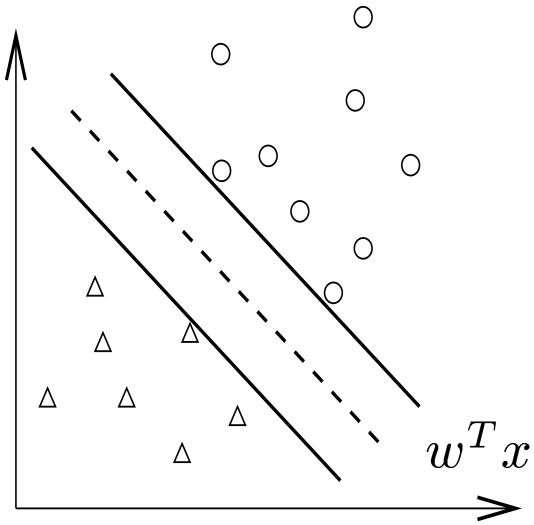
Support Vector Machines

- Training vectors : $x_i, i = 1, \dots, l$
- Consider a simple case with two classes:

  Define a vector y

$$y_i = \begin{cases} 1 & \text{if } x_i \text{ in class } 1 \\ -1 & \text{if } x_i \text{ in class } 2, \end{cases}$$

- A hyperplane which separates all data



- A separating hyperplane: $w^T x + b = 0$

$$
\begin{aligned}
(w^T x_i) + b > 0 \quad & \text{if } y_i = 1 \\
(w^T x_i) + b < 0 \quad & \text{if } y_i = -1
\end{aligned}
$$

- Decision function $f(x) = \text{sign}(w^T x + b)$, x: test data

  Variables: w and b : Need to know coefficients of a plane

  Many possible choices of w and b

- Select w, b with the maximal margin.
  Maximal distance between $w^T x + b = \pm 1$
  Vapnik's statistical learning theory. (will be discussed later)

$$\begin{aligned} (w^T x_i) + b &\geq 1 \quad \text{if } y_i = 1 \\ (w^T x_i) + b &\leq -1 \quad \text{if } y_i = -1 \end{aligned} \tag{1}$$

- Distance between $w^T x + b = 1$ and -1:

$$2/\|w\| = 2/\sqrt{w^T w}.$$

- $\max 2/\|w\| \equiv \min w^T w/2$

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ & y_i((w^T x_i) + b) \geq 1, \qquad \text{from (1)} \\ & i = 1, \ldots, l. \end{aligned}$$

Higher Dimensional Feature Spaces

- Earlier we tried to find a linear separating hyperplane

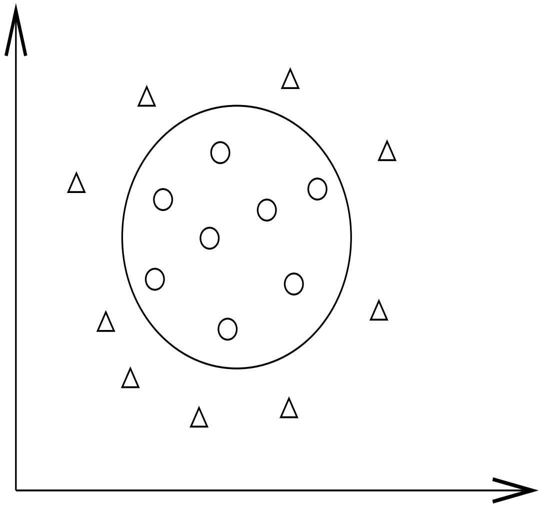
  Data may not be linear separable

- Non-separable case: allow training errors

$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C(\sum_{i=1}^{l} \xi_i) \\
& y_i((w^T x_i) + b) \geq 1 - \xi_i, \\
& \xi_i \geq 0,\ i = 1, \ldots, l
\end{aligned}
$$

- $\xi_i > 1$, x_i not on the correct side of the separating plane
- C: large penalty parameter, most ξ_i are zero

- Nonlinear case: linear separable in other spaces ?



- Higher dimensional (maybe infinite) feature space

$$\phi(x) = (\phi_1(x), \phi_2(x), \ldots).$$

- Example: $x \in R^3, \phi(x) \in R^{10}$

$$\begin{aligned}\phi(x) \quad = \quad & (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, \\ & x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)\end{aligned}$$

- Why higher dimensional spaces: a classic result by Cover [1965]
- A standard problem [Cortes and Vapnik, 1995]:

$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C\left(\sum_{i=1}^{l} \xi_i\right) \\
& y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\
& \xi_i \geq 0, \; i = 1, \ldots, l
\end{aligned}
$$

- Other variants (though similar); Example:

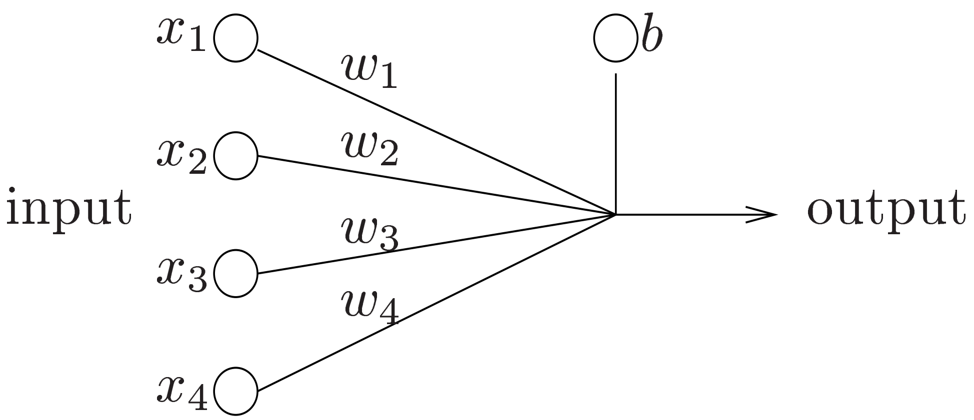
$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}w^T w + C\left(\sum_{i=1}^{l} \xi_i^2\right) \\
& y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\
& \xi_i \geq 0, \; i = 1, \ldots, l
\end{aligned}
$$

Neural Netrowks and Support Vector Machines

- Neural Networks:
- Starts from linear separating hyperplane as well

  Perceprton: a linear hyperplane separating all data

  Single-layer perceptron



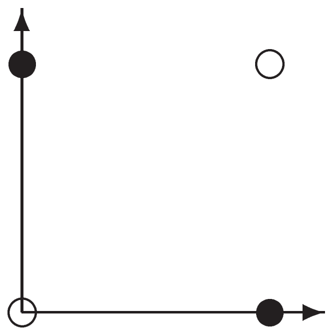
Decision function

$$\text{sgn}(w^T x + b)$$

- Data not linearly separable: multi-layer structure

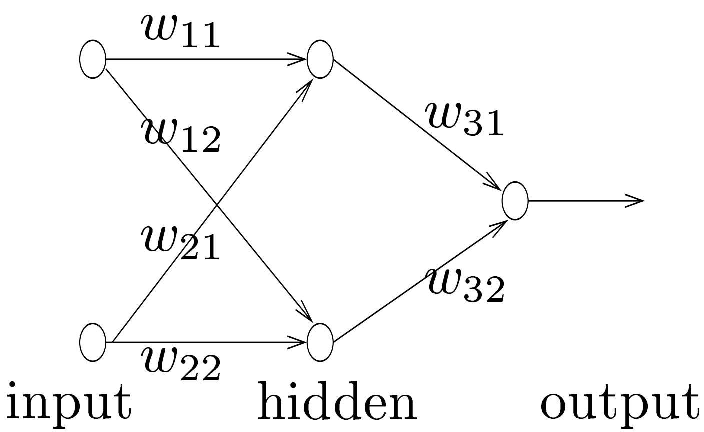
  Multi-layer perceptron

- XOR problem



  Not linearly separable

- More weights



Two exclusive OR

- Optimization problem

  Minimize training error

  Subject to connections between levels i and $i-1$

- Using more complicated structures for linearly non-separable data

- Starting here SVM differs from NN

Finding the Decision Function

- Finding w and b from the standard SVM form

  w: a vector in a high dimensional space $\Rightarrow$ maybe infinite variables

- The dual problem

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\
& 0 \le \alpha_i \le C, i = 1, \ldots, l \\
& y^T\alpha = 0,
\end{aligned}
$$

  where $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$ and $e = [1, \ldots, 1]^T$

$$w = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)$$

- Primal and dual : optimization theory. Not trivial.

  Infinite dimensional programming.

- A finite problem:

  #variables = #training data

- $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$ needs a closed form

  Efficient calculation of high dimensional inner products

- Example: $x_i \in R^3, \phi(x_i) \in R^{10}$

$$\begin{aligned}\phi(x_i) \quad &= \quad (1, \sqrt{2}(x_i)_1, \sqrt{2}(x_i)_2, \sqrt{2}(x_i)_3, (x_i)_1^2,\\ &\quad (x_i)_2^2, (x_i)_3^2, \sqrt{2}(x_i)_1(x_i)_2, \sqrt{2}(x_i)_1(x_i)_3, \sqrt{2}(x_i)_2(x_i)_3),\end{aligned}$$

Then $\phi(x_i)^T \phi(x_j) = (1 + x_i^T x_j)^2$.

- Popular methods: $\phi(x_i)^T\phi(x_j) =$

$$e^{-\gamma\|x_i - x_j\|^2}, \text{ (Radial Basis Function)}$$

$$(x_i^T x_j/a + b)^d \text{ (Polynomial kernel)}$$

$$\tanh(a x_i^T x_j + b)$$

- Decision function:

$$\begin{aligned} & w^T\phi(x) + b \\ = & \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)^T\phi(x) + b \end{aligned}$$

No need to have w

- > 0: 1st class, < 0: 2nd class
- Only $\phi(x_i)$ of $\alpha_i > 0$ used

$$\alpha_i > 0 \Rightarrow \text{ support vectors}$$

Support Vectors: More Important Data