Perceptrons

- linear discriminant functions for classification

. linear discriminant functions:

$$g(\underline{x}) = w_0 + \sum_{i=1}^{d} w_i x_i$$

Let

$\underline{x} = [1, x_1, \cdots, x_d]^T$ and $\underline{w} = [w_0, w_1, \cdots, w_d]^T$.

Then,

$g(\underline{x}) = \underline{w}^T \underline{x}$.

. binary classification:

two classes, c_1 and c_2.

if $\underline{w}^T \underline{x}_i > 0$, output of classifier = c_1

otherwise, output of classifier = c_2

if $\underline{x}_i$ belongs to c_1, $\underline{w}^T \underline{x}_i > 0$.

if $\underline{x}_i$ belongs to c_2, $\underline{w}^T \underline{x}_i < 0$ or $\underline{w}^T (-\underline{x}_i) > 0$.

So, let's change $\underline{x}_i$ into $-\underline{x}_i$, if $\underline{x}_i$ belongs to c_2.

Then,

$\underline{w}^T \underline{x}_i > 0$ if $\underline{x}_i$s are classified correctly.

. Perceptron criterion function:

$$J_P(\underline{w}) = \sum_{\underline{x} \in X(\underline{w})} -\underline{w}^T \underline{x}$$

where

$X(\underline{w})$ represents a set of samples misclassified by $\underline{w}$.

. gradient descent algorithm

The weights are updated as follows:

$$\underline{w}_{k+1} = \underline{w}_k - \rho_k \nabla J_P(\underline{w}_k)$$

where ρ_k represents a positive scaling factor (learning rate) and

$$\nabla J_P(\underline{w}_k) = \frac{\partial J_P}{\partial \underline{w}}\Big|_{\underline{w} = \underline{w}_k}.$$

(1) batch mode

$$\underline{w}_{k+1} = \underline{w}_k - \rho_k \sum_{\underline{x} \in X_k} \underline{x}$$

where X_k represents a set of samples misclassified by $\underline{w}_k$.

(2) on-line mode (or incremental mode)

Training samples $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$ are cyclically applied.

Training Rule:

1) arbitrary set w_1

2) $\underline{w}_{k+1} = \underline{w}_k - \rho_k \underline{x}_k$ for $k \geqq 1$

where $\underline{w}_k^T \underline{x}_k \leqq 0$

cf. $\rho_k = 1$ in Rosenblatt's Perceptron.

. convergence of on-line mode

Let w^* is a solution vector. Then,

$$w_{k+1} - w^* = w_k - w^* + \rho_k x_k \quad \text{and}$$

$$\| w_{k+1} - w^* \|^2 = \| w_k - w^* \|^2 + 2\rho_k (w_k - w^*)^T x_k + \rho_k^2 \| x_k \|^2.$$

Since $w_k^T x_k \leqq 0$,

$$\| w_{k+1} - w^* \|^2 \leqq \| w_k - w^* \|^2 - 2\rho_k w^{*T} x_k + \rho_k^2 \| x_k \|^2.$$

The term $-2\rho_k w^{*T} x_k + \rho_k^2 \| x_k \|^2$ is negative when

$$0 < \rho_k < \frac{2w_k^T x_k}{\| x_k \|^2}. \text{ (convergence condition for } \rho_k)$$

Therefore, the optimal learning rate is $\rho_k^* = \dfrac{w_k^{*T} x_k}{\| x_k \|^2}$.

By substituting the optimal learning rate, we get

$$\| w_{k+1} - w^* \|^2 \leqq \| w_k - w^* \|^2 - \frac{(w^{*T} x_k)^2}{\| x_k \|^2}.$$

. The upper bound of the number of updates

Let $\beta^2 = \max_i \| x_i \|^2$ and $\gamma = \min_i w^{*T} x_i$.

Then, after k updates,

$$\| w_{k+1} - w^* \|^2 \leqq \| w_1 - w^* \|^2 - k\frac{\gamma^2}{\beta^2}.$$

If $w_1 = 0$, k can be bounded by

$$k_0 = \frac{\beta^2 \| w^* \|^2}{\gamma^2}.$$

That is, Perceptron converges within the finite number of steps.

. adaptation of ρ_k

It can be shown if the samples are linearly separable and ρ_k satisfies the following conditions:

$$\lim_{k \to \infty} \rho_k = 0,$$

$$\lim_{m \to \infty} \sum_{k=1}^{m} \rho_k = \infty, \text{ and } \lim_{m \to \infty} \sum_{k=1}^{\infty} \rho_k^2 < \infty,$$

w_k converges to a solution vector satisfying $w^T x_i > 0 \ \ \forall i$.

example.

$$\rho_k = \frac{1}{k}$$

. optimal choice of ρ_k

The weight update rule:

$$w_{k+1} = w_k - \rho_k \nabla J(w_k).$$

Taylor series expansion of $J(w)$ around w_k:

$$J(w) \approx J(w_k) + \nabla J^T(w - w_k) + \frac{1}{2}(w - w_k)^T H(w - w_k)$$

where $H = \frac{\partial^2 J}{\partial w_i \partial w_j}\Big|_{w = w_k}$.

Then, $J(w_{k+1}) \approx J(w_k) - \rho_k \| \nabla J \|^2 + \frac{1}{2}\rho_k^2 \nabla J^T H \nabla J$.

Therefore, the optimal choice of ρ_k is $\rho_k = \frac{\| \nabla J \|^2}{\nabla J^T H \nabla J}$.

Reference: Pattern Classification, and Scene Analysis, chapter 5.