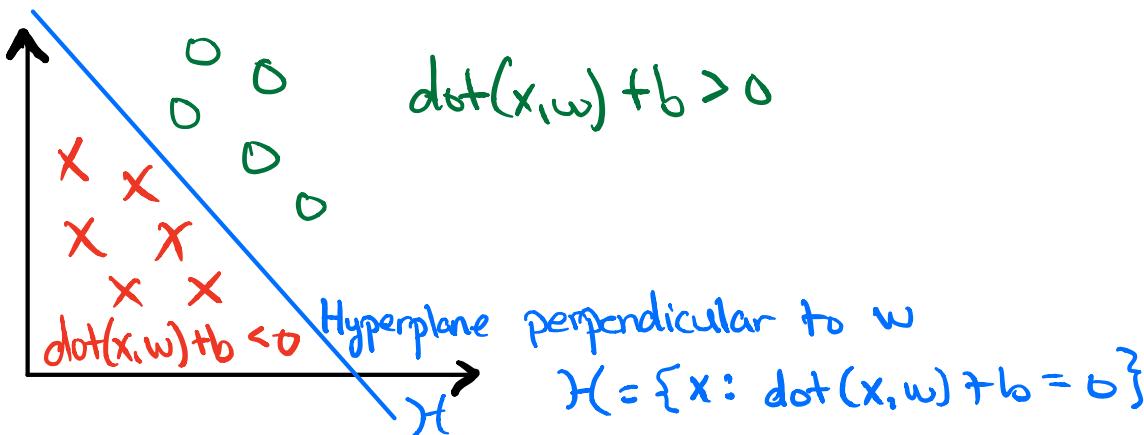


Assumptions:

- Binary classification (i.e. $y_i \in \{-1, 1\}$)
- Data is linearly separable

Classifier

$$h(x_i) = \text{sign}(\underline{w}^T x_i + b)$$



b is the bias term (w/o it the hyperplane that w defines would always have to go through the origin).

Dealing w/ b can be a PAIN!

So instead we 'absorb' it into the feature vector \underline{w} by adding one additional constant dimension.

Under this convention,

$$\begin{array}{l} \underline{x}_i \xrightarrow{\quad} \underline{x}_i \text{ becomes } \begin{bmatrix} \underline{x}_i \\ 1 \end{bmatrix} \\ \underline{w} \xrightarrow{\quad} \underline{w} \text{ becomes } \begin{bmatrix} \underline{w} \\ b \end{bmatrix} \end{array} \quad \left. \begin{array}{l} \text{what happened? No bias} \\ \text{term} \Rightarrow \text{hyperplane always} \\ \text{through origin!} \\ (\text{see figure}) \end{array} \right\}$$

Verify that

$$\begin{bmatrix} \underline{x}_i \\ 1 \end{bmatrix}^T \begin{bmatrix} \underline{w} \\ b \end{bmatrix} = \underline{w}^T \underline{x}_i + b$$

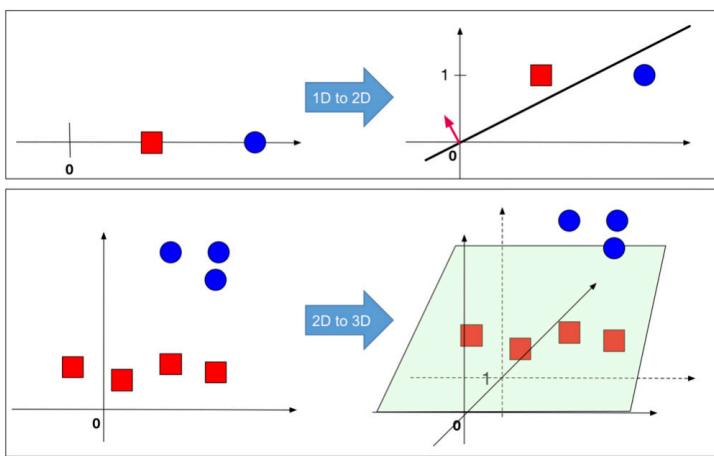
Using this, we can simplify the above formulation of $h(\underline{x}_i)$ to

$$h(\underline{x}_i) = \text{sign}(\bar{w}^T \bar{x})$$

think! This is intuitive!

if > 0 classify point as one label

if < 0 classify point as other label



(Left:) The original data is 1-dimensional (top row) or 2-dimensional (bottom row). There is no hyper-plane that passes through the origin and separates the red and blue points.

(Right:) After a constant dimension was added to all data points such a hyperplane exists.

Observation: $y_i (\bar{w}^T \bar{x}_i) > 0 \Leftrightarrow \underline{x}_i$ is classified correctly

where 'classified correctly' means that \underline{x}_i is on the correct side of the hyperplane defined by \bar{w} .

Also, note that the left side depends on $y_i \in \{-1, +1\}$
(would NOT work if $y_i \in \{0, +1\}$)

An aside: in higher dimensional spaces where points are very far apart (as seen in k-NN) this almost always holds

Perceptron Algorithm

Let's look at how we can get a hyperplane \vec{w} which separates the data

Initialize $\vec{w} = \vec{0}$ \leftarrow misclassifies EVERYTHING

while True do

$m = 0 \leftarrow$ # misclassifications

for $(x_i, y_i) \in \mathcal{D}$ do \leftarrow for (data, label) pair in dataset

if $y_i(\vec{w}_T \cdot \vec{x}_i) \leq 0$ then \leftarrow If pair is misclassified

$\begin{bmatrix} w \\ b \end{bmatrix} + y \begin{bmatrix} x_i \\ 1 \end{bmatrix} \quad \vec{w} \leftarrow \vec{w} + y \vec{x} \leftarrow$ update weight vector \vec{w}

$w + y x_i + b$ \leftarrow end if

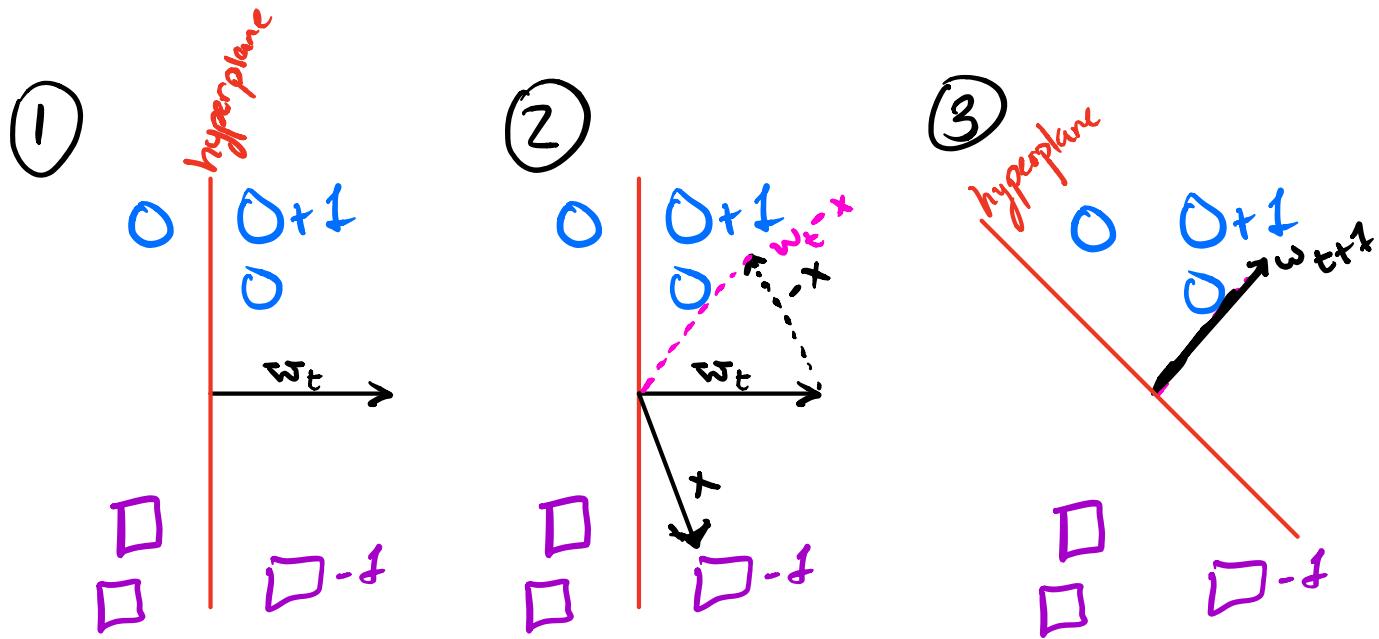
end for

if $m = 0$ then \leftarrow if no misclassifications, sufficient \vec{w}
break \leftarrow break out of while loop

end if

end while

Geometric Intuition



- ① Initially: hyperplane defined by w_t misclassifies one \square and one 0
- ② The misclassified \square (x) is chosen and used for an update.
Because its label is -1 we subtract x from w_t .
- ③ $w_{t+1} = w_t - x$ is updated as the new hyperplane; our algorithm has converged.

Perceptron Convergence

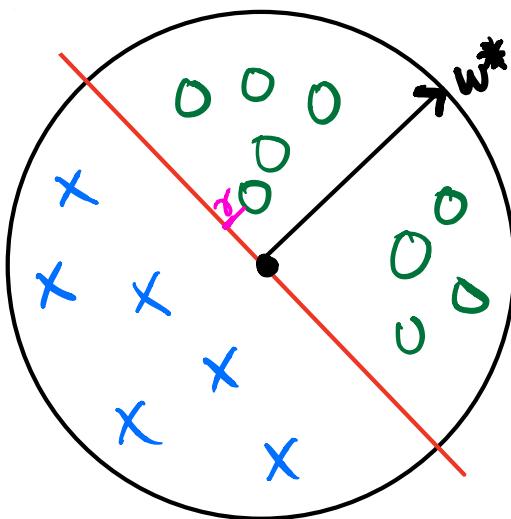
Suppose $\exists \underline{w}^*$ such that $y_i(\underline{x}^T \underline{w}^*) > 0 \nabla (x_i, y_i) \in D$

Now suppose we rescale each data point and \underline{w}^* such that

$$\|\underline{w}^*\| = 1 \quad \text{and} \quad \|\underline{x}_i\| \leq 1 \quad \forall x_i \in D$$

Define the margin γ of the hyperplane \underline{w}^* as

$$\gamma = \min_{(x_i, y_i) \in D} (|\underline{x}_i^T \underline{w}^*|)$$



Setup:

- All inputs \underline{x}_i live within the unit sphere
- $\exists \underline{w}^*$ w/ $\|\underline{w}^*\| = 1$ (*i.e.* \underline{w}^* lies exactly on unit sphere)
- γ is the distance from this **hyperplane** to the closest data point

Theorem: If all the above holds, then the perceptron make at **MOST** $1/\gamma^2$ mistakes.

Proof:

Consider the effect of an update (\underline{w} becomes $\underline{w} + y \underline{x}$) on the two terms $\underline{w}^T \underline{w}^*$ and $\underline{w}^T \underline{w}$

- $y(\underline{x}^T \underline{w}) \leq 0$: holds bc \underline{x} is misclassified by \underline{w}
(otherwise we wouldn't make the update)
- $y(\underline{x}^T \underline{w}^*) > 0$: Holds bc \underline{w}^* is a separating hyperplane and classifies all points correctly

① Consider the impact of an update on $\underline{w}^T \underline{w}^*$:

$$(\underline{w} + y \underline{x})^T \underline{w}^* = \underline{w}^T \underline{w}^* + y(\underline{x}^T \underline{w}^*) \geq \underline{w}^T \underline{w}^* + \gamma$$

γ follows from the fact that, for \underline{w}^* , the distance from the hyperplane defined by \underline{w}^* to \underline{x} must be AT LEAST γ (i.e. $y(\underline{x}^T \underline{w}^*) = |\underline{x}^T \underline{w}^*| \geq \gamma$)

So, for each update, $\underline{w}^T \underline{w}^*$ grows by AT LEAST γ

② Consider the impact of an update on $\underline{w}^T \underline{w}$:

$$\begin{aligned}
 (\underline{w} + y \underline{x})^T (\underline{w} + y \underline{x}) &= \underline{w}^T \underline{w} + \boxed{y \underline{w}^T \underline{x} + y \underline{x}^T \underline{w}} + \boxed{y^2 \underline{x}^T \underline{x}}
 \end{aligned}$$

$\geq y(\underline{w}^T \underline{x}) < 0 \quad 0 \leq \quad \leq$

$$\begin{aligned}
 &\leq \underline{w}^T \underline{w} + 1
 \end{aligned}$$

\leq follows from

- (i) $z_y(\underline{w}^T \underline{x}) < 0$ as we had to make an update $\Rightarrow x$ misclassified
- (ii) $0 \leq y^2(\underline{x}^T \underline{x}) \leq 1$ as $y^2 = 1$ and $\underline{x}^T \underline{x} \leq 1$ ($\|\underline{x}\| \leq 1$)

So, for each update $\underline{w}^T \underline{w}$ grows by AT MOST 1

③ Thus, after M updates, the following must hold:

- (i) $\underline{w}^T \underline{w}^* \geq M\gamma$
- (ii) $\underline{w}^T \underline{w} \leq M$

The proof follows

$$\begin{aligned} M\gamma &\leq \underline{w}^T \underline{w}^* && \text{from 3i} \\ &= \|\underline{w}\| \cos(\theta) && \text{definition of inner product} \\ &\leq \|\underline{w}\| && \cos \theta \leq 1 \\ &= \sqrt{\underline{w}^T \underline{w}} && \text{definition of } \|\underline{w}\| \\ &\leq \sqrt{M} && \text{from 3ii} \\ &\Rightarrow M\gamma \leq \sqrt{M} \\ M^2\gamma^2 &\leq M \\ M &\leq \frac{1}{\gamma^2} && \leftarrow M \text{ is bounded above by a constant} \end{aligned}$$