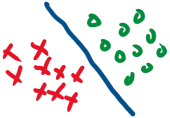
ECE595 / STAT598: Machine Learning I
Lecture 06 Linear Separability

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Overview



Linearly Separable?

Lecture 6

yes

Linear Discriminant Analysis

- Bayesian decision
- logistic regression
- support vector machine
- deep neural network

no

Feature Analysis

- PCA
- hand-crafted features
- deep features

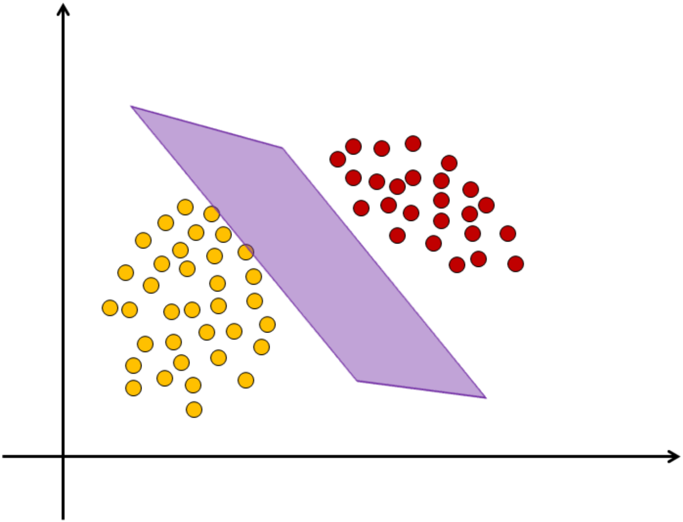
Outline

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?

Supervised Classification

The goal of supervised classification is to construct a **decision boundary** such that the two classes can be (maximally) **separated**.



Terminology

- **Input vectors**: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$.
  - E.g., images, speech, EEG signal, rating, etc
- **Input space**: $\mathcal{X}$. Every $\boldsymbol{x}_n \in \mathcal{X}$.
- **Labels** $y_1, y_2, \ldots, y_N$.
- **Label space**: $\mathcal{Y}$. Every $y_n \in \mathcal{Y}$.
  - If labels are binary, e.g., $y_n = \pm 1$, then
  $$\mathcal{Y} = \{+1, -1\}.$$
  - Labels are arbitrary. $\{+1, -1\}$ and $\{0, 1\}$ has no difference.
- **Target function** $f : \mathcal{X} \to \mathcal{Y}$. Unknown.
  - Relationship:
  $$y_n = f(\boldsymbol{x}_n).$$
- **Hypothesis** $h : \mathcal{X} \to \mathcal{Y}$. Ideally, want
$$h(\boldsymbol{x}) \approx f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

Binary Case

If we restrict ourselves to binary classifier, then

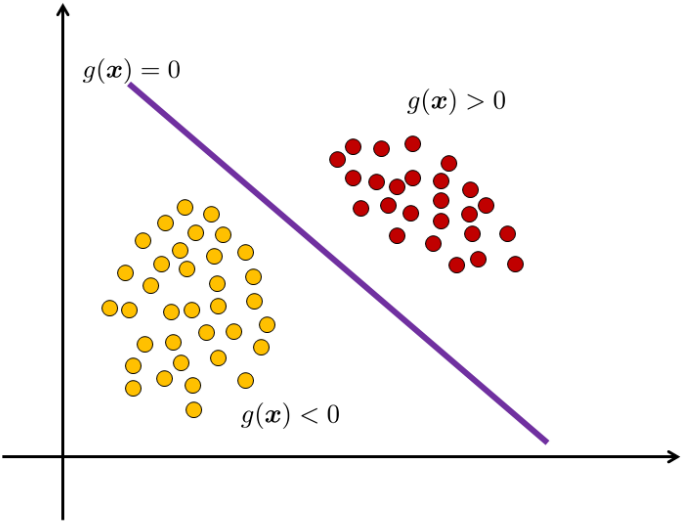
$$h(\boldsymbol{x}) = \begin{cases} 1, & \text{if} \quad g(\boldsymbol{x}) > 0 \\ 0, & \text{if} \quad g(\boldsymbol{x}) < 0 \\ \text{either}, & \text{if} \quad g(\boldsymbol{x}) = 0 \end{cases}$$

- $g : \mathcal{X} \to \mathbb{R}$ is called a **discriminant function**.
- $g(\boldsymbol{x}) > 0$: $\boldsymbol{x}$ lives on the positive side of g.
- $g(\boldsymbol{x}) < 0$: $\boldsymbol{x}$ lives on the negative side of g.
- $g(\boldsymbol{x}) = 0$: The decision boundary.
- You can also claim

$$h(\boldsymbol{x}) = \begin{cases} +1, & \text{if} \quad g(\boldsymbol{x}) > 0 \\ -1, & \text{if} \quad g(\boldsymbol{x}) < 0 \\ \text{either}, & \text{if} \quad g(\boldsymbol{x}) = 0 \end{cases}$$

No difference as far as decision is concerned.

Binary Case



Linear Discriminant Function

A linear discriminant function takes the form

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

- $\boldsymbol{w} \in \mathbb{R}^d$: linear coefficients
- $w_0 \in \mathbb{R}$: bias / offset
- Define the overall **parameter**

$$\boldsymbol{\theta} = \{\boldsymbol{w}, w_0\} \in \mathbb{R}^{d+1}.$$

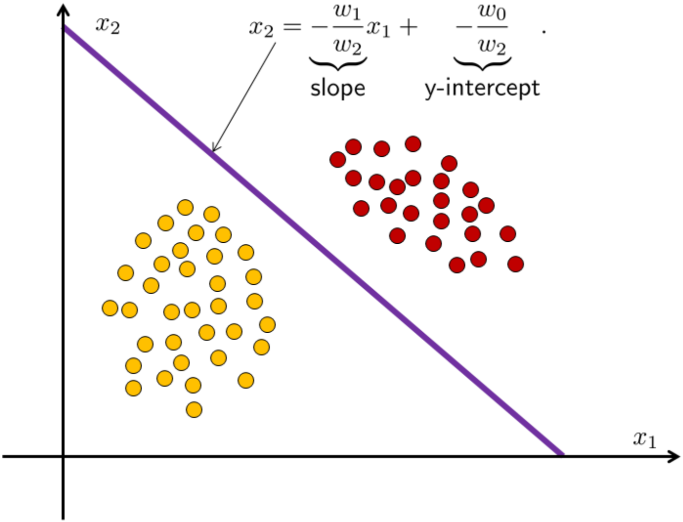
- Example:
  - If $d = 2$, then

$$g(\boldsymbol{x}) = w_2 x_2 + w_1 x_1 + w_0.$$

  - $g(\boldsymbol{x}) = 0$ means

$$x_2 = \underbrace{-\frac{w_1}{w_2}}_{\text{slope}} x_1 + \underbrace{-\frac{w_0}{w_2}}_{\text{y-intercept}}.$$

Linear Discriminant Function



Outline

Goal: Understand the geometry of linear separability.

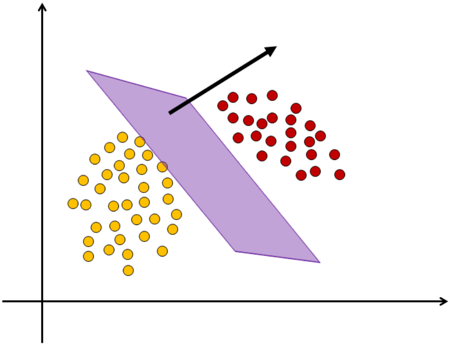
- Notations
  - Input Space, Output Space, Hypothesis
  - Discriminant Function
- Geometry of Discriminant Function
  - Separating Hyperplane
  - Normal Vector
  - Distance from Point to Plane
- Linear Separability
  - Which set is linearly separable?
  - Separating Hyperplane Theorem
  - What if theorem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

Linear Discriminant Function

- In high-dimension,

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

  is a hyperplane.

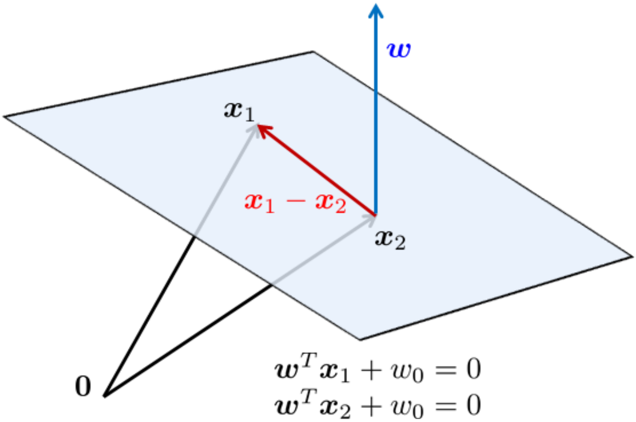


- **Separating Hyperplane**:

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

- $\boldsymbol{x} \in \mathcal{H}$ 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?



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:

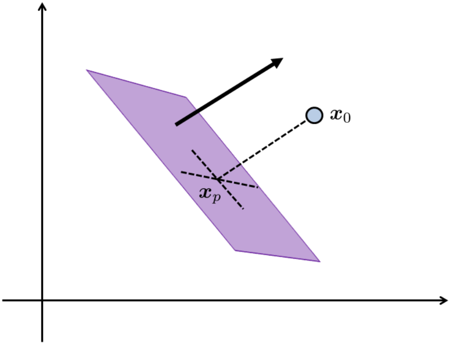
$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0, \quad \text{and} \quad \boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0.$$

- 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 $\boldsymbol{x}_0$ to $g(\boldsymbol{x}) = 0$



- 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 $\eta > 0$,

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

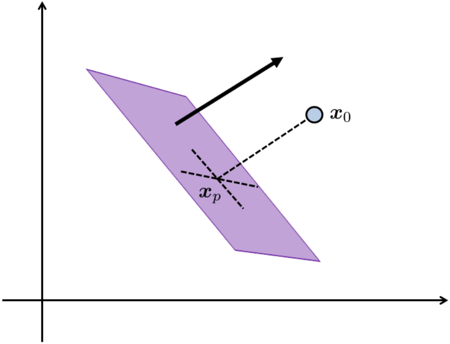
- $\boldsymbol{x}_p$ is on $\mathcal{H}$. So

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

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 $\mathbf{x}_0$ to $g(\mathbf{x}) = 0$



- So distance is

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

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

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

Conclusion:

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

Distance from $\boldsymbol{x}_0$ to $g(\boldsymbol{x}) = 0$

Alternative Solution:

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

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

- 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 $\boldsymbol{x}_0$ to $g(\boldsymbol{x}) = 0$

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:

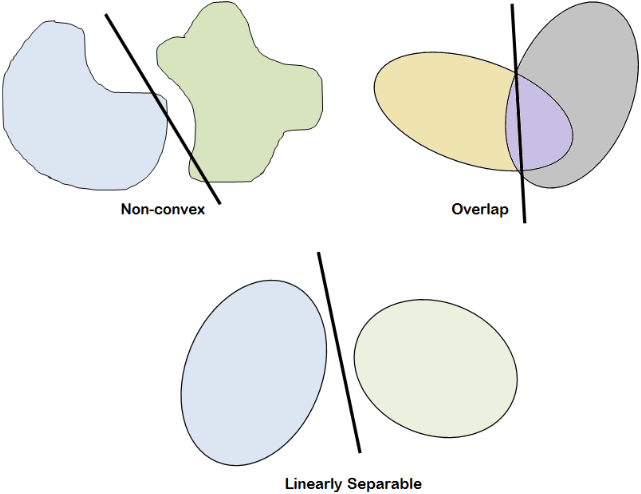
$$
\boldsymbol{x}_p \quad = \quad \boldsymbol{x}_0 \quad - \quad \underbrace{\frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2}}_{\text{distance}} \quad \cdot \quad \underbrace{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}}_{\text{normal vector}}
$$

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?

Which one is Linearly Separable? Which one is Not?



Separating Hyperplane Theorem

Can we always find a separating hyperplane?

- No.
- Unless the classes are linearly separable.
- If convex and not overlapping, then yes.

Theorem (Separating Hyperplane Theorem)

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two closed convex sets such that $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. Then, there exists a linear function

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

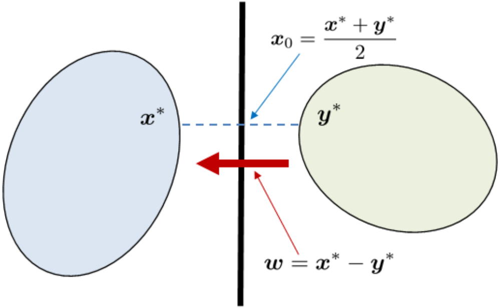
such that $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{C}_1$ and $g(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in \mathcal{C}_2$.

Remark: The theorem above provides sufficiency but not necessity for linearly separability.

Separating Hyperplane Theorem

Pictorial "proof":

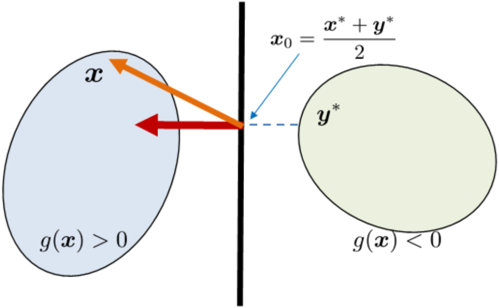
- Pick two points $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ s.t. the distance between the sets is minimized.
- Define the mid-point as $\boldsymbol{x}_0 = (\boldsymbol{x}^* + \boldsymbol{y}^*)/2$.
- Draw the separating hyperplane with normal $\boldsymbol{w} = \boldsymbol{x}^* - \boldsymbol{y}^*$
- Convexity implies any inner product must be positive.



Separating Hyperplane Theorem

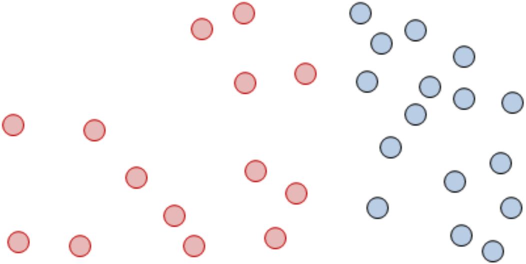
Pictorial "proof":

- Pick two points $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ s.t. the distance between the sets is minimized.
- Define the mid-point as $\boldsymbol{x}_0 = (\boldsymbol{x}^* + \boldsymbol{y}^*)/2$.
- Draw the separating hyperplane with normal $\boldsymbol{w} = \boldsymbol{x}^* - \boldsymbol{y}^*$
- Convexity implies any inner product must be positive.



Linearly Separable?

- I have data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$.
- Closed. Convex. Non-overlapping.
- Separating hyperplane theorem: I can find a line.
- Victory?
- Not quite.



When Theory Fails

Theorem (Separating Hyperplane Theorem)

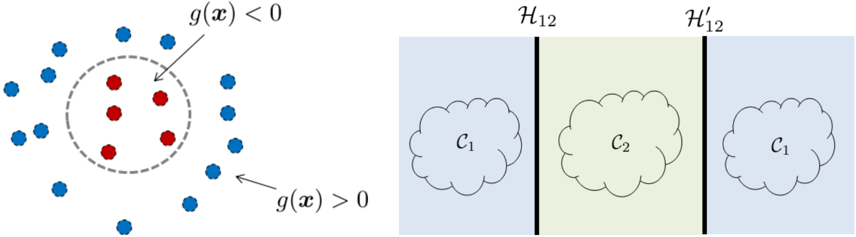
Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two closed convex sets such that $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. Then, there exists a linear function

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

such that $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{C}_1$ and $g(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in \mathcal{C}_2$.

- Finding a separating hyperplane for **training set** does not imply it will work for the **testing set**.
- Separating hyperplane theorem is more often used in **theoretical analysis** by assuming properties of the testing set.
- If a dataset is linearly separable, then you are guaranteed to find a perfect classifier. Then you can say how good is the classifier you designed compared to the perfect one.

Linear Classifiers Do Not Work



Example 1 Example 2

- **Intrinsic geometry** of the two classes could be bad.
- The training set could be **lack of training samples**.
- Solution 1: Use non-linear classifiers, e.g.,
  $g(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W} \boldsymbol{x} + \boldsymbol{w}^T \boldsymbol{x} + \omega_0$.
- Solution 2: Kernel method, e.g., Radial basis function.
- Solution 3: Extract features, e.g., $g(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$.

Reading List

Separating Hyperplane:

- Duda, Hart and Stork's *Pattern Classification*, Chapter 5.1 and 5.2.
- Princeton ORFE-523, Lecture 5 on Separating hyperplane http://www.princeton.edu/~amirali/Public/Teaching/ORF523/S16/ORF523_S16_Lec5_gh.pdf
- Cornell ORIE-6300, Lecture 6 on Separating hyperplane https://people.orie.cornell.edu/dpw/orie6300/fall2008/Lectures/lec06.pdf
- Caltech, Lecture Note http://www.its.caltech.edu/~kcborder/Notes/SeparatingHyperplane.pdf

Appendix

Proof of Separating Hyperplane Theorem

- Conjecture: Let's see if this is the correct hyperplane

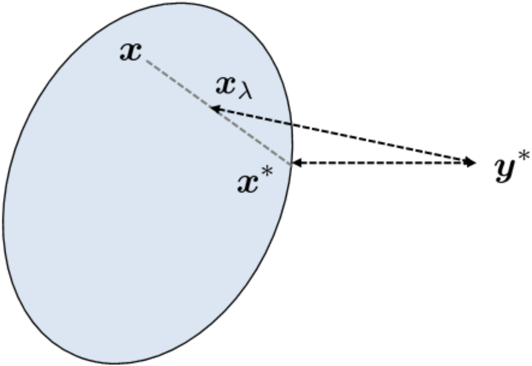
$$
\begin{aligned}
g(\boldsymbol{x}) &= \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) \\
&= (\boldsymbol{x}^* - \boldsymbol{y}^*)^T \left(\boldsymbol{x} - \frac{\boldsymbol{x}^* + \boldsymbol{y}^*}{2} \right) \\
&= (\boldsymbol{x}^* - \boldsymbol{y}^*)^T \boldsymbol{x} - \frac{\|\boldsymbol{x}^*\|^2 - \|\boldsymbol{y}^*\|^2}{2}
\end{aligned}
$$

- According to picture, we want $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{C}_1$.
- Suppose not. Assume

$$
g(\boldsymbol{x}) = (\boldsymbol{x}^* - \boldsymbol{y}^*)^T \boldsymbol{x} - \frac{\|\boldsymbol{x}^*\|^2 - \|\boldsymbol{y}^*\|^2}{2} < 0.
$$

See if we can find a contradiction.

Proof of Separating Hyperplane Theorem



- $\mathcal{C}_1$ is convex.
- Pick $\boldsymbol{x} \in \mathcal{C}_1$
- Pick $\boldsymbol{x}^* \in \mathcal{C}_1$
- Let $0 \leq \lambda \leq 1$
- Construct a point

$$\boldsymbol{x}_\lambda = (1-\lambda)\boldsymbol{x}^* + \lambda\boldsymbol{x}.$$

- Convex means

$$\boldsymbol{x}_\lambda \in \mathcal{C}_1$$

So we must have

$$\|\boldsymbol{x}_\lambda - \boldsymbol{y}^*\| \geq \|\boldsymbol{x}^* - \boldsymbol{y}^*\|$$

Proof of Separating Hyperplane Theorem

- Pick an arbitrary point $\boldsymbol{x} \in \mathcal{C}_1$.
- $\boldsymbol{x}^*$ is fixed already.
- Pick $\boldsymbol{x}_\lambda$ along the line connecting $\boldsymbol{x}$ and $\boldsymbol{x}^*$.
- Convexity implies $\boldsymbol{x}_\lambda \in \mathcal{C}_1$.
- So $\|\boldsymbol{x}_\lambda - \boldsymbol{y}^*\| \geq \|\boldsymbol{x}^* - \boldsymbol{y}^*\|$. If not, something is wrong.
- Let us do some algebra:

$$
\begin{aligned}
\|\boldsymbol{x}_\lambda - \boldsymbol{y}^*\|^2 &= \|(1-\lambda)\boldsymbol{x}^* + \lambda\boldsymbol{x} - \boldsymbol{y}^*\|^2 \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^* + \lambda(\boldsymbol{x} - \boldsymbol{x}^*)\|^2 \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 + 2\lambda(\boldsymbol{x}^* - \boldsymbol{y}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) + \lambda^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 + 2\lambda\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}^*) + \lambda^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2.
\end{aligned}
$$

- Remember: $\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) < 0$.

Proof of Separating Hyperplane Theorem

$$
\begin{aligned}
\|\boldsymbol{x}_\lambda - \boldsymbol{y}^*\|^2 &= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 + 2\lambda \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}^*) + \lambda^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \\
&< \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 + 2\lambda(\boldsymbol{w}^T\boldsymbol{x}_0 - \boldsymbol{w}^T\boldsymbol{x}^*) + \lambda^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 + 2\lambda\left[\left(\frac{\|\boldsymbol{x}^*\|^2 - \|\boldsymbol{y}^*\|^2}{2}\right) - \boldsymbol{w}^T\boldsymbol{x}^*\right] \\
&\quad + \lambda^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 - \lambda\underbrace{\|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2}_{=A} + \lambda^2\underbrace{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2}_{=B} \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 - \lambda A + \lambda^2 B \\
&= \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 - \lambda(A - \lambda B).
\end{aligned}
$$

Now, pick an $\boldsymbol{x}$ such that $A - \lambda B > 0$. Then $-\lambda(A - \lambda B) < 0$.

$$
\lambda < \frac{A}{B} = \frac{\|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2}{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2}.
$$

Proof of Separating Hyperplane Theorem

Therefore, if we choose λ such that $A - \lambda B > 0$, i.e.,

$$\lambda < \frac{A}{B} = \frac{\|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2}{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2},$$

then $-\lambda(A - \lambda B) < 0$, and so

$$\begin{aligned}\|\boldsymbol{x}_\lambda - \boldsymbol{y}^*\|^2 &< \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2 - \lambda(A - \lambda B) \\ &< \|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2\end{aligned}$$

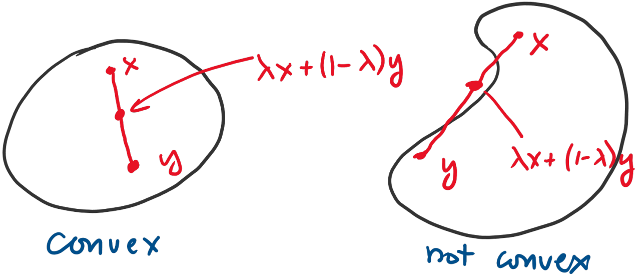
Contradiction, because $\|\boldsymbol{x}^* - \boldsymbol{y}^*\|^2$ should be the smallest!

Conclusion:

- If $\boldsymbol{x} \in \mathcal{C}_1$, then $g(\boldsymbol{x}) > 0$.
- By symmetry, if $\boldsymbol{x} \in \mathcal{C}_2$, then $g(\boldsymbol{x}) < 0$.
- And we have found the separating hyperplane $(\boldsymbol{w}, w_0)$.

Q&A 1: What is a convex set?

- A set C is convex if the following condition is met.
- Pick $\boldsymbol{x} \in C$ and $\boldsymbol{y} \in C$, and let $0 < \lambda < 1$. If $\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ is also in C for any $\boldsymbol{x}$, $\boldsymbol{y}$ and λ, then C is convex.
- Basically, it says that you can pick two points and draw a line. If the line is also in the set, then the set is convex.



Q&A 2: Is there a way to check whether two sets are linearly separable?

- No, at least I do not know.
- The best you can do is to check whether a **training set** is linearly separable.
- To do so, solve the **hard SVM**. If you can solve it with zero training error, then you have found one. If the hard SVM does not have a solution, then the training set is not separable.
- Checking the **testing set** is impossible unless you know the distributions of the samples. But if you know the distributions, you can derive formula to check linear separability.
- For example, Gaussians are not linearly separable because no matter how unlikely you can always find a sample that lives in the wrong side. Uniform distributions are linearly separable.
- Bottom line: Linear separability, in my opinion, is more of a theoretical tool to describe the **intrinsic property** of the problem. It is not for computational purposes.

Q&A 3: If two sets are not convex, how do I know if it is linearly separable?

- You can look at the convex hull.
- A convex hull is the smallest convex set that contains the original set.
- If the convex hulls are not overlapping, then linearly separable.
- For additional information about convex sets, convex hulls, you can check Chapter 2 of https://web.stanford.edu/class/ee364a/lectures.html