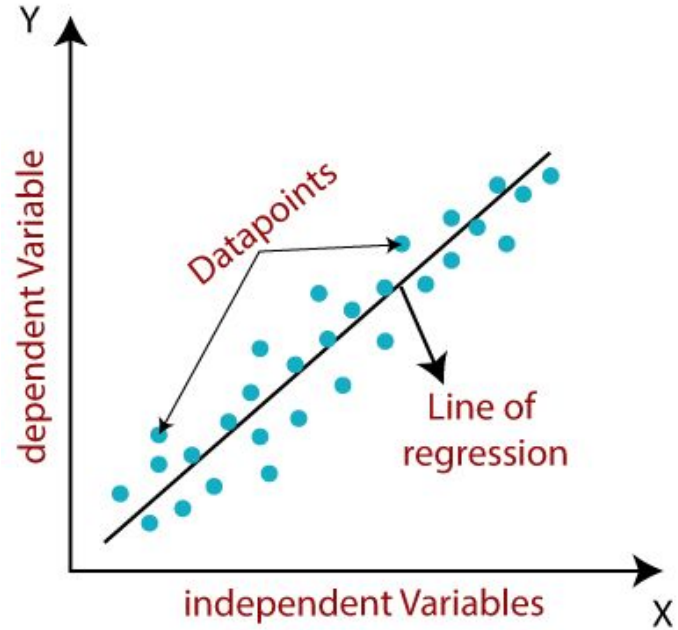


Fig. 2

- One blue ball is in the boundary of the red ball. That blue ball is an **outlier** of blue balls. In such cases, the SVM algorithm ignores the outlier and finds the best hyperplane that maximizes the margin (as in Fig 2).
- SVM is robust to outliers.

Linear Regression

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the **dependent variable** and one or more **independent variables/features** by fitting a linear equation to observed data.
- Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.
- The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



$$y = \beta_0 + \beta_1 X$$

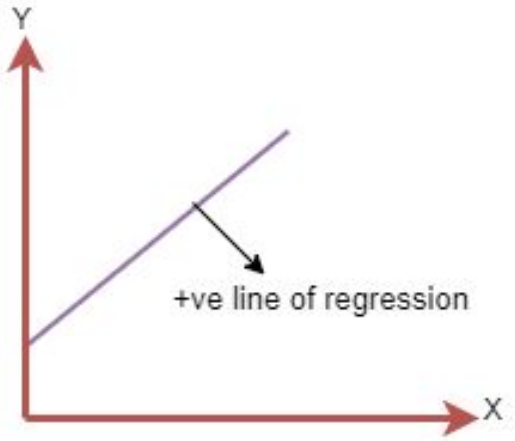
where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Linear Regression Line

- A linear line showing the relationship between the dependent and independent variables is called a **regression line**.
- A regression line can show two types of relationship:

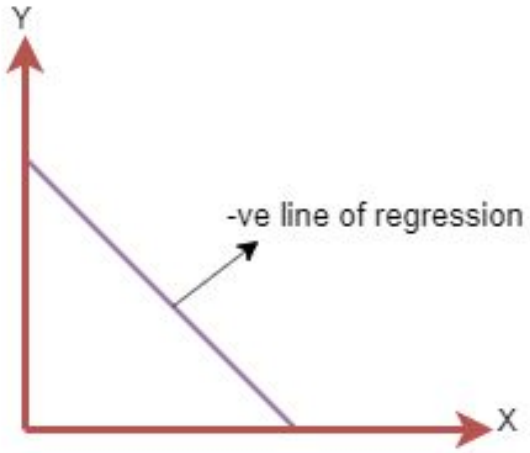
1. **Positive Linear Relationship:**
   If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

2. Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Finding the best fit line

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.
- The best fit line will have the least error.
- **Residuals:** The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.