20. Support vector machines

- A support vector machine classifies data into two classes.
- Data
  - The i-th data vector $\boldsymbol{x}(i)$
  - The class that $\boldsymbol{x}(i)$ belongs to is represented by $y(i)$.

$$y(i) = \begin{cases} 1, & \boldsymbol{x}(i) \text{ belongs to class 1,} \\ -1, & \boldsymbol{x}(i) \text{ belongs to class 0.} \end{cases}$$

As before we use vector $\boldsymbol{w}$ to classify the data.

If $\boldsymbol{w}^T\boldsymbol{x}(i) - b > 0$ then $\boldsymbol{x}(i)$ belongs to class 1.

If $\boldsymbol{w}^T\boldsymbol{x}(i) - b < 0$ then $\boldsymbol{x}(i)$ belongs to class 0.

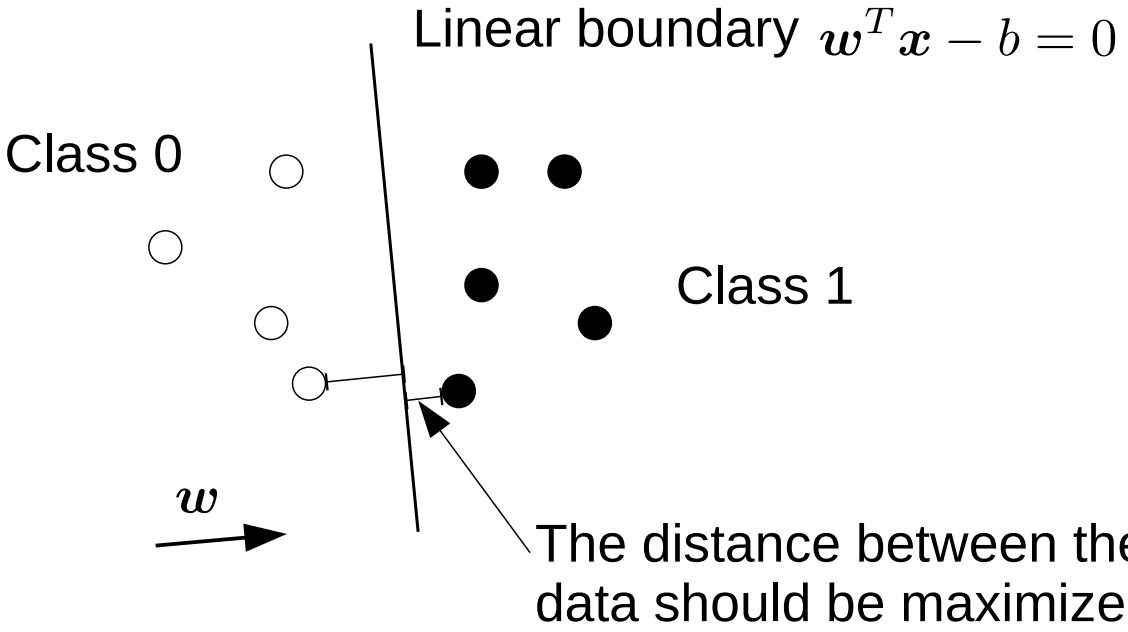
$$\Updownarrow$$

$$y(i)(\boldsymbol{w}^T\boldsymbol{x}(i) - b) > 0$$

We want to derive good $\boldsymbol{w}$ and $\boldsymbol{b}$.

20. Support vector machines

- Good $\boldsymbol{w}$ and b?
  - The data should be well away from the boundary.



The distance between the boundary and the data $\boldsymbol{x}(i)$ is

$$\frac{|\boldsymbol{w}^T\boldsymbol{x}(i) - b|}{\|\boldsymbol{w}\|},$$

therefore $\min_i \frac{|\boldsymbol{w}^T\boldsymbol{x}(i) - b|}{\|\boldsymbol{w}\|}$ is to be maximized.

20. Support vector machines

The boundary is represented as

$$\boldsymbol{w}^T\boldsymbol{x} - b = 0,$$

which is equivalent to

$$(k\boldsymbol{w})^T\boldsymbol{x} - (kb) = 0,$$

for any nonzero k.

So, let us choose k such that $\min_i |k\boldsymbol{w}^T\boldsymbol{x}(i) - kb| = 1$ holds.

Now, let us re-define $k\boldsymbol{w}$ as $\boldsymbol{w}$ and kb as b. Then, maximization of

$$\min_i \frac{|\boldsymbol{w}^T\boldsymbol{x}(i) - b|}{\|\boldsymbol{w}\|}$$

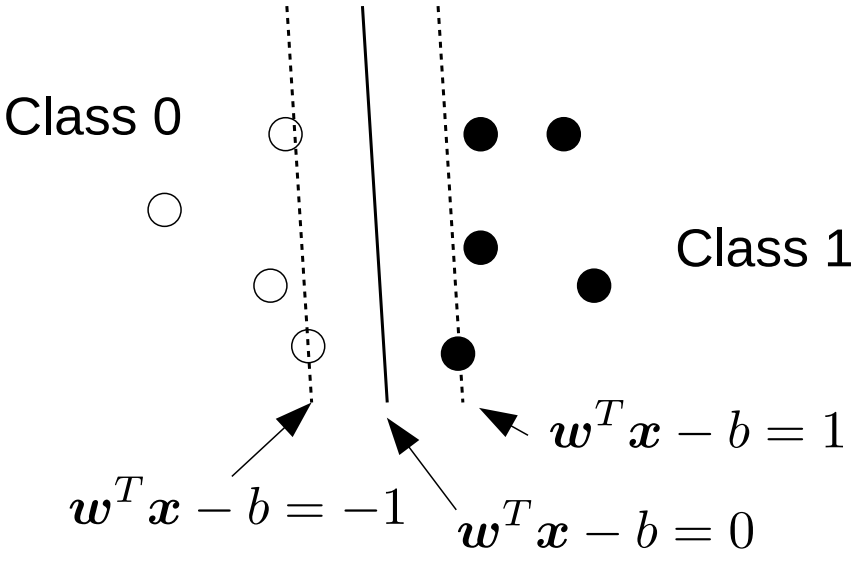
is achieved by minimizing $\|\boldsymbol{w}\|$.

20. Support vector machines

- Obtaining good $\boldsymbol{w}$ and b
  - To summarize the above discussion, good $\boldsymbol{w}$ and b are obtained by solving the following optimization problem:

$$\text{Minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 \text{ subject to } y(i)(\boldsymbol{w}^T\boldsymbol{x}(i) - b) \geq 1.$$

  - The result obtained by solving the optimization problem is called hard margin model of support vector machine.



20. Support vector machines

- When some data cannot be correctly classified by a linear boundary
  - Loosen the classification condition

$$y(i)(\boldsymbol{w}^T\boldsymbol{x}(i) - b) \geq 1$$

to

$$y(i)(\boldsymbol{w}^T\boldsymbol{x}(i) - b) \geq 1 - \varepsilon_i, \quad \varepsilon_i > 0.$$

A large ε_i allows more mis-classification of data. Therefore a smaller ε_i is more preferable. We can formulate the following problem:

$$\text{Minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \varepsilon_i,\ C > 0$$

$$\text{subject to } y(i)(\boldsymbol{w}^T\boldsymbol{x}(i) - b) \geq 1 - \varepsilon,\ \varepsilon > 0,$$

which gives the soft margin model of support vector machine.

20. Support vector machines

- When data are not well separated by a linear boundary
  - When the dimension of the data vector is high, it is easy for the data to be separated by a linear boundary.
    If the dimension is larger then the number of data, the data can be always separated by a linear boundary.
    Increasing the dimensionality is a promising way to separate data.
  - For the above purpose, a nonlinear transform $\phi(\boldsymbol{x})$ is used to map the original data to a data vector of a higher dimension.
  - Design of a support vector machine is done based on the inner production of data vectors. When the inner product of mapped data vectors is directly calculated using data vectors themselves, i.e., if there is a function K such that
    $$\phi(\boldsymbol{x}_1)^T\phi(\boldsymbol{x}_2) = K(\boldsymbol{x}_1, \boldsymbol{x}_2)$$
    then the design becomes easier.

20. Support vector machines

- Function K is called kernel.
  - A typical kernel is

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}\right).$$