Supervised learning - Classification (Demo)

Support Vector Machines (SVM)

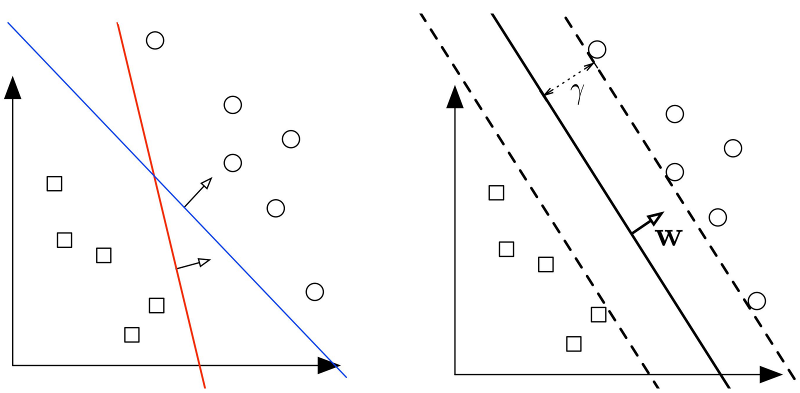
STAT 32950-24620

Spring 2025 (wk6)

Support Vector Machines (SVM)

SVM is a classification method with two outstanding characters:

- Maximizing classification margin.
- Readily generalizable using the kernel method*.



Review: Distance of a point to a line

The distance of a point $x_o \in \mathbb{R}^p$ to a line $w'x + b = 0$ can be written as

$$d = d(x_o, L) = \frac{|w'x_o + b|}{\|w\|} \tag{1}$$

where

$x, w \in \mathbb{R}^p$ are vectors.

$|\cdot|$ is the absolute value,

$\|\cdot\|$ is the vector norm.

The most common norm is the Euclidean norm, or the 2-norm.

Signed distance of a point to a line

The **signed distance** of a vector x to a line $w'x + b = 0$ is defined as

$$\frac{w'x + b}{\|w\|} \tag{2}$$

which is also called **directional distance** of point x to line $w'x + b = 0$.

In higher dimensional space with $x \in \mathbb{R}^p$, $p > 2$,

the equation $w'x + b = 0$ represents a hyperplane.

Vector $w \in \mathbb{R}^p$ in $w'x + b = 0$ is the **normal vector** of the hyperplane.

SVM for linear separable 2-classes

First consider the simplest case that there exists a linear classifier.

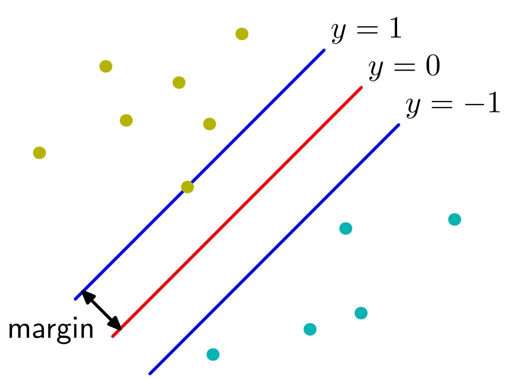
There is a line or hyperplane that completely separate the points in 2 classes.

SVM aims for the linear classifier maximizing the margin between the 2 classes.

SVM classifier formulation

Denote the class label of a training point x as y, with values $y = 1$ or $y = -1$.

Margin



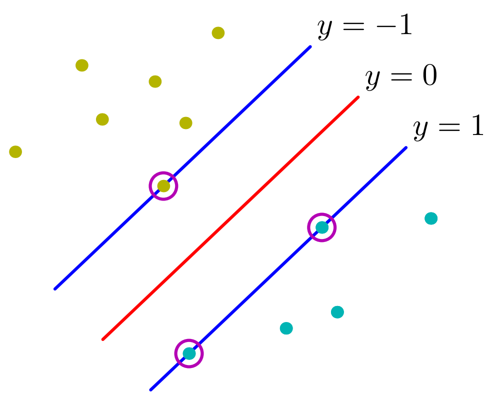
Courtesy of C. Bishop.

Properties of the SVM classification hyperplane

Properties of the SVM classification hyperplane H: $w'x + b = 0$:

- H divides the two classes.
- $w'x + b > 0$ for x in class $y = 1$, and $w'x + b < 0$ for $y = -1$.
- There is $c > 0$, such that there are **supporting vectors** on the **margin hyperplanes** $w'x + b = \pm c$:
  - there are vectors x with $w'x + b = c, y = 1$, and
  - there are vector x with $w'x + b = -c, y = -1$.
- Other vectors x should have $|w'x + b| > c$.

Support Vectors



Courtesy of C. Bishop.

Conventional SVM parameterization

The margin hyperplanes $w'x + b = \pm c$ is equivalent to $(w/c)'x + b/c = \pm 1$, which can be written as $w^{*\prime}x + b^* = \pm 1$.

We can rescale to express the margin hyperplanes as $w'x + b = \pm 1$.

Then the SVM formulation becomes

$$w'x + b \begin{cases} \geq 1, & y = 1 \\ \leq -1, & y = -1 \end{cases}$$

Combine the two inequalities, the SVM classifier can be stated as

$$y(w'x + b) \geq 1 \tag{3}$$

with the objective to maximize the margin.

Margin size

If x_1 is a supporting vector with $w'x_1 + b = 1$, and

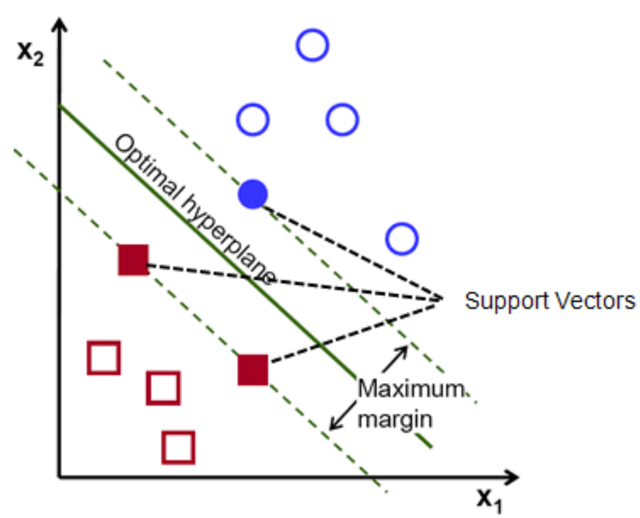
x_2 is a supporting vector with $w'x_2 + b = -1$,

the distance between the two margin hyperplane $w'x + b = \pm 1$ is

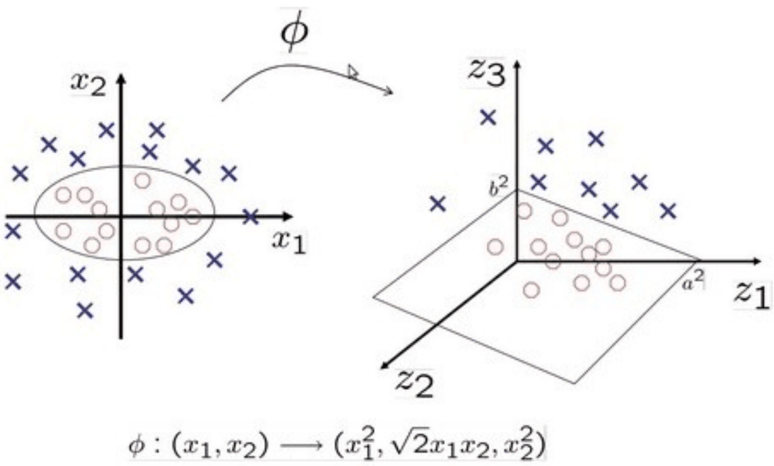
$$\left| \frac{w'}{\|w\|}(x_2 - x_1) \right| = \frac{|w'x_2 - w'x_1|}{\|w\|} = \frac{|(1-b) - (-1-b)|}{\|w\|} = \frac{2}{\|w\|}$$

This is the quantity SVM aims to maximize.

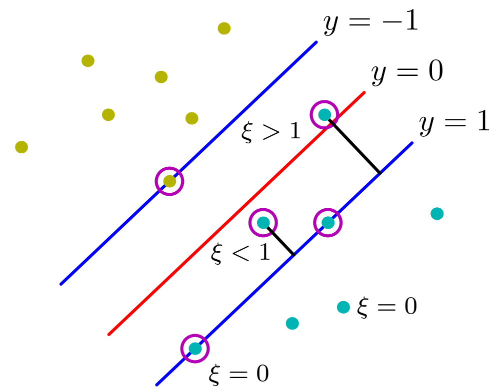
Maximized margin



Kernel method example*



Soft margin example*



Courtesy of C. Bishop.

Formulation of soft margin with slack variable*

For each feature point on the wrong side of the margins, let ξ denote the distance of the point to its margin.

The objective:

$$minimize_{w,b,\xi_i} \left(\|w\|^2/2 + C\sum_{i=1}^{n} \xi_i \right)$$

under the constraints

$$y_i(w^T x_i + b) \geq 1 - \sum_{i1}^{n} \xi_i, \qquad \xi_i \geq 0, \quad i = 1, \cdots, n.$$

Remarks about basic SVM

- Linear boundaries with some optimal-separation theoretical properties.
- Transform to higher dimensions to obtain linearly separation (kernel function).
- Based on a theoretical model of learning explicitly, with guaranteed performance.
- Not affected by local minima.
- Do not suffer from the curse of dimensionality.
- Quadratic program, doable.
- Optimization algorithm instead of greedy search.
- The kernel function has to be handpicked.
- Integrated into other high performers such as deep neural network.