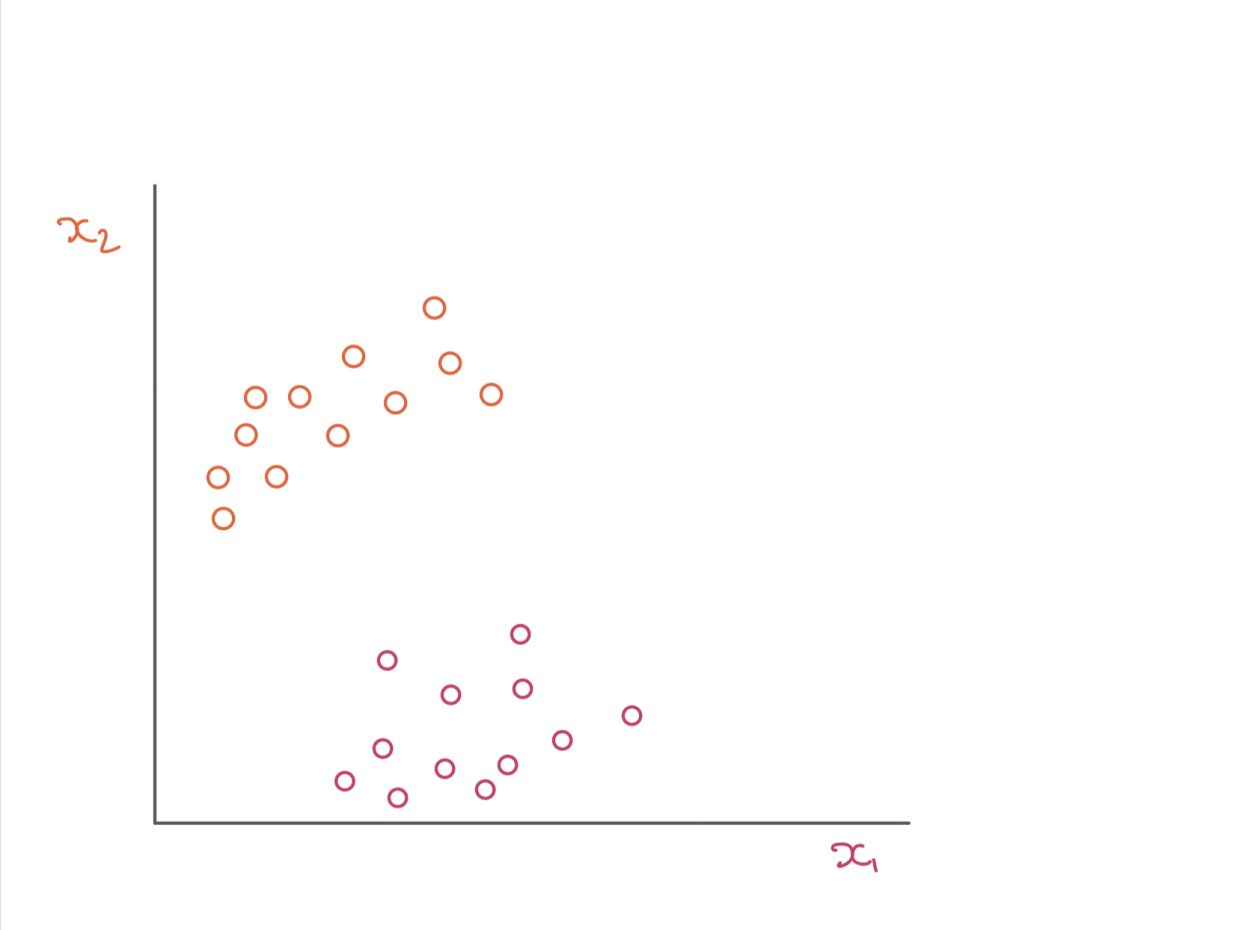
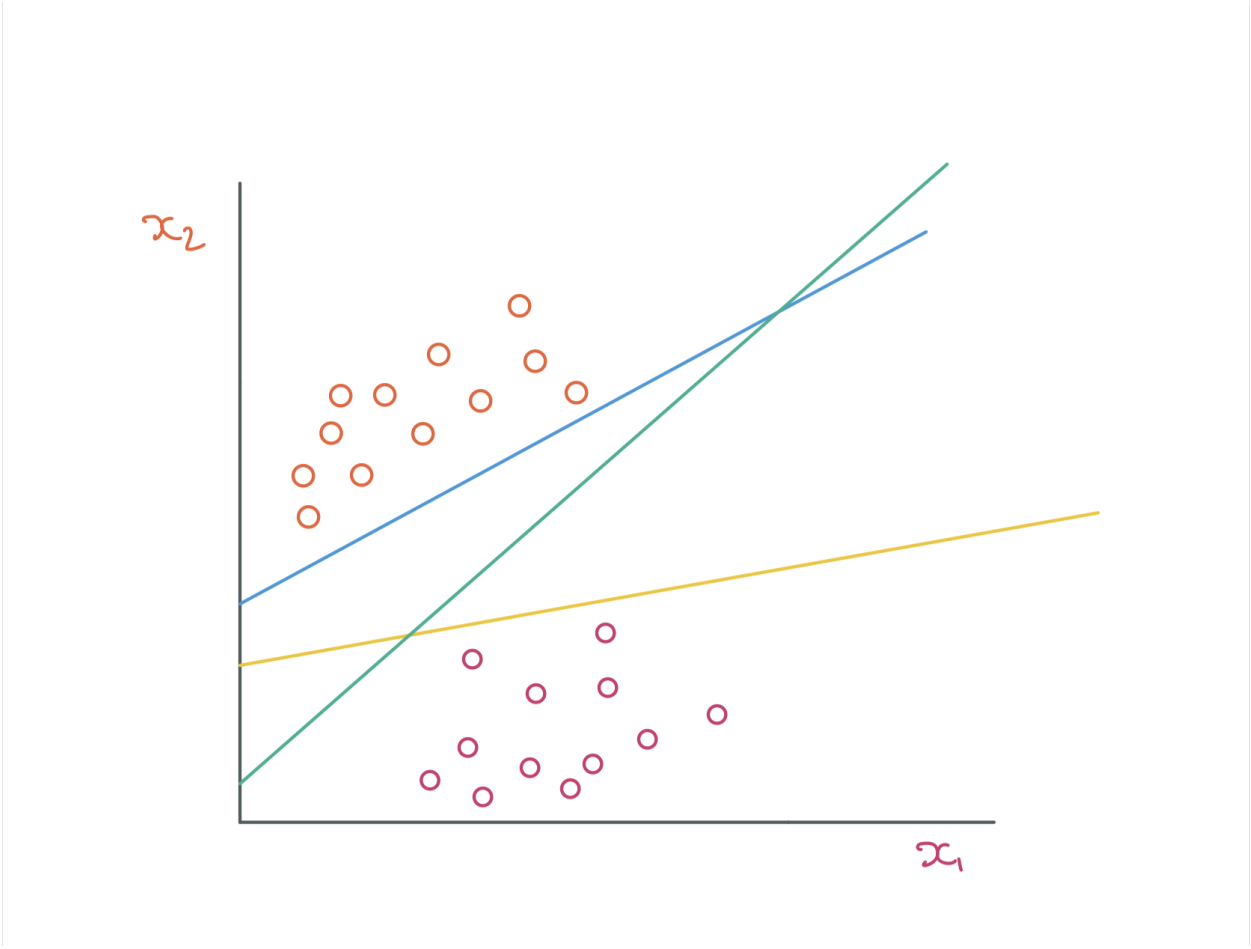
Support Vector Machines Part 1

Let's say we have a bunch of points in 2D plane that we want to separate into two clusters like in the following figure



It is not too difficult to see that a line can separate the two group of points also known as **training data points**. In fact for this specific case we can have multiple lines separating the two clusters as shown below.



Also notice a line is just a 1D plane (sounds weird). A common terminology is to just use the word hyperplane. So a 2D plane is just a hyperplane in 3 dimensions and similarly a line is a hyperplane in 2 dimensions.

A hyperplane in n dimensions can be described by

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n = 0$$

For the 2D case in the above figure, the points that lie on the line will satisfy the above equation. But what about points that do not lie on the line. For them we would get either

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n > 0$ or $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n < 0$

depending on which side of the line the points lie.

And since we are dividing the training data into two classes, say $y = +1$ for orange points and $y = -1$ for maroon data. Then let's say

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n > 0$ if $y = 1$

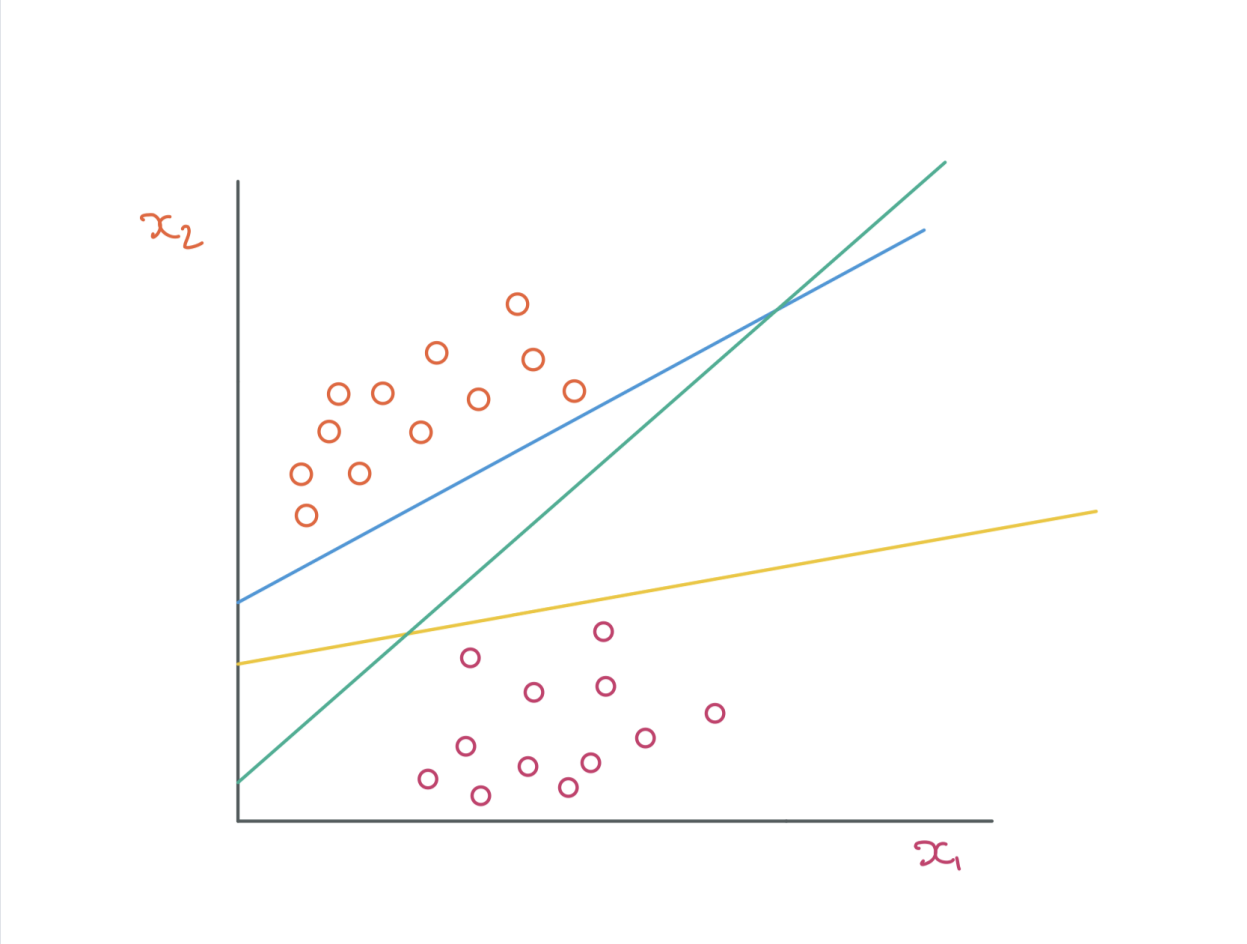
And

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n < 0$ if y = -1

We can combine this and say

> $$y_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n) > 0$$

Let's look at this again and ask



Which of these lines should we pick as our classifier? Is any one of them "better" than the other? How do we decide which separator to use?

The same problem can be generalized to higher dimension, so in case of points in 3D space, what 2D plane should we pick as our classifier?

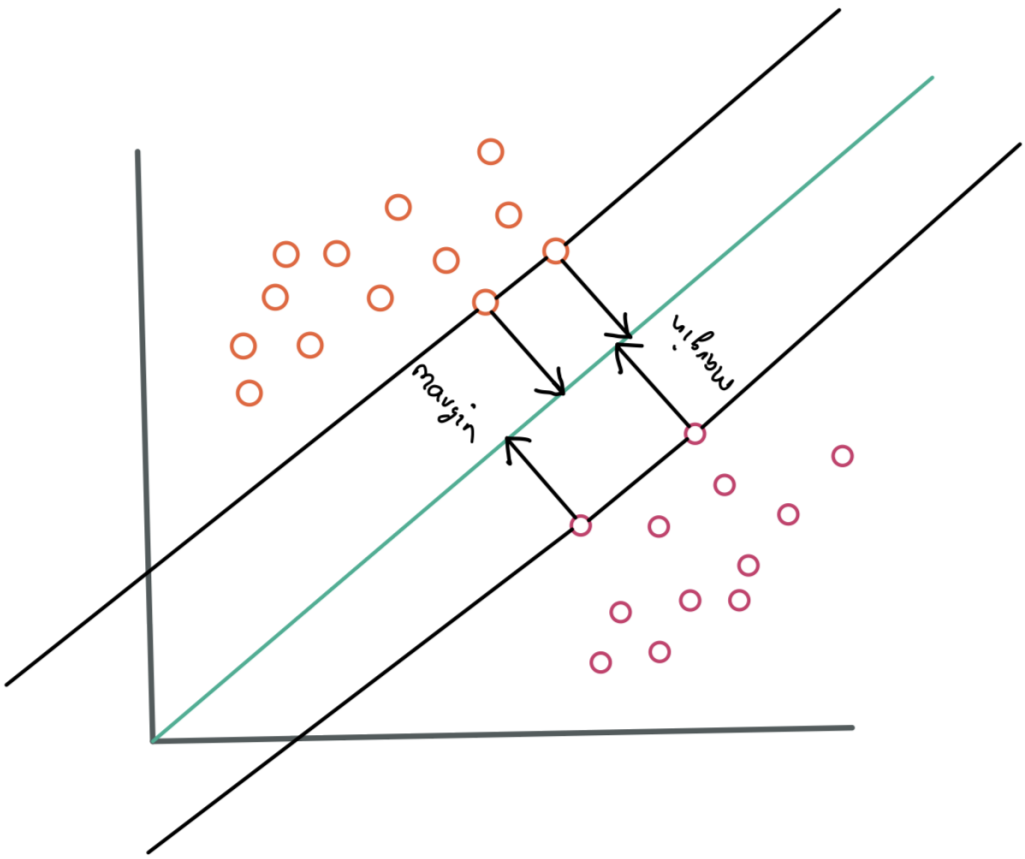
A natural choice is to pick a hyperplane that lies as far from the training observations as possible. So in the above case, if we have to pick one of these three, we will prefer the green one.

We can calculate the perpendicular distance between the training data points and the hyperplane, the smallest of these distances is known as **margin**.

The hyperplane for which this margin is **largest** is called **maximal margin hyperplane**. In the above case, green line is the maximal margin hyperplane.

Now given a new test observation, based on which side of the maximal margin hyperplane it lies, we can classify it to one or the other class.

For a better understanding of all this let's take a look at the following figure.



The distance between the closest data points and the green line is called margin. The green line itself is called maximal margin hyperplane. The two orange and two maroon data points are called **support vectors.** They are called support vectors because changing their position will change the position of the maximal margin hyperplane (green line). What's really important to understand here is that this maximal margin hyperplane only depends on these support vectors, if we were to move any other observation points, it will not affect the green separator line, as long as they do not cross boundary defined by the support vectors.

In the next part, we will look at some of the limitations of maximal margin classifier and discuss it's extension called **support vector classifier.**

References

1. An introduction to Statistical learning by Gareth James et al