1 defination

Perceptron is one of supervised learning algorithms which learns a binary classifier based on a linear predictor function: a function that maps its input $\vec{x}$ to an out put value

$$f(x) = sgn(\vec{w} * \vec{x} + b)$$

where $\vec{w}$ is a vector of real-valued weights, and b is the *bias*, in addition, $\vec{w} * \vec{x} + b = 0$ is called decision boundary. This algorithm will not terminate if the training set is not linearly separable, that is to say, there exists at least one line which separates positive dots and negative dots, as we will see in next part. One example of this is the Boolean exclusive-or problem.

2 algorithm

Our goal is to minimize the sum of distance of all dots that are misclassified:

$$L(\vec{x}) = -\sum_{x_i \in M} y_i(\vec{w} * \vec{x}_i + b)$$

Note that we omitted a constant $\frac{1}{||\vec{w}||_2}$ for we know eventually $L(\vec{x})$ will be zero if the training set is linearly separable. And then we know whether a stochastic gradient descent method or Lagrange dual method could be appllied to fit these perameters.

Algorithm 1 stochastic method

Input: training set, learning rate η **Output:** $\vec{w}, b$

$\vec{w} \leftarrow \vec{w}_0$
$b \leftarrow b_0$
while exists x_i st. $f(x_i) \neq y_i$ **do**
    $\vec{w} \leftarrow \vec{w} + \eta y_i x_i$
    $b \leftarrow b + \eta y_i$
end while

Algorithm 2 Lagrange dual method

Input: training set, learning rate η **Output:** $\vec{w}, b$

$\vec{\alpha} \leftarrow \vec{\alpha}_0$
$b \leftarrow b_0$
while existsx_i st. $f(x_i) \neq y_i$ **do**
    $\vec{\alpha} \leftarrow \vec{\alpha} + \eta$
    $b \leftarrow b + \eta y_i$
end while

In fact, I do not think this could serve as a typical example of Lagrange duality. First, if we define α_i to be the times we misclassified x_i, then it is not differentiatable, and again, we do not get rid of b, which equals $\sum_{i=1}^{N} \alpha_i * y_i$.

3 convergency

Theorem 1 *Suppose our training set is linearly separable, so that:*
there exists one hyperplain w_{opt} *that corrcetly separates our data set and satisfies* $\|w_{opt}\|_2 = 1$*, and*

that there exists $\gamma > 0$ *that all distance to this hyperplain is greater than* γ
the time that we misclassify is less than $(max\| \begin{pmatrix} x_i \\ 1 \end{pmatrix} \|)^2/\gamma^2$

Proof 3.1 *t.b.c.*