ECE595 / STAT598: Machine Learning I
Lecture 19.1: Support Vector Machine - Concept of Margin

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

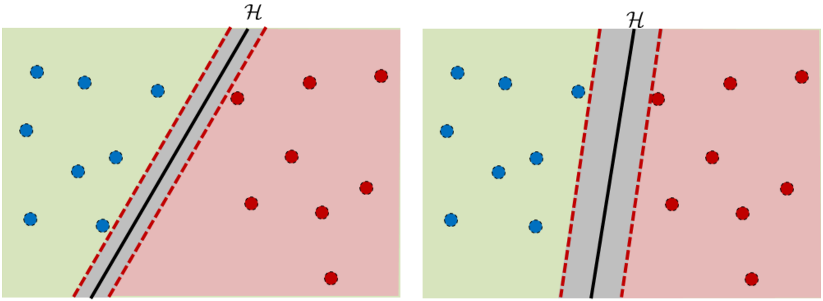
Support Vector Machine

- Lecture 19 SVM 1: The Concept of Max-Margin
- Lecture 20 SVM 2: Dual SVM
- Lecture 21 SVM 3: Kernel SVM

This lecture: Support Vector Machine 1

- Concept of Margin
  - Distance from point to plane
  - Margin
  - Max Margin Classifier
- SVM
  - SVM via Optimization
  - Programming SVM
  - Visualization

Margin and Max-Margin Classifier



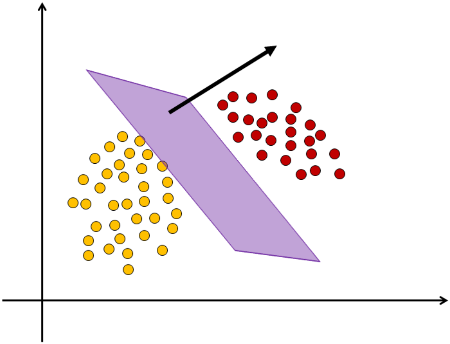
- **Margin**: Smallest gap between the two classes
- **Max-Margin Classifier**: A classifier that maximizes the margin
- **What do we need?**
  - How to measure the distance from a point to a plane?
  - How to formulate a max margin problem?
  - How to solve the max margin problem?

Recall: Linear Discriminant Function

- In high-dimension,

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

is a hyperplane.

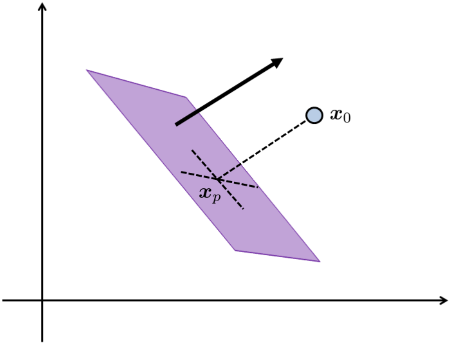


- **Separating Hyperplane**:

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

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

Recall: Distance from x_0 to $g(x) = 0$



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

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

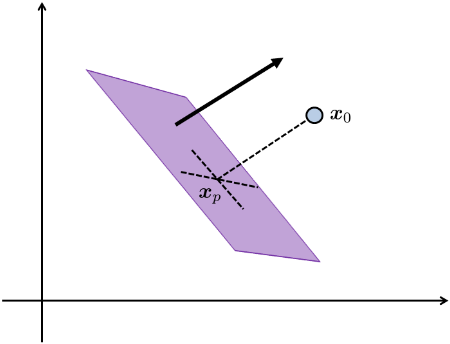
- $\mathbf{x}_p$ is on $\mathcal{H}$. So

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

Therefore, we can show that

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

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



- 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:

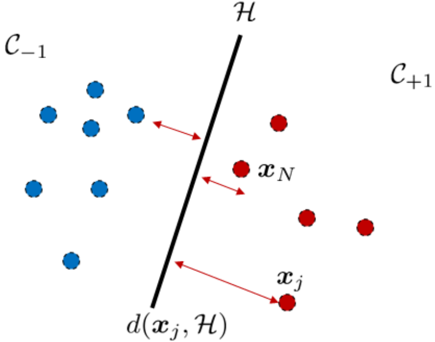
$$\boldsymbol{x}_p = \boldsymbol{x}_0 - \underbrace{\frac{g(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2}}_{\text{distance}} \cdot \underbrace{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}}_{\text{normal vector}}$$

Unsigned Distance

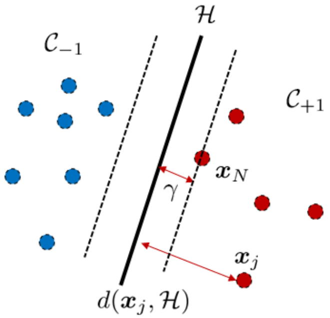
- We define the distance between a data point $\boldsymbol{x}_j$ and a separating hyperplane as

$$d(\boldsymbol{x}_j, \mathcal{H}) = \frac{|g(\boldsymbol{x}_j)|}{\|\boldsymbol{w}\|_2} = \frac{|\boldsymbol{w}^T \boldsymbol{x}_j + w_0|}{\|\boldsymbol{w}\|_2}.$$

- $d(\boldsymbol{x}_j, \mathcal{H})$ is called **unsigned** distance



Margin

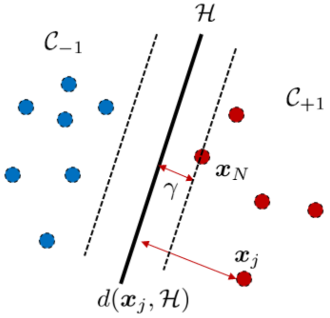


- Among all the unsigned distances, pick the smallest one.
- Margin: γ such that

$$\gamma = \min_{j=1,\ldots,N} \; d(\boldsymbol{x}_j, \mathcal{H}).$$

- Without loss of generality, assume $\boldsymbol{x}_N$ is the closest point.

More about Margin



- Margin: γ such that

$$\gamma = \min_{j=1,\dots,N} \; d(\boldsymbol{x}_j, \mathcal{H}).$$

- γ depends on $(\boldsymbol{w}, w_0)$.
- γ always exist because training set is finite.
- $\gamma \geq 0$, and is zero when $\boldsymbol{x}_N$ is on the boundary.

Signed Distance

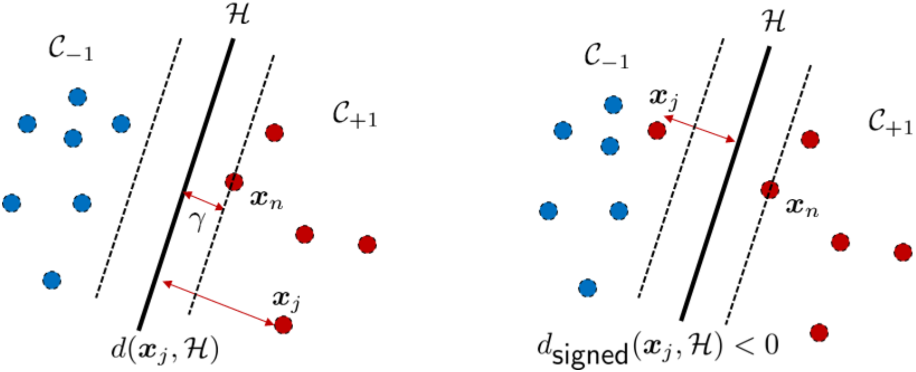
- $d(\boldsymbol{x}_j, \mathcal{H})$ is unsigned
- So γ does not tell whether a point $\boldsymbol{x}_j$ is correctly classified or not
- Assume that the labels are defined as $y_j \in \{-1, +1\}$
- Then define a **signed distance**

$$d_{\text{signed}}(\boldsymbol{x}_j, \mathcal{H}) = y_j \left(\frac{\boldsymbol{w}^T \boldsymbol{x}_j + w_0}{\|\boldsymbol{w}\|_2} \right) = \begin{cases} \geq 0, & \text{correctly classify } \boldsymbol{x}_j \\ < 0, & \text{incorrectly classify } \boldsymbol{x}_j. \end{cases}$$

- Recall perceptron loss:

$$\mathcal{L}(\boldsymbol{x}_j) = \max\left\{ -y_j(\boldsymbol{w}^T \boldsymbol{x}_j + w_0), 0 \right\}$$

Unsigned VS Signed Distance



- Unsigned distance: Just the distance
- Signed distance: Distance plus whether on the correct side

Max-Margin Objective

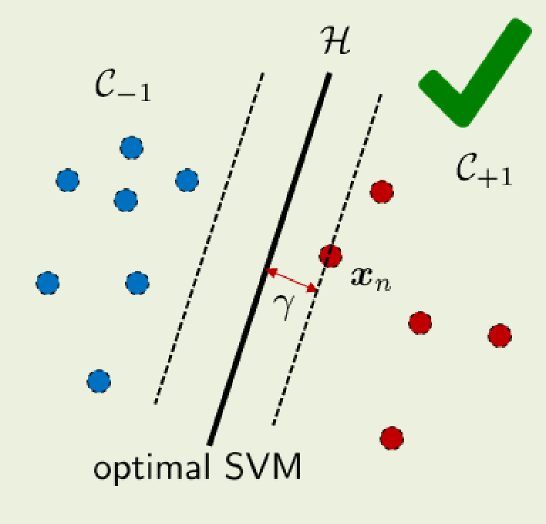
- **Assumptions**: Linearly separable.
- This means

$$y_j \left(\frac{\boldsymbol{w}^T \boldsymbol{x}_j + w_0}{\|\boldsymbol{w}\|_2} \right) \geq \gamma, \quad j = 1, \ldots, N.$$

- All training samples are correctly classified.
- All training samples are at lest γ from the boundary.
- So the max-margin classifier is

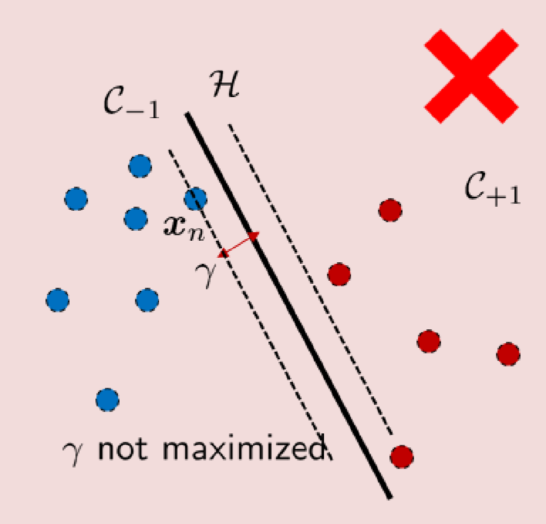
$$\begin{aligned} \underset{\boldsymbol{w}, w_0}{\text{maximize}} \quad & \gamma \\ \text{subject to} \quad & y_j \left(\frac{\boldsymbol{w}^T \boldsymbol{x}_j + w_0}{\|\boldsymbol{w}\|_2} \right) \geq \gamma, \quad j = 1, \ldots, N. \end{aligned}$$

Good or Bad?



$\mathcal{H}$
$\mathcal{C}_{-1}$
$\mathcal{C}_{+1}$
$\boldsymbol{x}_n$
γ
optimal SVM

Good or Bad?



$\mathcal{H}$
$\mathcal{C}_{-1}$
$\mathcal{C}_{+1}$
$\boldsymbol{x}_n$
γ
γ not maximized

Good or Bad?