9. Perception

Signed distance of $x$ to hyperplane

$$w(x - x_0) = \frac{w^T x - w^T x_0}{||w||} = \frac{wx + w_0}{||w||}$$

For class $C_1$, $w^Tx + w_0 > 0$ for all points $x_i$ in $C_1$ that are correctly classified.

For class $C_2$, $w^Tx + w_0 < 0$ for all points $x_i$ in $C_2$ that are correctly classified.

For correctly classified points,
$$y_i (w^T x_i + w_0) > 0$$

For misclassified points,
$$y_i (w^T x_i + w_0) < 0$$

Perception criterion:
$$\text{minimize} \ D(w, w_0) = - \sum_{i \in M} y_i (w^T x_i + w_0)$$

$M$ index the misclassified points.

$$\nabla_w D(w, w_0) = - \sum_{i \in M} y_i x_i = - \left( \sum_{i \in C_1} x_i - \sum_{i \in C_2} x_i \right)$$

$$\nabla_{w_0} D(w, w_0) = - \sum_{i \in M} y_i$$

$$= -(N_1 - N_2) \text{ where } N_i \text{ is the number of misclassified points in } C_i$$

Perception Update Rule

$$\begin{bmatrix} w \\ w_0 \end{bmatrix} \leftarrow \begin{bmatrix} w \\ w_0 \end{bmatrix} + \eta \begin{bmatrix} y_i x_i \\ y_i \end{bmatrix}$$

Stochastic Gradient Descent — go in the direction of negative “gradient”, but only contribution to gradient by $x_i$ (step is taken after each $x_i$ is visited).
Perceptron pseudo-code

Repeat
    for \( i = 1 \) to \( n \) do
        if \( y_i (w \cdot x_i + w_0) < 0 \) then
            \( w \leftarrow w + \eta y_i x_i \)
            \( w_0 \leftarrow w_0 + \eta y_i \)
        end if
    end for
    Until there are no mistakes (misclassified) within for loop

Perceptron algorithm is guaranteed to find a \( w \) separating hyperplane if the data is linearly separable.

Drawbacks of Perceptron
1. If data is linearly separable, the hyperplane output depends on the order in which points are presented to the algorithm.
2. No. of iterations might be large
3. If classes are not linearly separable, then algorithm will not converge — cycles can develop that are not easy to detect.