CME 216, ME 343 - Spring 2020

Eric Darve, ICME

Stanford University

Support vector machine (SVM) is one of the simplest methods for classification.

It forms a stepping block to neural networks and deep learning.

SVM is a method for binary classification.

That is, we are given a point $x \in \mathbb{R}^n$, and we want to predict a label with possible values $+1$ or -1.

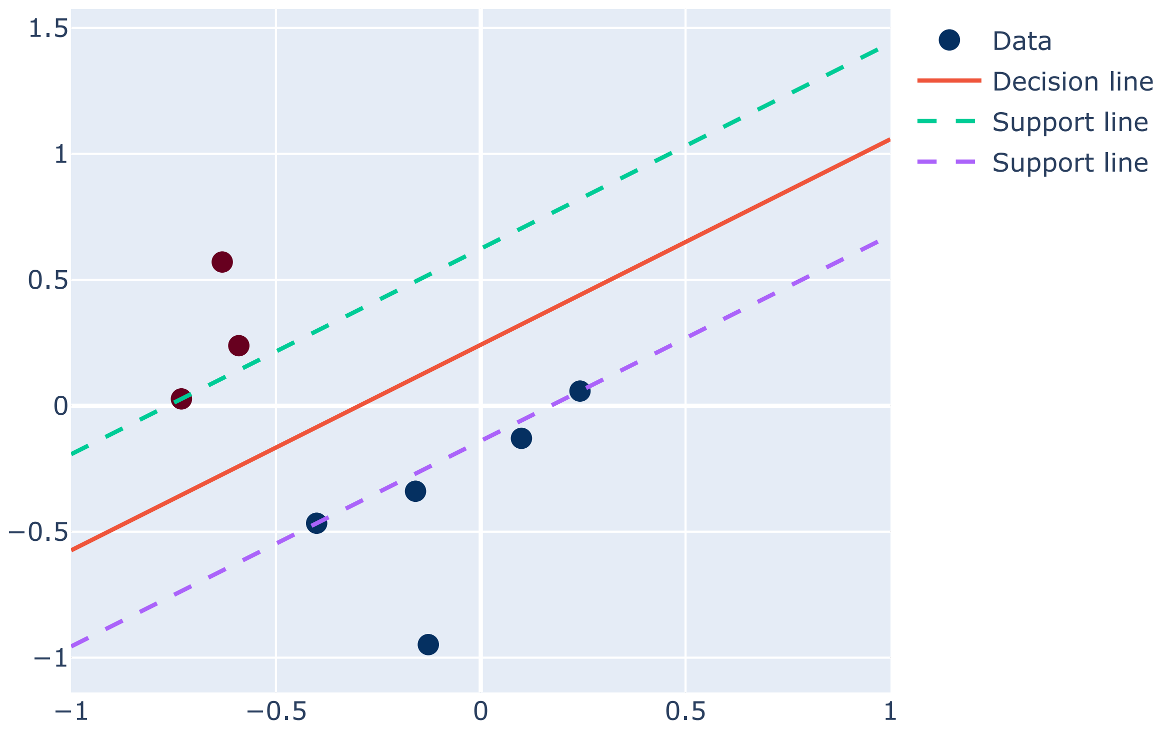
In SVM, the space $\mathbb{R}^n$ of possible x is subdivided into 2 half-space by a hyperplane: a line in 2D or a plane in 3D.

On one side of the hyperplane, the label is 1 and is -1 on the other.

-1 labels from $+1$ is simply $y = x$.

The -1 labels are in the top left and the $+1$ in the bottom right.

SVM



Data
Decision line
Support line
Support line
1.5
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

However, if the training data that we are given are only the colored dots, we cannot exactly determine the separating line $y = x$.

So instead, based on the observed data (the dots), we ask what the best hyperplane we can find is.

The hyperplane is defined by a normal vector w and a bias b:

$$w^T x + b = 0$$

Once the hyperplane is found, the classification is given as follows.

If $w^T x + b > 0$ then we predict that the label is 1, otherwise the label is -1.

In SVM, the best hyperplane is defined as the one that has the largest margin, that is the one which the points are the farthest from.

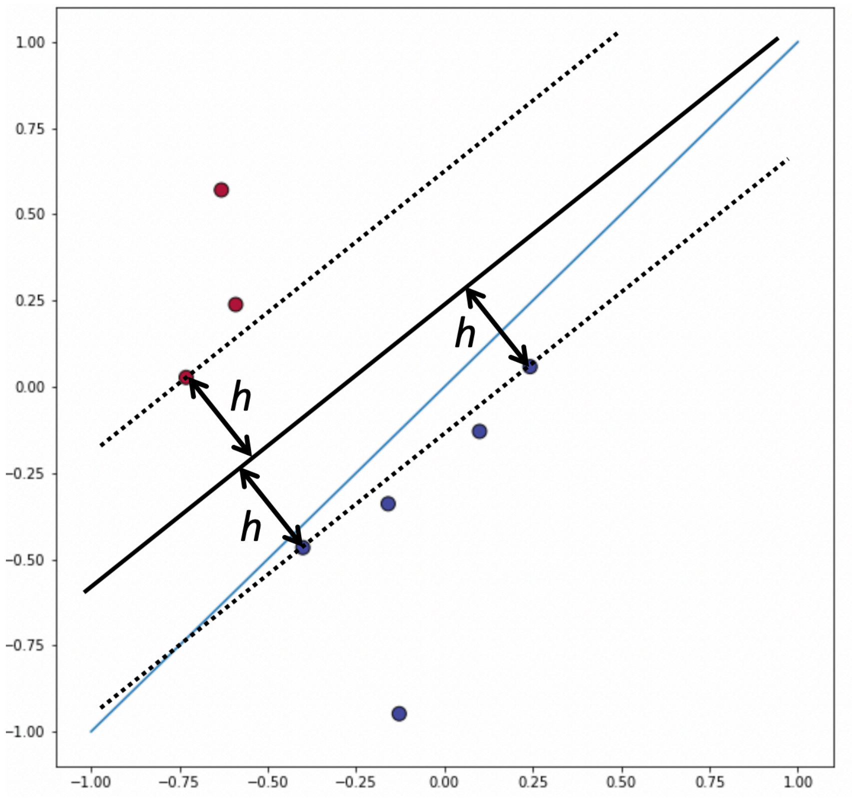
This makes the classifier more robust and accurate.

In the figure below, the "exact" solution is

$$y = x$$

(the solid blue line), but this is unknown to us.

Instead, we observe only the colored dots. Based on this, the best line of separation is the black solid line in the middle.



1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
h
h
h

The goal in SVM is to determine the equation of the black solid line such that the distance h is maximum.

By definition, no training points (the colored dots) can reside between the dashed lines. The black solid line must be equidistant from the two dashed lines.

Let's now see how this can be formulated mathematically.

The first step is calculating the distance of a point to the separating hyperplane (black solid line above).

This hyperplane will be defined by the equation
$w^Tx + b = 0$

w: vector in $\mathbb{R}^n$

b: scalar

Take a point x not on the hyperplane.

How far is it from the hyperplane?

You can prove that the distance δ is

$$\delta = \frac{|w^T x + b|}{|w|}$$

We then need to search for (w, b) that makes δ as large as possible.

The division by $|w|$ indicates that there is a scaling invariance in this problem.

We can multiply w and b by any constant C and the hyperplane/classifier is the same.

$$\textbf{maximize}\ \delta = \frac{|w^T x + b|}{|w|}$$

is simplified to

To normalize, we choose w and b such that

- $y_i(w^T x_i + b) \geq 1$ for all i, and
- there must be some i for which

$$y_i(w^T x_i + b) = 1$$

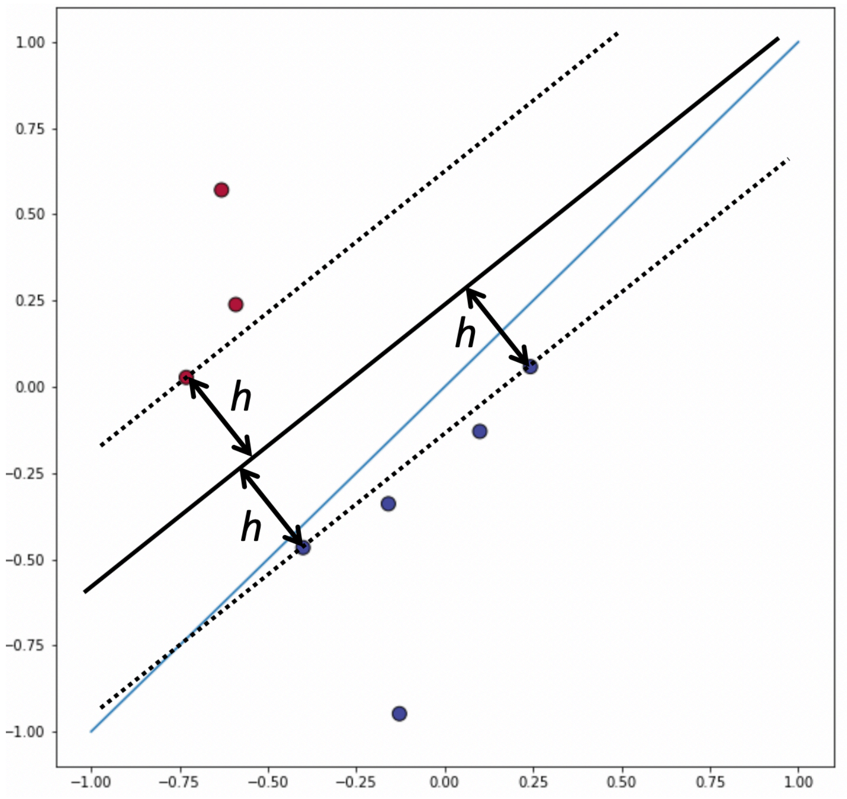
This happens for the point(s) that is closest to the plane.

The condition $y_i(w^T x_i + b) = 1$ may look strange.

But, recall that:

- when the label $y_i = -1$, we expect $w^T x_i + b < 0$, and
- when $y_i = 1$, we expect $w^T x_i + b > 0$.

So at least in terms of the signs, this equation makes sense.



1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
h
h
h

Let's denote by ρ the distance of the point closest to the hyperplane. Then:

$$\rho(w, b) = \min \frac{|w^T x_i + b|}{|w|}$$

With our normalization this becomes simple:

$$\rho(w, b) = \frac{1}{|w|}$$

since $|w^T x_i + b| = 1$ for the nearest point.

We search for w such that ρ is maximum.

Equivalently $|w|$ minimum with $|w^T x_i + b| = 1$.

To solve this problem, we can re-write it as a quadratic programming problem.

We don't need to know what this is in detail but what matters is that there are efficient methods to solve this type of problem.

Since we want to maximize the distance of the nearest point ρ, we minimize $|w|_2$.

We search for

$$(w, b) = \mathbf{argmin}_{w,b} \frac{1}{2} |w|_2^2$$

subject to the constraint

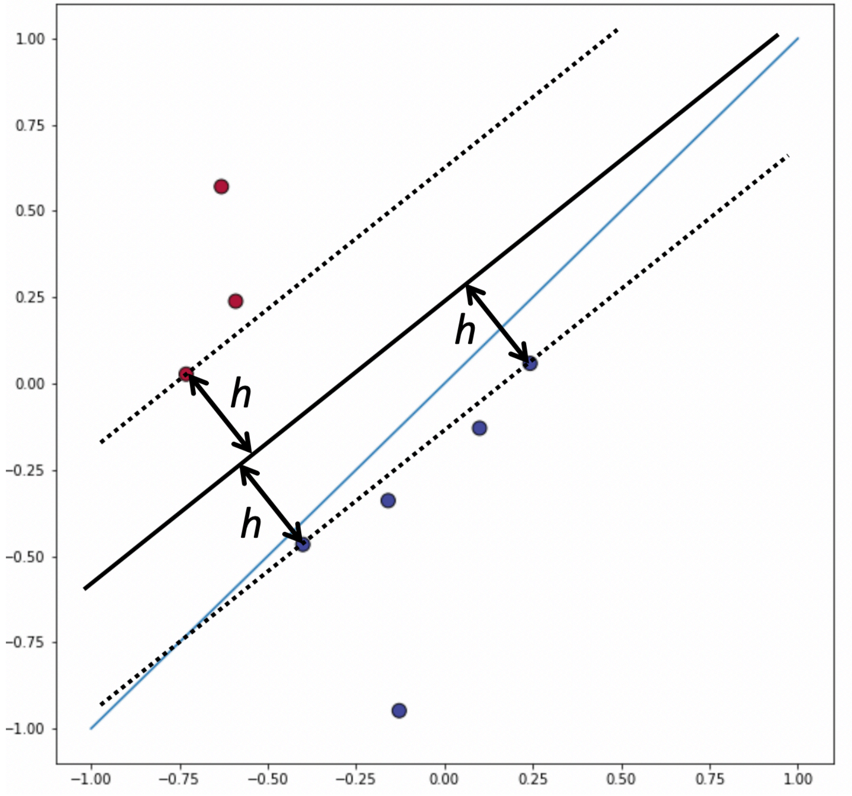
$$y_i(w^T x_i + b) \geq 1$$

For the optimal solution, there must be at least one x_i for which

$$y_i(w^T x_i + b) = 1$$

These x_is are called **support vectors.**

See Section 5.7.2 in [Deep Learning]



1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
h
h
h

In the previous figure, the red and blue dots lying on the dashed lines are the support vectors.

The black solid line is our best guess $w^T x + b = 0$ where x is a point in the plane $x = (x_1, x_2)$.