# ECE595 / STAT598: Machine Learning I
## Lecture 6.2: Linear Separability - Geometry of Discriminant Function

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

# Outline

**Goal**: Understand the geometry of linear separability.

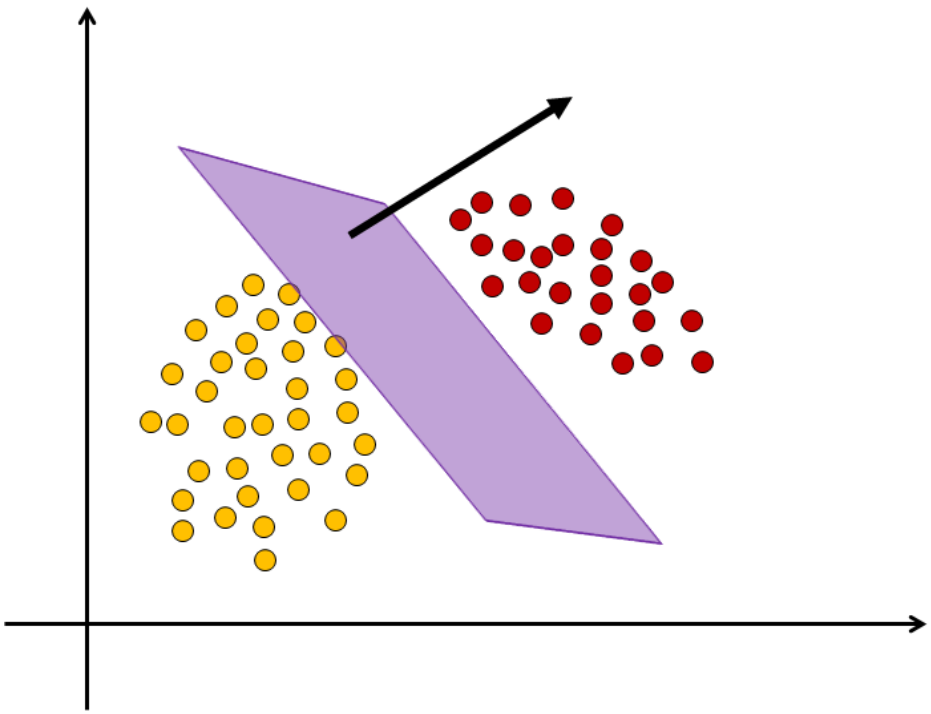
- Notations
  - Input Space, Output Space, Hypothesis
  - Discriminant Function
- Geometry of Discriminant Function
  - Separating Hyperplane
  - Normal Vector
  - Distance from Point to Plane
- Linear Separability
  - Which set is linearly separable?
  - Separating Hyperplane Theorem
  - What if theorem fails?

# Linear Discriminant Function

- In high-dimension,

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0.$$

is a hyperplane.

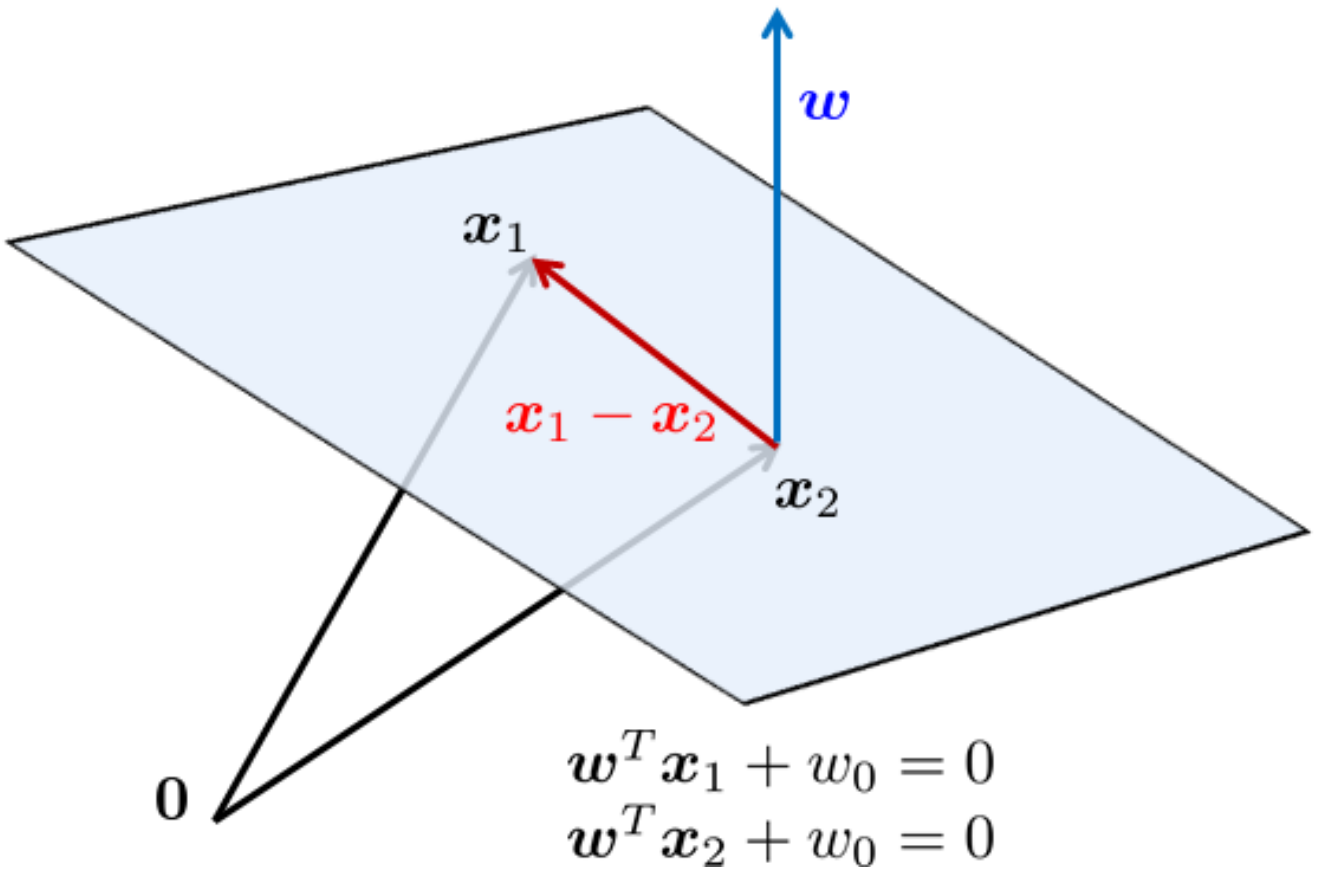


- **Separating Hyperplane**:

$$\begin{aligned}\mathcal{H} &= \{\boldsymbol{x} \mid g(\boldsymbol{x}) = 0\} \\ &= \{\boldsymbol{x} \mid \boldsymbol{w}^T \boldsymbol{x} + w_0 = 0\}\end{aligned}$$

- $\boldsymbol{x} \in \mathcal{H}$ means $\boldsymbol{x}$ is on the decision boundary.
- $\boldsymbol{w}/\|\boldsymbol{w}\|_2$ is the **normal vector** of $\mathcal{H}$.

# Why is $\boldsymbol{w}$ the Normal Vector?




$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0$$
$$\boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0$$

# Why is $\boldsymbol{w}$ the Normal Vector?

- Pick $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from $\mathcal{H}$.
- So $g(\boldsymbol{x}_1) = 0$ and $g(\boldsymbol{x}_2) = 0$.
- This means:

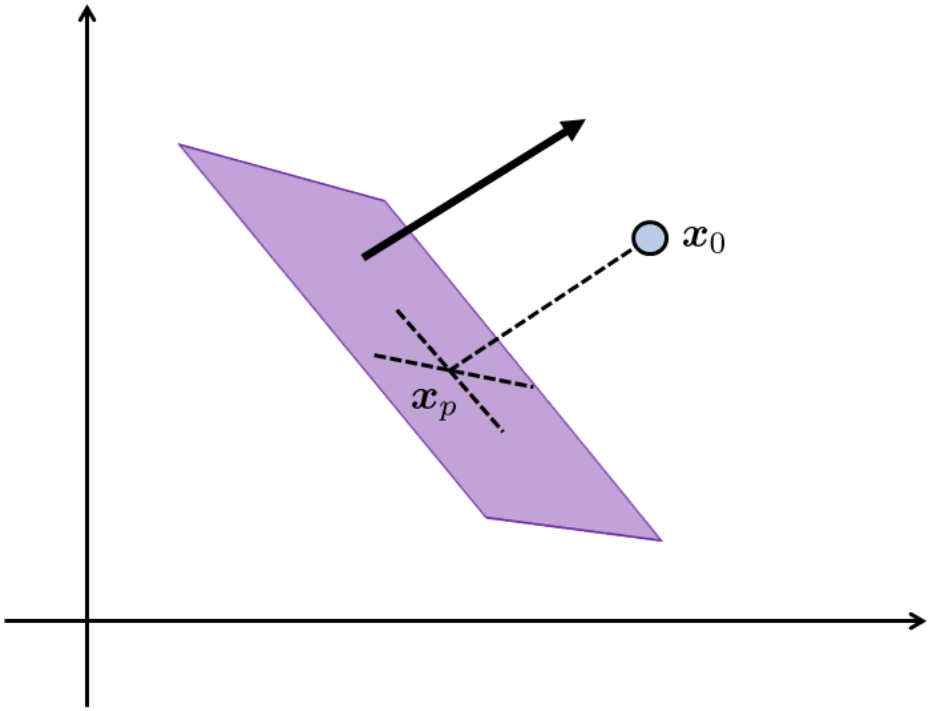
$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0, \quad \text{and} \quad \boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0.$$

- Consider the difference vector $\boldsymbol{x}_1 - \boldsymbol{x}_2$.
- $\boldsymbol{x}_1 - \boldsymbol{x}_2$ is the tangent vector on the surface of $\mathcal{H}$.
- Check

$$\boldsymbol{w}^T (\boldsymbol{x}_1 - \boldsymbol{x}_2) = (\boldsymbol{w}^T \boldsymbol{x}_1 + w_0) - (\boldsymbol{w}^T \boldsymbol{x}_2 + w_0) = 0.$$

- So $\boldsymbol{w}$ is perpendicular to $\boldsymbol{x}_1 - \boldsymbol{x}_2$, hence it is the normal.
- Normalize $\boldsymbol{w}/\|\boldsymbol{w}\|_2$ so that it has unit norm.

# Distance from $x_0$ to $g(x) = 0$




- Pick a point $\boldsymbol{x}_p$ on $\mathcal{H}$
- $\boldsymbol{x}_p$ is the closest point to $\boldsymbol{x}_0$
- $\boldsymbol{x}_0 - \boldsymbol{x}_p$ is the normal direction
- So, for some scalar $\eta > 0$,

$$\boldsymbol{x}_0 - \boldsymbol{x}_p = \eta \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$$

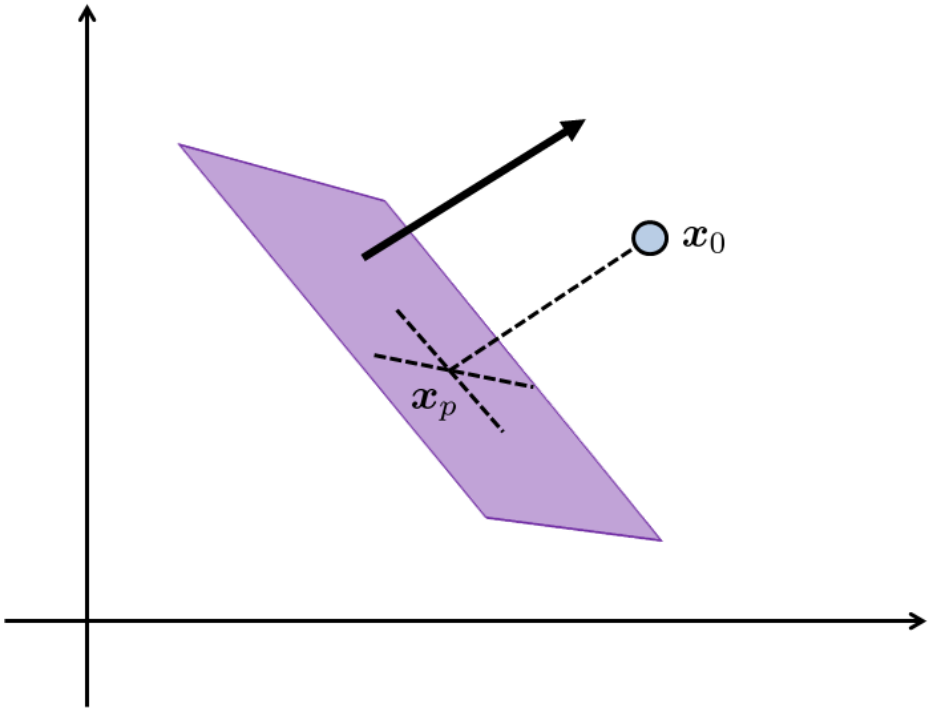
- $\boldsymbol{x}_p$ is on $\mathcal{H}$. So

$$g(\boldsymbol{x}_p) = \boldsymbol{w}^T \boldsymbol{x}_p + w_0 = 0$$

Therefore, we can show that

$$\begin{aligned} g(\boldsymbol{x}_0) &= \boldsymbol{w}^T \boldsymbol{x}_0 + w_0 \\ &= \boldsymbol{w}^T \left( \boldsymbol{x}_p + \eta \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \right) + w_0 \\ &= g(\boldsymbol{x}_p) + \eta \|\boldsymbol{w}\|_2 = \eta \|\boldsymbol{w}\|_2. \end{aligned}$$

# Distance from $x_0$ to $g(x) = 0$




- So distance is

$$\eta = \frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2}$$

- The closest point $\boldsymbol{x}_p$ is

$$\begin{aligned}\boldsymbol{x}_p &= \boldsymbol{x}_0 - \eta \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \\ &= \boldsymbol{x}_0 - \frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2} \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}.\end{aligned}$$

**Conclusion**:

$$\boldsymbol{x}_p = \boldsymbol{x}_0 - \underbrace{\frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2}}_{\text{distance}} \cdot \underbrace{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}}_{\text{normal vector}}$$

# Distance from $x_0$ to $g(x) = 0$

**Alternative Solution**:

We can also obtain the same result by solving the optimization:

$$\boldsymbol{x}_p = \underset{\boldsymbol{x}}{\text{argmin}} \quad \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \quad \text{subject to} \quad \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0.$$

- Let Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \lambda) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2 - \lambda(\boldsymbol{w}^T\boldsymbol{x} + w_0)$$

- Stationarity condition implies

$$\begin{aligned} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \lambda) &= (\boldsymbol{x} - \boldsymbol{x}_0) - \lambda\boldsymbol{w} &= \boldsymbol{0}, \\ \nabla_{\lambda}\mathcal{L}(\boldsymbol{x}, \lambda) &= \boldsymbol{w}^T\boldsymbol{x} + w_0 &= 0. \end{aligned}$$

# Distance from $x_0$ to $g(x) = 0$

Let us do some derivation:

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda) &= (\boldsymbol{x} - \boldsymbol{x}_0) - \lambda \boldsymbol{w} &= \boldsymbol{0}, \\
\nabla_{\lambda} \mathcal{L}(\boldsymbol{x}, \lambda) &= \boldsymbol{w}^T \boldsymbol{x} + w_0 &= 0.
\end{aligned}$$

- This gives

$$\begin{aligned}
& \boldsymbol{x} &= \boldsymbol{x}_0 + \lambda \boldsymbol{w} \\
\Rightarrow \quad & \boldsymbol{w}^T \boldsymbol{x} + w_0 &= \boldsymbol{w}^T (\boldsymbol{x}_0 + \lambda \boldsymbol{w}) + w_0 \\
\Rightarrow \quad & 0 &= \boldsymbol{w}^T \boldsymbol{x}_0 + \lambda \|\boldsymbol{w}\|^2 + w_0 \\
\Rightarrow \quad & 0 &= g(\boldsymbol{x}_0) + \lambda \|\boldsymbol{w}\|^2 \\
\Rightarrow \quad & \lambda &= -\frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|^2} \\
\Rightarrow \quad & \boldsymbol{x} &= \boldsymbol{x}_0 + \left( -\frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|^2} \right) \boldsymbol{w}.
\end{aligned}$$

- Therefore, we arrive at the same result:

$$\boldsymbol{x}_p = \boldsymbol{x}_0 - \underbrace{\frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2}}_{\text{distance}} \cdot \underbrace{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}}_{\text{normal vector}}$$