収束定理

- **Theorem**: (Novikoff) Let S be a non-trivial training set, and let
$$R = \max_{1 \le i\ell} \|\boldsymbol{x}_i\|.$$
Suppose that there exists a vector $\boldsymbol{w}_{opt}$ such that $\|\boldsymbol{w}_{opt}\| = 1$ and
$$y_i(\langle \boldsymbol{w}_{opt}, \boldsymbol{x}_i \rangle + b_{opt}) \ge \gamma$$
for $1 \le i\ell$. Then the number of mistakes made by the on-line perceptron algorithm on S is at most
$$\left(\frac{2R}{\gamma}\right)^2.$$
- 間違いの回数に上限 $\Longrightarrow$ 有限回の繰り返しで収束
- 線形識別可能性を仮定
- 線形識別可能でなければ修正を繰り返し収束しない
- 線形識別可能性の判断には使えない
- 途中で打ち切った場合，そのときの重みが最適かどうかわからない

証明

Define $\hat{\boldsymbol{x}}_i = (\boldsymbol{x}_i^T, R)^T, \hat{\boldsymbol{w}} = (\boldsymbol{w}^T, b/R)^T$. Let $\hat{\boldsymbol{w}}_{t-1}$ be the augmented weight vector prior to the tth mistake. The tth update is performed when

$$y_i \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{x}}_i \rangle = y_i(\langle \boldsymbol{w}_{t-1}, \boldsymbol{x}_i \rangle + b_{t-1}) \leq 0$$

where $(\boldsymbol{x}_i, y_i)$ is the point incorrectly classified by $\hat{\boldsymbol{w}}_{t-1}$. The update is the following:

$$\hat{\boldsymbol{w}}_t = \begin{bmatrix} \boldsymbol{w}_t \\ b_t/R \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{t-1} \\ b_{t-1}/R \end{bmatrix} + \eta y_i \begin{bmatrix} \boldsymbol{x}_i \\ R \end{bmatrix} = \hat{\boldsymbol{w}}_{t-1} + \eta y_i \hat{\boldsymbol{x}}_i$$

where $b_t = b_{t-1} + \eta y_i R^2$. Since the margin is γ, we have

$$\langle \hat{\boldsymbol{w}}_t, \hat{\boldsymbol{w}}_{opt} \rangle = \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{w}}_{opt} \rangle + \eta y_i \langle \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{w}}_{opt} \rangle \geq \langle \hat{\boldsymbol{w}}_{i-1}, \hat{\boldsymbol{w}}_{opt} \rangle + \eta\gamma$$

and this implies that

$$\langle \hat{\boldsymbol{w}}_t, \hat{\boldsymbol{w}}_{opt} \rangle \geq t\eta\gamma$$

Similarly, we have

$$\begin{aligned} \|\hat{\boldsymbol{w}}_t\|^2 &= \|\hat{\boldsymbol{w}}_{t-1}\|^2 + 2\eta y_i \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{x}}_i \rangle + \eta^2 \|\hat{\boldsymbol{x}}_i\|^2 \\ &\geq \|\hat{\boldsymbol{w}}_{t-1}\|^2 + \eta^2 \|\hat{\boldsymbol{x}}_i\|^2 = \|\hat{\boldsymbol{w}}_{t-1}\|^2 + \eta^2(\|\boldsymbol{x}_i\|^2 + R^2), \end{aligned}$$

which implies $\|\hat{\boldsymbol{w}}_t\|^2 \geq \|\hat{\boldsymbol{w}}_{t-1}\|^2 + \eta^2 R^2$. Thus we have

$$\|\hat{\boldsymbol{w}}_t\|^2 \geq 2t\eta^2 R^2.$$

The two inequalities combined give the 'squeezing' relations

$$\|\boldsymbol{w}_{opt}\| \sqrt{2t\eta} R \geq \|\boldsymbol{w}_{opt}\| \|\boldsymbol{w}_t\| \geq \langle \hat{\boldsymbol{w}}_t, \hat{\boldsymbol{w}}_{opt} \rangle \geq t\eta\gamma,$$

which together imply the bound (note $\|\hat{\boldsymbol{w}}_{opt}\|^2 \leq \|\boldsymbol{w}_{opt}\|^2 + 1 = 2$)

$$t \leq 2\left(\frac{R}{\gamma}\right)^2 \|\hat{\boldsymbol{w}}_{opt}\|^2 \leq \left(\frac{2R}{\gamma}\right)^2.$$

- **Formulation**: (Optimization problem) Given functions f, g_i, $i = 1, \ldots, k$, and h_i, $i = 1, \ldots, m$, defined on a domain $\Omega \subseteq R^n$, optimization problem is formalized as follows:

$$\begin{array}{lll} \text{minimize} & f(\boldsymbol{w}), & \boldsymbol{w} \in \Omega \\ \text{subject to} & g_i(\boldsymbol{w}) \leq 0, & i = 1, \ldots, k \\ & h_i(\boldsymbol{w}) = 0, & i = 1, \ldots, m \end{array}$$

  where f is called the *objective function*, and the remaining relations are called, respectively, the *inequality* and *equality constraints*.

- **Definition**: (Convexity) A real-valued function $f(\boldsymbol{w})$ is called *convex* for $\boldsymbol{w} \in R^n$ if, $\forall \boldsymbol{w}, \boldsymbol{u} \in R^n$, and for any $\theta \in (0, 1)$,

$$f(\theta \boldsymbol{w} + (1 - \theta)\boldsymbol{u}) \leq \theta f(\boldsymbol{w}) + (1 - \theta) f(\boldsymbol{u})$$

- **Theorem**: (Fermat) A necessary condition for $\boldsymbol{w}^*$ to be a minimum of a function $f(\boldsymbol{w})$ is

$$\frac{\partial f(\boldsymbol{w}^*)}{\partial \boldsymbol{w}} = \mathbf{0}.$$

  This condition, together with convexity of f, is also sufficient.

- **Definition**: (Lagrangian) Given an optimization problem with objective function $f(\boldsymbol{w})$, and equality constraints $h_i(\boldsymbol{w}) = 0$, $i = 1, ..., m$, we define the *Lagrangian function* as

$$L(\boldsymbol{w}, \boldsymbol{\beta}) = f(\boldsymbol{w}) + \sum_{i=1}^{m} \beta_i h_i(\boldsymbol{w})$$

  where the coefficients β_i are called the *Lagrange multipliers*.

- **Theorem**: (Lagrange) A necessary condition for a point $\boldsymbol{w}^*$ to be a minimum of $f(\boldsymbol{w})$ subject to $h_i(\boldsymbol{w}) = 0$, $i = 1, \ldots, m$, is

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{w}} &= \mathbf{0} \\ \frac{\partial L(\boldsymbol{w}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} &= \mathbf{0} \end{aligned}$$

  for some values $\boldsymbol{\beta}^*$. The above conditions are also sufficient provided that $L(\boldsymbol{w}, \boldsymbol{\beta}^*)$ is a convex function of $\boldsymbol{w}$.

Kuhn-Tucker 法

- **Definition**: (generalized Lagrangian) Given an optimization problem with domain $\Omega \subseteq R^n$,

$$\begin{array}{lll} \text{minimize} & f(\boldsymbol{w}), & \boldsymbol{w} \in \Omega \\ \text{subject to} & g_i(\boldsymbol{w}) \leq 0, & i = 1, \ldots, k \\ & h_i(\boldsymbol{w}) = 0, & i = 1, \ldots, m \end{array}$$

we define the *generalized Lagrangian function* as

$$\begin{aligned} L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= f(\boldsymbol{w}) + \sum_{i=1}^{k} \alpha_i g_i(\boldsymbol{w}) + \sum_{i=1}^{m} \beta_i h_i(\boldsymbol{w}) \\ &= f(\boldsymbol{w}) + \boldsymbol{\alpha}^T \boldsymbol{g}(\boldsymbol{w}) + \boldsymbol{\beta}^T \boldsymbol{h}(\boldsymbol{w}) \end{aligned}$$

- **Theorem**: (Kuhn-Tucker) Sufficient conditions for a point $\boldsymbol{w}^*$ to be an optimum are the existence of $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ such that

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{w}} &= \mathbf{0}, \\ \frac{\partial L(\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} &= \mathbf{0}, \\ \alpha_i^* g_i(\boldsymbol{w}^*) &= 0, \quad i = 1, \ldots, k, \\ g_i(\boldsymbol{w}^*) &\leq 0, \quad i = 1, \ldots, k, \\ \alpha_i^* &\geq 0, \quad i = 1, \ldots, k. \end{aligned}$$

- 例題1

$$\begin{array}{ll} \text{maximize} & x + y \\ \text{subject to} & x^2 + y^2 \leq 1, \quad x \geq 0, \quad y \geq 0 \end{array}$$

- 例題2

$$\begin{array}{ll} \text{maximize} & (x-1)^2 + (y-1)^2 \\ \text{subject to} & x + 2y \leq 1, \quad x \geq 0, \quad y \geq 0 \end{array}$$

Kuhn-Tucker 法 (例題)

- 解答 1:

$$
\begin{aligned}
f(\boldsymbol{x}) &= -x-y \\
g_1(\boldsymbol{x}) &= x^2+y^2-1 \\
g_2(\boldsymbol{x}) &= -x \\
g_3(\boldsymbol{x}) &= -y
\end{aligned}
$$

とおく．また $D=\frac{\partial}{\partial \boldsymbol{x}}$ と書くことにすると，

$$
\begin{aligned}
Df(\boldsymbol{x}) &= (-1,-1) \\
Dg_1(\boldsymbol{x}) &= (2x,2y) \\
Dg_2(\boldsymbol{x}) &= (-1,0) \\
Dg_3(\boldsymbol{x}) &= (0,-1)
\end{aligned}
$$

となる．したがって最適解は

$$
\begin{aligned}
0 &= (-1,-1)+\alpha_1(2x,2y)+\alpha_2(-1,0)+\alpha_3(0,-1) \\
0 &= \alpha_1(x^2+y^2-1) \\
0 &= \alpha_2(-x) \\
0 &= \alpha_3(-y) \\
\alpha_i &\geq 0
\end{aligned}
$$

をみたす.

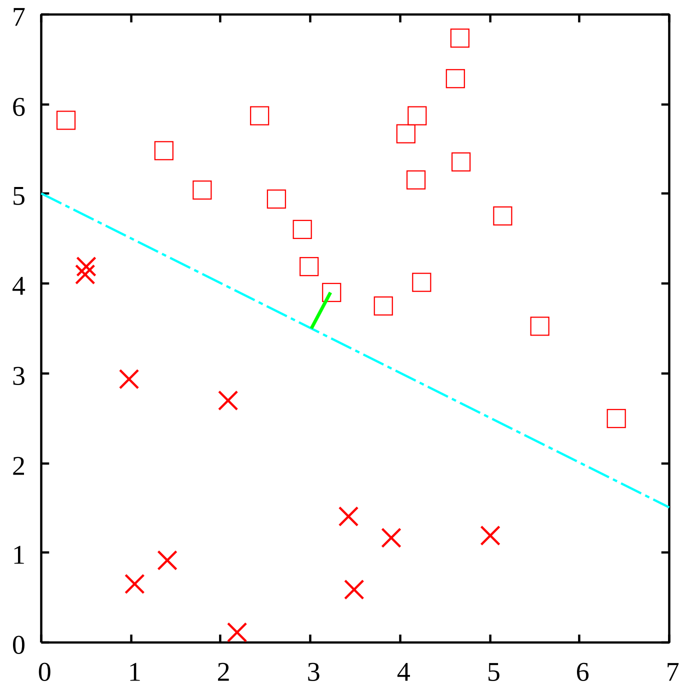
1. $x>0, y>0$ のとき．$\alpha_2=\alpha_3=0$ であるから $2\alpha_1 x = 2\alpha_1 y = 1$．これを第 2 式に代入し，$\alpha_1>0$ を考慮すると $\alpha_1=1/\sqrt{2}, x=1/\sqrt{2}, y=1/\sqrt{2}$ を得る．
2. $x=0$ のとき，第 1 式より $\alpha_2=-1$ となり不適.
3. $y=0$ のとき，第 1 式より $\alpha_3=-1$ となり不適.

したがって，求める解は $(x,y)=(1/\sqrt{2},1/\sqrt{2})$.

- 解答 2: 省略．求める解は $(x,y)=(0,0)$.

マージン

- Functional margin: $\gamma = y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)$



- Scale $\boldsymbol{w}$ and b so that: $y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 \quad \forall i$
- Support vectors: $\boldsymbol{x}^+, \boldsymbol{x}^-$

$$\langle \boldsymbol{w}, \boldsymbol{x}^+ \rangle + b = 1, \quad \langle \boldsymbol{w}, \boldsymbol{x}^- \rangle + b = -1$$

- Geometric margin: d

$$\begin{aligned} d &= \frac{1}{2} \left(\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{x}^+ \right\rangle \right) - \left(\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{x}^- \right\rangle \right) \\ &= \frac{1}{2\|\boldsymbol{w}\|} \left(\langle \boldsymbol{w}, \boldsymbol{x}^+ \rangle \right) - \left(\langle \boldsymbol{w}, \boldsymbol{x}^- \rangle \right) = \frac{1}{\|\boldsymbol{w}\|} \end{aligned}$$

最大マージン識別器 (Primal form)

Proposition: Given a linearly separable training sample

$$S = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell))$$

the hyperplane $(\boldsymbol{w}, b)$ that solves the optimization problem

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{w},b} & \langle \boldsymbol{w}, \boldsymbol{w} \rangle \\ \text{subject to} & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad \text{for} \quad i = 1, \ldots, \ell \end{array}$$

realizes the maximal margin hyperplane with geometric margin $\gamma = 1/\|\boldsymbol{w}\|$.

- Lagrangian

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle - \sum_{i=1}^{\ell} \alpha_i [y_i(\langle \boldsymbol{w}_i, \boldsymbol{x}_i \rangle + b) - 1]$$

where $\alpha_i \geq 0$ are Lagrange multipliers. Imposing stationarity condition, we have

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}, b, \alpha)}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{\ell} y_i \alpha_i \boldsymbol{x}_i = \mathbf{0}, \\ \frac{\partial L(\boldsymbol{w}, b, \alpha)}{\partial b} &= \sum_{i=1}^{\ell} y_i \alpha_i = 0. \end{aligned} \tag{1}$$

Substituting these into the primal to obtain

$$L(\boldsymbol{w}, b, \alpha) = \sum_{j=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} y_i \alpha_i \sum_{j=1}^{\ell} y_j \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

最大マージン識別器 (Dual form)

Proposition: Given a linearly separable training sample

$$S = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell))$$

and suppose the parameters $\boldsymbol{\alpha}^*$ solve the following quadratic optimization problem:

$$\begin{aligned}
\text{maximize} \quad & W(\boldsymbol{\alpha}) = \sum_{j=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\
\text{subject to} \quad & \sum_{i=1}^{\ell} y_i \alpha_i = 0, \\
& \alpha_i \geq 0 \ \text{ for } \ i = 1, \ldots, \ell
\end{aligned}$$

Then the weight vector $\boldsymbol{w} = \sum_{i=1}^{\ell} y_i \alpha_i^* \boldsymbol{x}_i$ realizes the maximal margin hyperplane with geometric margin $\gamma = 1/\|\boldsymbol{w}^*\|$.

- **Remark 1**: The value of b

$$b^* = -\frac{1}{2}\left(\max_{y_i=-1}(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle) + \min_{y_i=1}(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle)\right)$$

- **Remark 2**: Karush-Kuhn-Tucker conditions state that the optimal solutions $\boldsymbol{\alpha}^*, (\boldsymbol{w}^*, b^*)$ must satisfy

$$\alpha_i^*[y_i(\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle + b^*) - 1] = 0$$

Only for inputs $\boldsymbol{x}_i$ for which the functional margin is one (and therefore lie closest to the hyperplane), the corresponding α_i^* are non-zero. All the other parameters α_i^* are zero.

- **Remark 3**: The optimal hyperplane can be expressed in terms of **support vectors**

$$f(\boldsymbol{x}, \boldsymbol{\alpha}^*, b^*) = \sum_{i=1}^{\ell} y_i \alpha_i \langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + b^* = \sum_{i \in \mathrm{SV}} y_i \alpha_i \langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + b^*$$

Points that are not support vectors have no influence.

- **Remark 4**: Another important consequence of the Karush-Kuhn-Tucker complementarity condition is that for $j \in \mathrm{sv}$,

$$y_j f(\boldsymbol{x}_j, \boldsymbol{\alpha}^*, b^*) = y_j \left(\sum_{i \in \mathrm{SV}} y_i \alpha_i \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle + b^* \right) = 1,$$

and therefore

$$\begin{aligned}
\langle \boldsymbol{w}^*, \boldsymbol{w}^*\rangle &= \sum_{i,j=1}^{\ell} y_i y_j \alpha_i^* \alpha_j^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \\
&= \sum_{j \in \mathrm{SV}} \alpha_j^* y_j \sum_{i \in \mathrm{SV}} y_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \\
&= \sum_{j \in \mathrm{SV}} \alpha_j^* (1 - y_j b^*) \\
&= \sum_{i \in \mathrm{SV}} \alpha_i^* \qquad (2)
\end{aligned}$$

i.e.,

$$\gamma = 1/\|\boldsymbol{w}^*\| = \left(\sum_{i \in \mathrm{SV}} \alpha_i^* \right)^{1/2}$$