3 Understanding Linear Separability [30 pts]

Definition 1 (Linear Program) *A* linear program *can be stated as follows:*

Let A be an $m \times n$ real-valued matrix, $\vec{b} \in \mathbb{R}^m$, and $\vec{c} \in \mathbb{R}^n$. Find a $\vec{t} \in \mathbb{R}^n$, that minimizes the linear function

$$\begin{aligned} & z(\vec{t}) = \vec{c}^{\,T}\vec{t} \\ \textit{subject to} \quad & A\vec{t} \geq \vec{b} \end{aligned}$$

In the linear programming terminology, $\vec{c}$ is often referred to as the *cost vector* and $z(\vec{t})$ is referred to as the *objective function.*[1] We can use this framework to define the problem of learning a linear discriminant function.[2]

The Learning Problem:[3] Let $\vec{x_1}, \vec{x_2}, \ldots, \vec{x_m}$ represent m samples, where each sample $\vec{x_i} \in \mathbb{R}^n$ is an n-dimensional vector, and $\vec{y} \in \{-1, 1\}^m$ is an $m \times 1$ vector representing the respective labels of each of the m samples. Let $\vec{w} \in \mathbb{R}^n$ be an $n \times 1$ vector representing the weights of the linear discriminant function, and θ be the threshold value.

We *predict* $\vec{x_i}$ to be a *positive* example if $\vec{w}^T\vec{x_i} + \theta \geq 0$. On the other hand, we *predict* $\vec{x_i}$ to be a *negative* example if $\vec{w}^T\vec{x_i} + \theta < 0$.

We hope that the learned linear function can separate the data set. That is,

$$y_i = \begin{cases} 1 & \text{if } \vec{w}^T\vec{x_i} + \theta \geq 0 \\ -1 & \text{if } \vec{w}^T\vec{x_i} + \theta < 0. \end{cases} \tag{1}$$

In order to find a good linear separator, we propose the following linear program:

$$\begin{aligned} \min \quad & \delta \\ \text{subject to} \quad & y_i(\vec{w}^T\vec{x_i} + \theta) \geq 1 - \delta, \qquad \forall(\vec{x_i}, y_i) \in D \\ & \delta \geq 0 \end{aligned} \tag{2}$$

where D is the data set of all training examples.

[1] Note that you don't need to really know how to solve a linear program since we will use Matlab to obtain the solution (although you may wish to brush up on Linear Algebra). Also see the appendix for more information on linear programming.

[2] This discussion closely parallels the linear programming representation found in *Pattern Classification*, by Duda, Hart, and Stork.

[3] Note that the notation used in the Learning Problem is **unrelated** to the one used in the Linear Program definition. You may want to figure out the correspondence.

(a) **(15 pts)** A data set $D = \{(\vec{x}_i, y_i)\}_{i=1}^m$ that satisfies condition (1) above is called *linearly separable.* Show that there is an optimal solution with $\delta = 0$, then D is linearly separable.

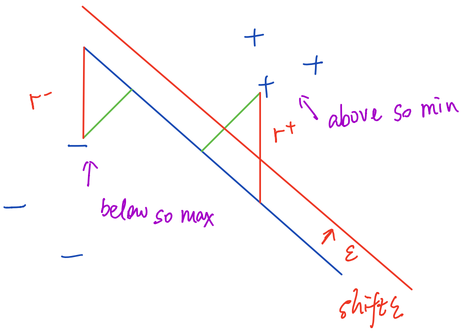
(b) **(5 pts)** What can we say about the linear separability of the data set if there exists a hyperplane that satisfies condition (2) with $\delta > 0$?

(c) **(5 pts)** An alternative LP formulation to (2) may be

$$\begin{array}{rll} \min & \delta & \\ \text{subject to} & y_i(\vec{w}^T\vec{x}_i + \theta) \geq -\delta, & \forall(\vec{x}_i, y_i) \in D \\ & \delta \geq 0 & \end{array}$$

Find the optimal solution to this formulation (independent of D) to illustrate the issue with such a formulation.

(a) when there is an optimal solution, $\delta = 0$.



Since r^+ is above the hyperplane, so it takes the min of $y_i = W^Tx_i + b$, where label $y = 1$

$$r^+ = \min_{\substack{(x_i, y_i)\in D \\ y=1}} W^Tx_i + b$$

Since r^- is below the hyperplane, so it takes the max of $y_j = W^Tx_j + b$, wher label $y = -1$.

$$r^- = \max_{\substack{(x_j, y_j)\in D \\ y=-1}} W^Tx_j + b$$

By defn of Linear Separability:

$$r^+ \geq 0 > r^-$$

In order to guarantee there is a minimum distance. Since $r^+ \geq 0 > r^-$, Let $\varepsilon \geq 0$ so $r^+ - \varepsilon \geq 0 > r^- - \varepsilon$. Now the new hyperplane $W^Tx + b - \varepsilon$ should also separate the data, and x_i and x_j has the same distance to $w^Tx + b - \varepsilon$.

To calculte the distance , use projection :

$$\text{Dist} = \frac{|W^T x_i + b - \varepsilon|}{||w||} = \frac{|w^T x_j^- + b - \varepsilon|}{||w||}$$

$$W^T x_i + b - \varepsilon = -(W^T x_j^- + b - \varepsilon)$$ since $W^T x_j^- + b - \varepsilon < 0$.

$$r^+ - \varepsilon = -(r^- - \varepsilon)$$ plug in r^+, r^-.

$$\varepsilon = \frac{r^+ - r^-}{2}.$$

Then $r^+_{new} = r^+ - \frac{r^+ - r^-}{2} = \frac{r^+ + r^-}{2}$ $= \min\limits_{\substack{(x_i, y_i) \in D \\ y=1}} W^T x_i + b - \varepsilon$

$r^-_{new} = r^- - \frac{r^+ - r^-}{2} = \frac{r^- - r^+}{2}$ $= \max\limits_{\substack{(x_j^-, y_j) \in D \\ y=-1}} W^T x_j^- + b - \varepsilon$

Then : $y(W^T x + b - \varepsilon) \geq \frac{r^+ - r^-}{2}$ $\forall (x, y) \in D$ -

Given that $r^+ > r^-$, $\varepsilon = \frac{r^+ + r^-}{2}$, let $\varepsilon' = \frac{r^+ - r^-}{2}$, and set $w' = \frac{w}{\varepsilon'}$, $\theta = \frac{b - \varepsilon}{\varepsilon'}$ and $\delta = 0$.

We got: $y(W^T x + \theta) \geq 1 - \delta$, $\forall (x, y) \in D$.

If $\delta = 0$, $y(w^T x + \theta) \geq 1 - \delta$,

Then $y(W^T x + \theta) \geq 1$, $\forall (x, y) \in D$.

If Label $y = 1$ $1 \cdot (W^T x + \theta) \geq 1$

$(W^T X + \theta) \geq 1 > 0$ $\forall (x, y) \in D, y = 1$

If Label $y = -1$ $-1 \cdot (W^T x + \theta) \geq 1$

$(W^T X + \theta) \leq -1 < 0$ $\forall (x, y) \in D, y = -1$

By defn of Linear Separability, it 's separable when $\delta = 0$.

(b) **(5 pts)** What can we say about the linear separability of the data set if there exists a hyperplane that satisfies condition (2) with $\delta > 0$?

(c) **(5 pts)** An alternative LP formulation to (2) may be

$$\begin{array}{rll} \min & \delta & \\ \text{subject to} & y_i(\vec{w}^T\vec{x}_i + \theta) \geq -\delta, & \forall(\vec{x}_i, y_i) \in D \\ & \delta \geq 0 & \end{array}$$

Find the optimal solution to this formulation (independent of D) to illustrate the issue with such a formulation.

(b) When $\delta > 0$: Case 1: $1-\delta > 0$, as we proved before, $y(w^Tx+\theta) > 1-\delta > 0$, then the data is Linear Separable.

Case 2: $1-\delta \leq 0$, $\delta \geq 1$, $y(w^Tx+\theta) > 1-\delta$ and may become negative. Thus, the given argument cannot be applied, as long as δ equal to any number which is greater than 1.

(c) When $w=\theta=\delta=0$, $y_i(w^Tx_i+\theta) \geq -\delta$ satisfied, and $w \cdot \theta \cdot \delta$ also satisfy the given agreement at very beginning of this problem. However, $w^Tx+\theta$ where $w=\theta=0$ is not a hyperplane at all. Even though there may have other solutions, this issue somehow cannot be avoided. Therefore, we'd better not to use

$y_i(w^Tx_i+\theta) \geq -\delta$.

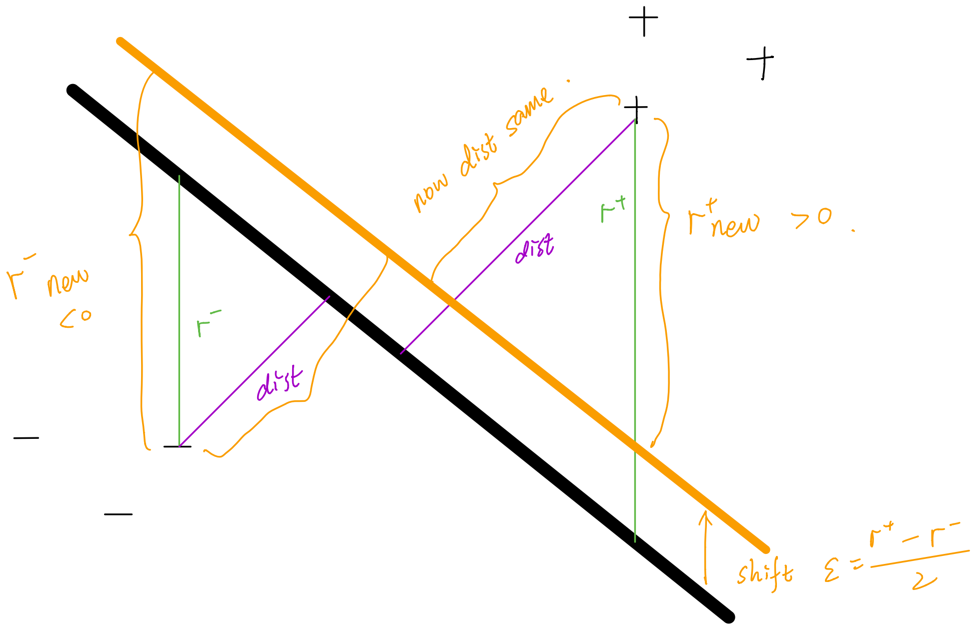
(d) $D = \{(x_1\ y_1), (x_2, y_2)\}$ $\quad \begin{cases} x_1^T = [1, 1, 1] & y_1 = 1 \\ x_2^T = [-1, -1, -1] & y_2 = -1 \end{cases}$

Since the given $(x_1 y_1)$ $(x_2\ y_2)$ is separable in the data set. So $\delta = 0$ for optimal solution.

$y_1(w^Tx_1+\theta) = w_1+w_2+w_3+\theta \geq 1$

$y_2(w^Tx_2+\theta) = -(-w_1-w_2-w_3+\theta) \geq 1$

$\therefore$ $w_1+w_2+w_3 \geq 1+|\theta|$, w, θ, δ is optimal.


+
+
+
now dist same .
r^+
dist
$r^+_{new} > 0$.
r^-_{new}
< 0
r^-
dist
−
−
−
shift $\varepsilon = \frac{r^+ - r^-}{2}$