CUNY School of Professional Studies

Document Classification And SVM

Keshaw K Sahay

DATA 607 – Data Scientist in context

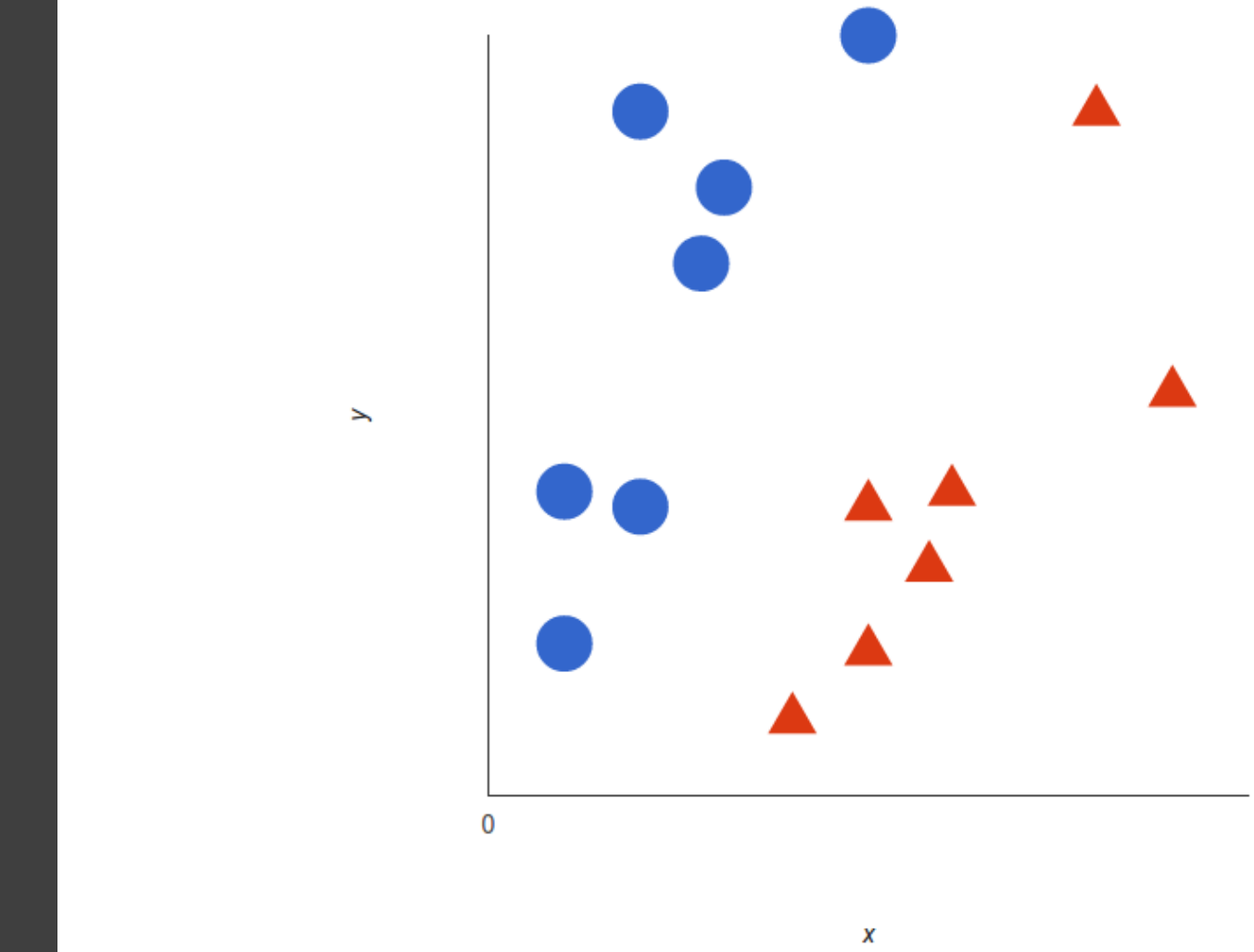
Introduction

In machine learning, **support vector machines (SVM)** are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis

Support Vector Machines Algorithm

- Let’s imagine we have two tags: *red* and *blue*, and our data has two features: x and y.
- We want a classifier that, given a pair of (x,y) coordinates, outputs if it’s either *red* or *blue*. We plot our already labeled training data on a plane:

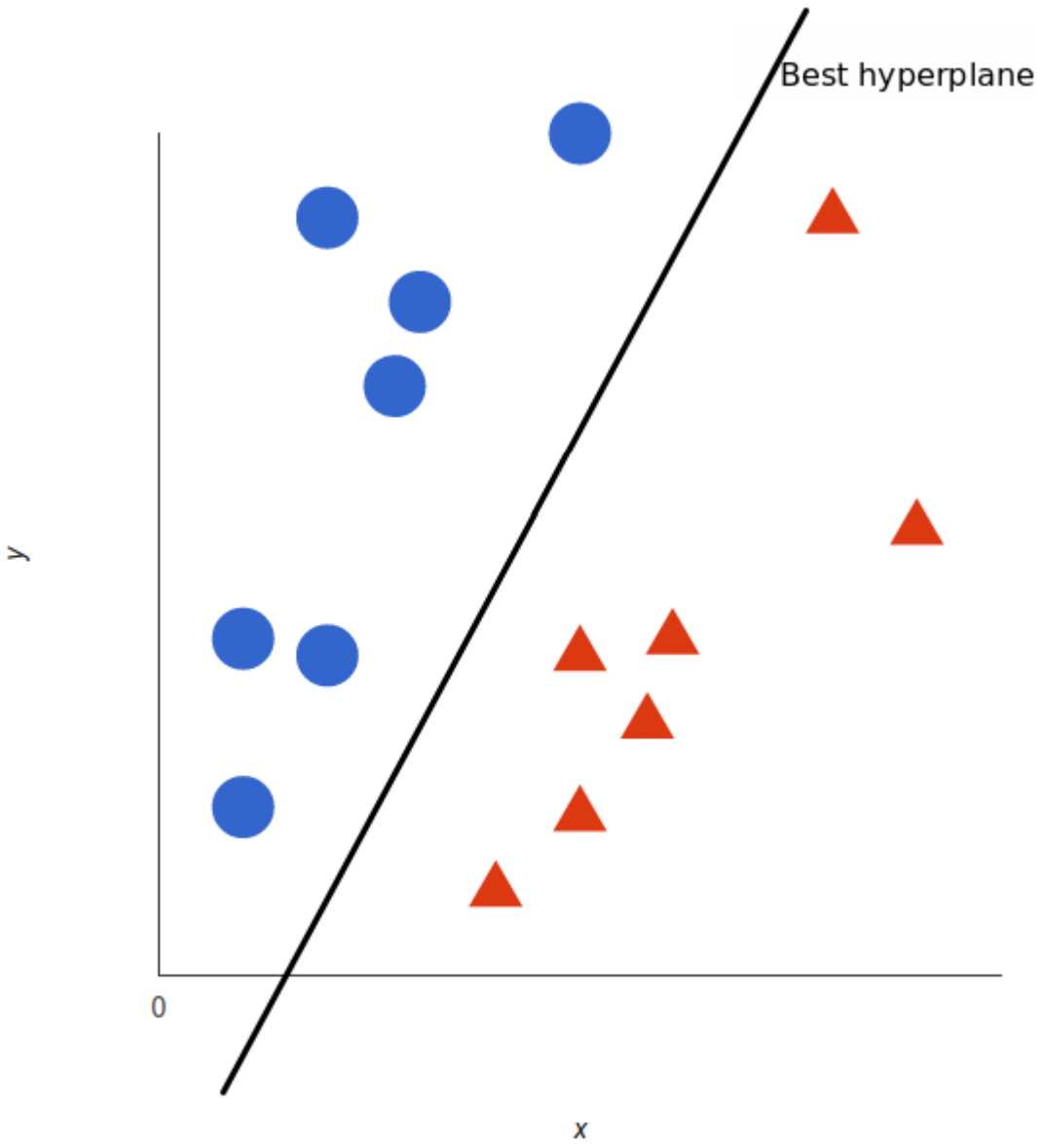
Linear Data



Support Vector Machines Algorithm

- A support vector machine takes these data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the tags.
- This line is the **decision boundary**: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red.

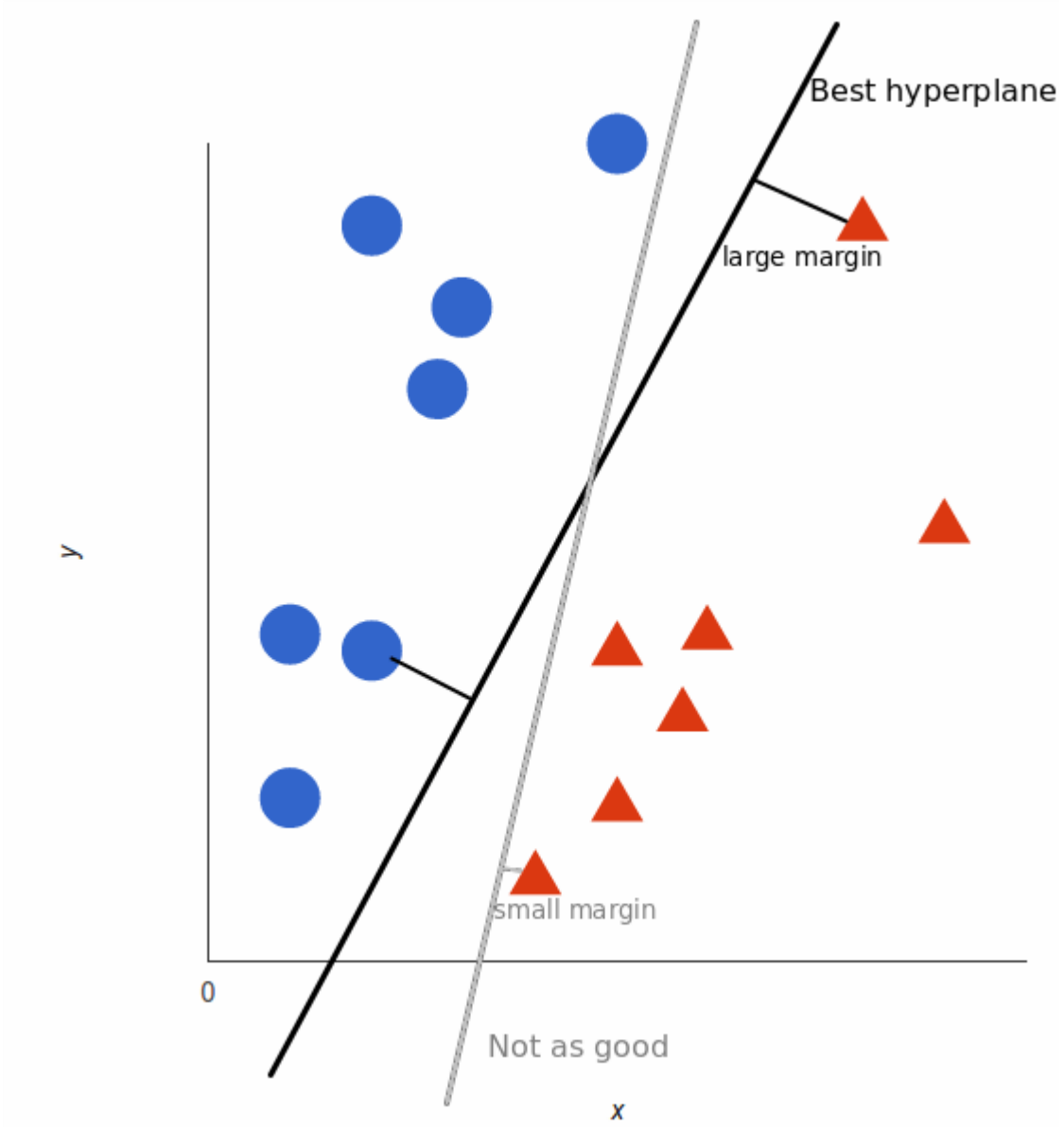
Linear Data



Support Vector Machines Algorithm

- But, what exactly is the best hyperplane?
- For SVM, it's the one that maximizes the margins from both tags. In other words: the hyperplane (remember it's a line in this case) whose distance to the nearest element of each tag is the largest.

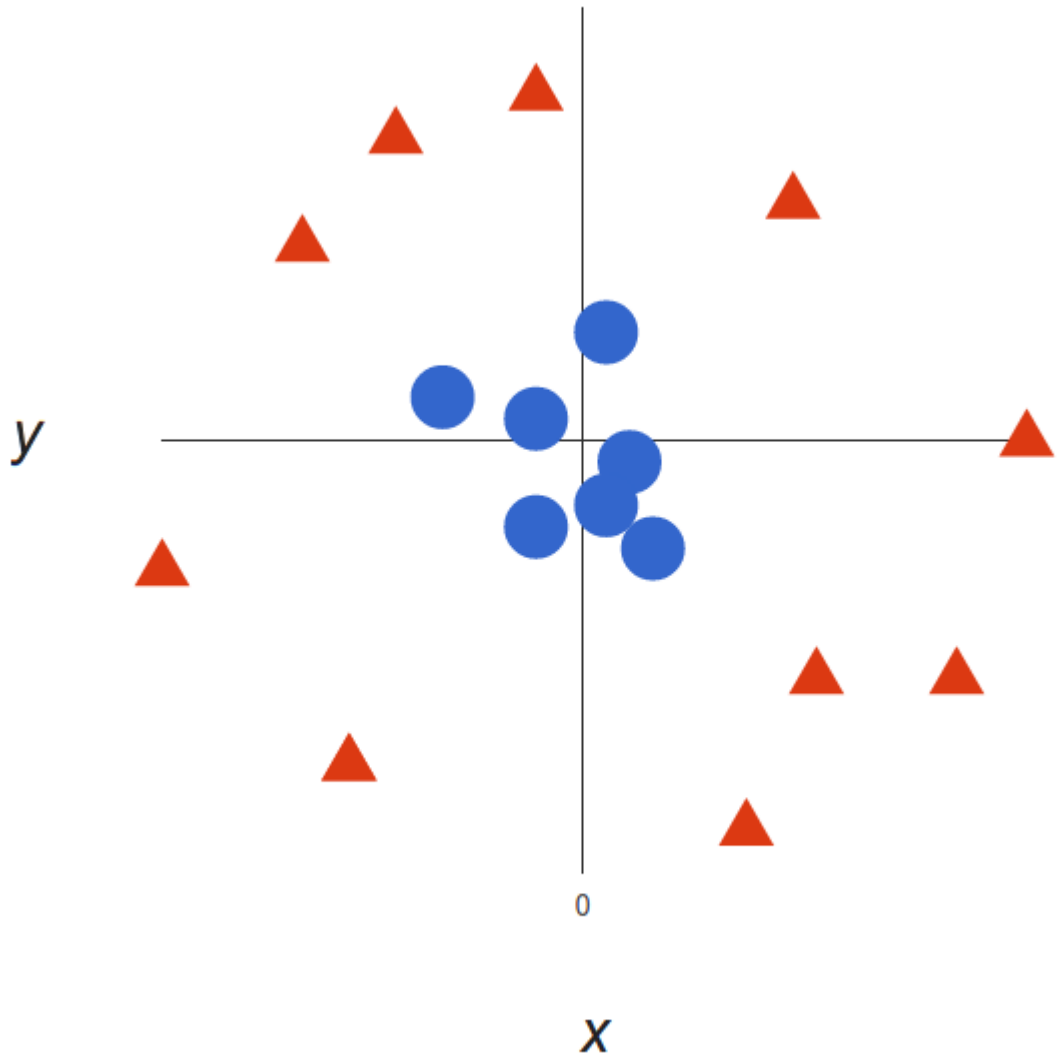
Linear Data



Support Vector Machines Algorithm

- Sadly, usually things aren't that simple
- It's pretty clear that there's not a linear decision boundary (a single straight line that separates both tags). However, the vectors are very clearly segregated, and it looks as though it should be easy to separate them

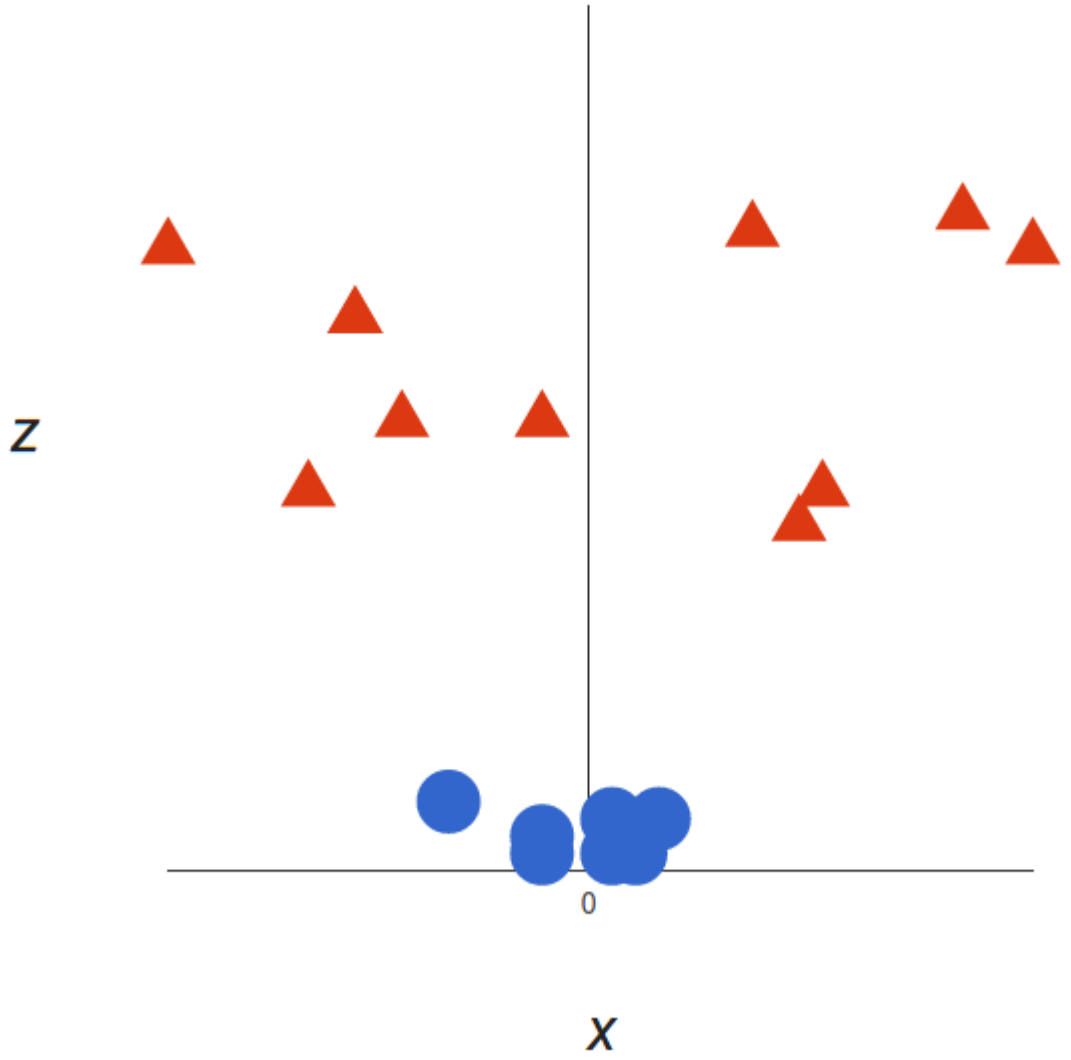
Non Linear Data



Support Vector Machines Algorithm

- So here's what we'll do: we will add a third dimension.
- Up until now, we had two dimensions: x and y. We create a new z dimension, and we rule that it be calculated a certain way that is convenient for us:
- $z=x^2+y^2$ (you'll notice that's the equation for a circle).

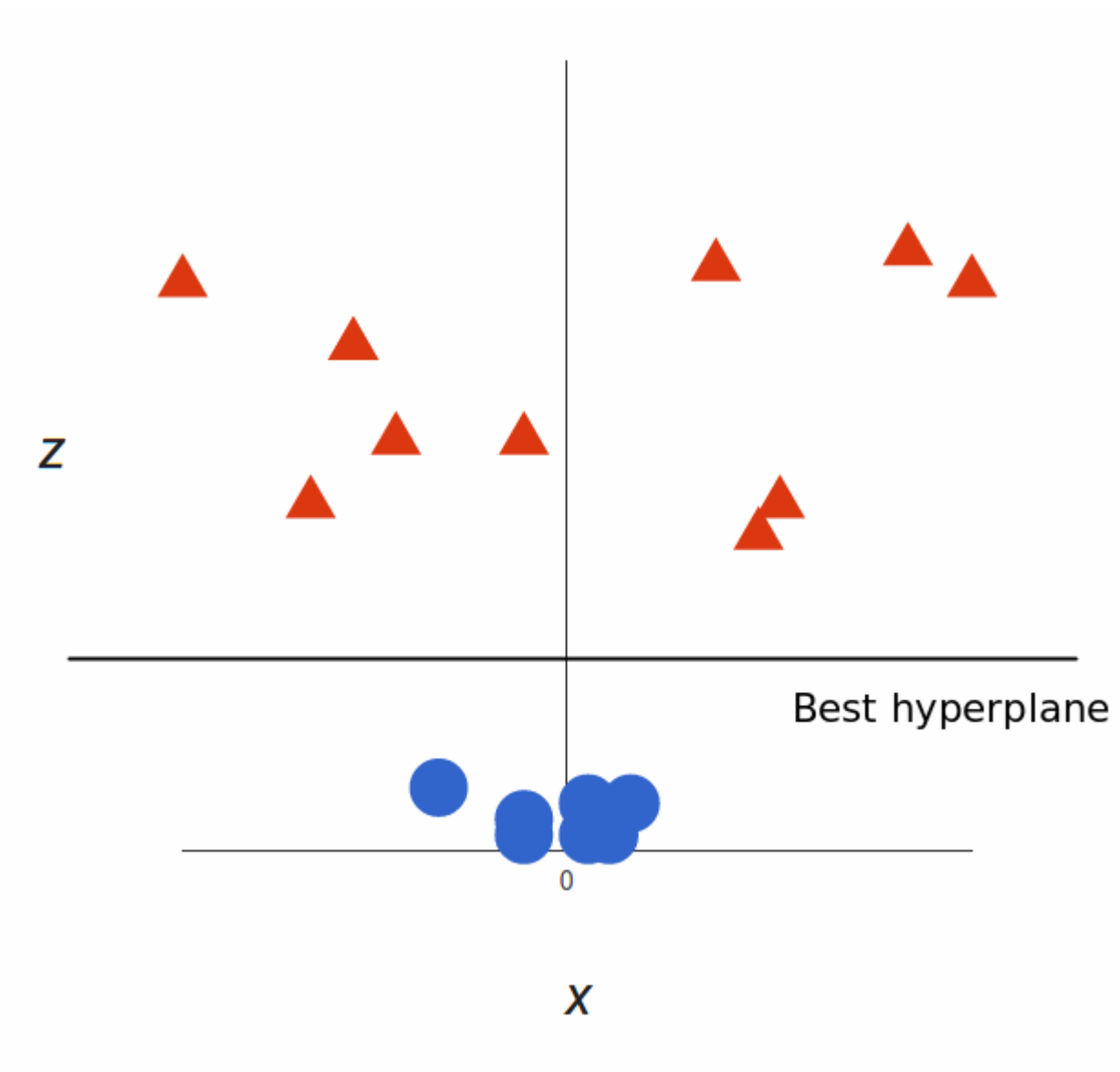
Non Linear Data



Support Vector Machines Algorithm

- Note that since we are in three dimensions now, the hyperplane is a plane parallel to the x axis at a certain z (let's say z=1).

Non Linear Data



SVM - Advantages

- **High Dimensionality**: SVM is an effective tool in high-dimensional spaces, which is particularly applicable to document classification and sentiment analysis where the dimensionality can be extremely large.
- **Memory Efficiency**: Since only a subset of the training points are used in the actual decision process of assigning new members, just these points need to be stored in memory (and calculated upon) when making decisions.
- **Versatility**: Class separation is often highly non-linear. The ability to apply new kernels allows substantial flexibility for the decision boundaries, leading to greater classification performance.

Conclusion

Support Vector Machines are a subclass of supervised classifiers that attempt to partition a feature space into two or more groups. They achieve this by finding an optimal means of separating such groups based on their known class labels:

- In simpler cases the separation "boundary" is linear, leading to groups that are split up by lines (or planes) in high-dimensional spaces.

- In more complicated cases (where groups are not nicely separated by lines or planes), SVMs are able to carry out non-linear partitioning. This is achieved by means of a kernel function.

Thank You