To simplify the explanations on SVM algorithm, the following will assume that we are dealing with binary classification problem. The possible label involved can only be either 1 or -1. Sometimes the label will be simply called a plus or minus. The multiple classes classification can be implemented simply by generating multiple SVM classifiers.

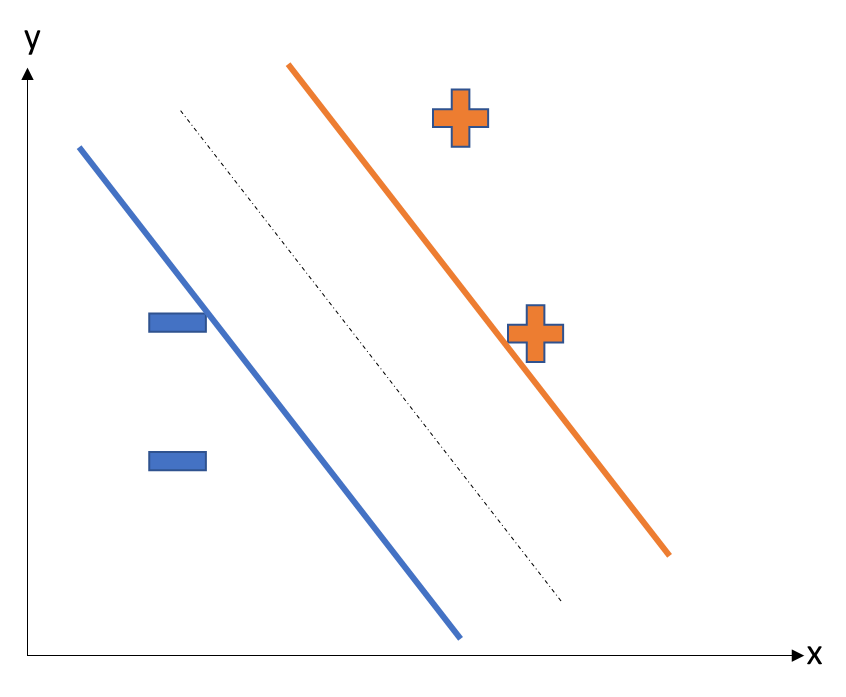


Figure 1 Example of Binary Classification

Suppose we have several samples that are labelled with minus and plus signs which represent two classes. We separate the samples with a plane which is represented by the dotted black line in figure 1. The orientation of this plane can be different as long as the plane can succeed to separate the two kinds of samples into different classes. The distance from the plane to blue line and to orange line is the same, it is called the margin, and the points what are exactly located on the blue lines and orange lines are called support vectors. This was also where the classifier name “Support Vector Machine” comes from. The objective is this algorithm is to find such support vectors to maximize the margin to minimize the generalization error.

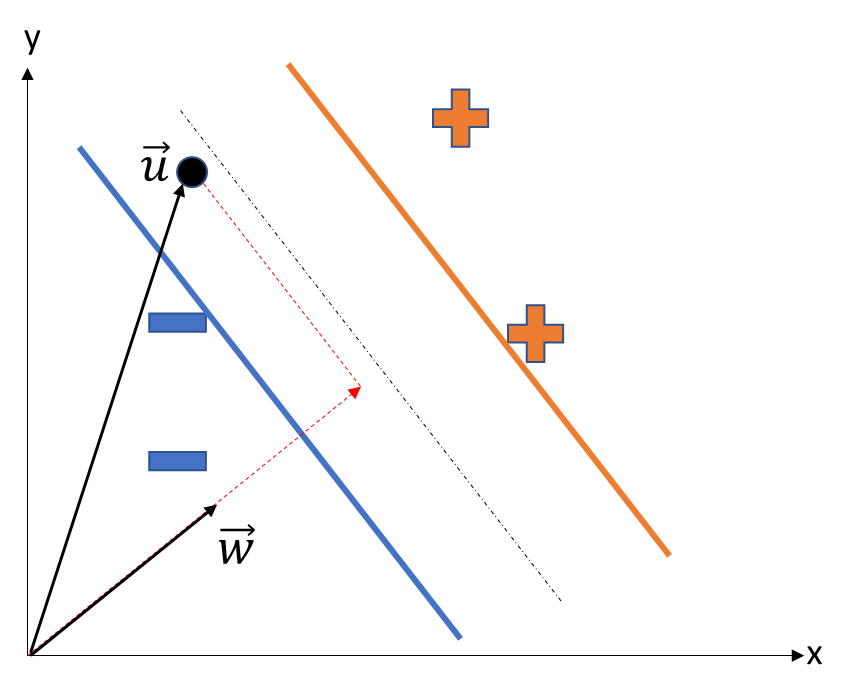


Figure 2 Classification of An Unknown Sample

In figure 2, we add a vector perpendicular to the plane (dotted black line). The length of this vector is random. In the coordinate system as shown in figure 2, the dotted black line can also be represented by the vector . When we need to classify the coming new sample , we just need to compute the length in the direction of , which is . If the length exceeds a certain constant , then this new sample will be classified as a plus sign, otherwise, it would be defined as minus. To formally define this, if the new point satisfies

Eq1

, then the point would be labelled as a plus sign. Otherwise, the point will be a minus sign. Equation 1 is also called the decision rule. Now what we are going to do is to find a suitable vector and b value to maximize the margin. To achieve this, we define

Eq2

for each sample between the blue line and orange line. From equation 2, we can derive the width of the margin as

Eq3

To find vector and b value that maximize the margin, we can simply define the objective function as

Eq4

With the constraints from eq1 and eq2, we need to introduce the Lagrange multipliers to solve this convex programming problem. In this way, the objective function become

Eq5

Equation 5 is also called Lagrange primal function, , and it becomes a dual problem. Set the partial derivative to be 0, we can easily get

Eq6

Eq7

By plugging equation 6 and equation 7 back to equation 5, we get the Lagrangian dual function

Eq8

Subject to ,

Equation 8 manifests that the Lagrangian function solely depends on the pair of sample points when Lagrange multipliers are fixed. This gives us the general algorithm for a SVM classifier:

|  |
| --- |
| **Algorithm 2** Support Vector Machine |
| Repeat till convergence {   1. Select some pair and to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum) 2. Reoptimize with respect to and , while holding all the other ’s fixed.   } |

Using the algorithm 2, we can specify the suitable for each sample . Once the Lagrange multipliers are fixed, the plane with maximal margin are decided. Then we can use the equation 1 and equation 6 combined to decide the label of an unknown point

Eq9

Above is the process of deciding a separating plane for separable sample set, which is also called hard margin. As for the non-separable sample set, we will create a soft margin that can tolerate misclassified samples. Figure 3 clearly stated the differences between hard margin and the soft margin.

The whole equation deduction is similar. I briefly introduce it here. We add a penalty parameter C of error term to the Lagrange primal function, the Lagrange primal function becomes

Eq10

Subject to ,

As we solve this quadratic problem, we can easily get the Lagrange dual function

Eq10

Subject to ,

The KKT conditions for the soft margin are

We can see that the penalty term limits the range of , this is how the value affects the final accuracy. In this project, the data sets are non-separable. This can be inferred from figure below. In the figure below, as we decrease the C value exponentially, the train accuracy and test accuracy both decrease when the C value reach at the interval (1E-5, 1E-7). This means if we keep reducing the penalty on the misclassified samples in the training process, we will expect a decrease of both training and test accuracies. It can be inferred that if C approaches 0, the accuracy will drop significantly for our problem.

