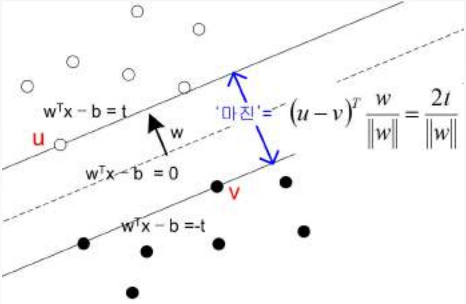
Hyperplane classifier

Given two sets $\{u_i|i \in U\}$ and $\{v_i|i \in V\}$, find a hyperplane (w, b) that separates two sets by a 'maximum' margin.



Suppose the supporting hyperplanes are $w^T x = b - t$ and $w^T x = b + t$, and the support vectors are $u \in U$, $v \in V$. Then the margin is $(v - u)^T \frac{w}{\|w\|} = 2\frac{t}{\|w\|}$.

If we assume $t > 0$, then the problem of finding hyperplane with a largest margin becomes a QP:

$$\begin{array}{rlll} & \max & \frac{t}{\|w\|_2} & (\Leftrightarrow \min \frac{\|w\|_2}{t}) \\ & \text{s.t.} & u_i^T w - b & \geq +t,\ i \in U, \\ & & v_i^T w - b & \leq -t,\ i \in V, \\ \Leftrightarrow & \min & \|w\|_2^2 & \\ & \text{s.t.} & u_i^T w - b & \geq +1,\ i \in U, \\ & & v_i^T w - b & \geq -1,\ i \in V. \end{array}$$

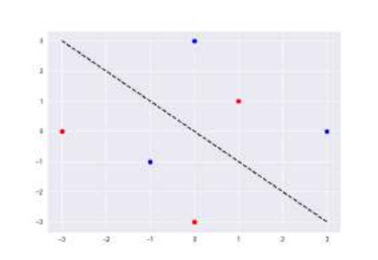
For a general case when two sets U and V are not separable by a hyperplane, we can allow error $\xi_i \geq 0$ for each i and add a total error penalty to the objective function:

$$\begin{array}{lll} \min & \|w\|_2^2 + \gamma \sum_i \xi_i & \\ \text{s.t.} & u_i^T w - b \geq +1 - \xi_i, & i \in U, \\ & v_i^T w - b \leq -1 + \xi_i, & i \in V, \\ & \xi_i \geq 0, & i \in U, V. \end{array}$$

Example 12.2

Suppose $U = \{(3,0), (0,3), (-1,-1)\}$, $V = \{(-3,0), (0,-3), (1,1)\}$ and we have chosen $\gamma = 1$.

$$
\begin{array}{lrl}
\min & w_1^2 + w_2^2 + \sum_{i=1}^{6} \xi_i & \\
\text{s.t} & (3w_1 - b) - 1 + \xi_1 & \geq 0, \\
& (3w_2 - b) - 1 + \xi_2 & \geq 0, \\
& (-w_1 - w_2 - b) - 1 + \xi_3 & \geq 0, \\
& -(-3w_1 - b - b) - 1 + \xi_4 & \geq 0, \\
& -(-3w_2 - b) - 1 + \xi_5 & \geq 0, \\
& -(w_1 + w_2 - b) - 1 + \xi_6 & \geq 0, \\
& \xi \geq 0, \forall i = 1, \cdots, 6. &
\end{array}
$$



$$\min \quad f(x) \tag{13.23}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable on an open domain dom$f$.

Assumption 13.1

There exists an optimal point x^* such that $p^* = f(x^*) = \inf_x f(x)$.

Since f is differentiable and convex, x^* is optimal if and only if

$$\nabla f(x^*) = 0. \tag{13.24}$$

Thus, solving (13.23) is the same as finding a solution of (13.24), a set of n equations in n variables $x_1, \ldots, x_n$.

- We can find a solution of (13.23) by either solving (13.24) analytically, or using an iterative method computing a sequence of points $x^{(0)}, x^{(1)}, \cdots \in$ dom$f$ with

$$f(x^{(k)}) \to p^* \ \text{ as } k \to \infty.$$

- An iterative algorithm normally requires a suitable starting point $x^{(0)}$ such that $x^{(0)} \in \mathrm{dom} f$, and $S = \{x \in \mathrm{dom} f | f(x) \leq f(x^{(0)})\}$ is closed.

Example 13.2

$$\min \quad \tfrac{1}{2}x^T P x + q^T x + r, \tag{13.25}$$

where P is a PSD matrix, $q \in \mathbb{R}^n$, and $r \in \mathbb{R}$.

- Any x^* satisfying $Px^* = -q$ is an optimal solution.
- If P is invertible, $x^* = -P^{-1}q$ is a unique optimal solution.
- If $Px = -q$ does not have a solution, (13.25) is unbounded below.

Example 13.3

$$\min \quad \|Ax - b\|_2^2 = x^T(A^T A)x - 2(A^T b)^T x + b^T b. \tag{13.26}$$

The optimality conditions $A^T A x^* = A^T b$ are called the normal equations of the least-square problem.

In iterative algorithms, we generate a minimizing sequence $x^{(k)}$, $k = 1, 2, \dots$

$$x^{(k+1)} = x^{(k)} + \sigma^{(k)} d^{(k)},\ \sigma^{(k)} > 0,$$

where, $d^{(k)}$ is called *search direction* at iteration k, and $\sigma^{(k)}$ *step size* at iteration k.

In descent method, sequence $x^{(k)}$, $k = 1, 2, \dots$ satisfies

$$f(x^{(k+1)}) < f(x^{(k)}).$$

Proposition 14.1

If f is convex, a method is descent if and only $\nabla f(x^{(k)})^T d^{(k)} < 0$.

A natural choice is then $d^{(k)} = -\nabla f(x^{(k)})$.

- Compute an initial point $x^{(0)}$.
- Until a stopping criterion is satisfied, generate x^k $k = 1, 2, \ldots$:

$$x^{(k+1)} = x^{(k)} - \sigma^{(k)} \nabla f(x^{(k)}).$$

where, $\sigma^{(k)} > 0$ is called the *step size* at iteration k.

- $\sigma^{(k)} = \sigma > 0$ fixed.
- $\sigma^{(k)} = \arg\min_{\sigma > 0} f\left(x^{(k)} - \sigma \nabla f(x^{(k)})\right)$. Not practical!
- $\sigma^{(k)} = \frac{\sigma}{\sqrt{k+1}}$, for a constant $\sigma > 0$.
- In exact line search e.g. Goldstein-Armijo's rule.