MAT8034: Machine Learning

Support Vector Machines

Fang Kong

https://fangkongx.github.io/Teaching/MAT8034/Spring2025/index.html

Part of slide credit: Stanford CS229

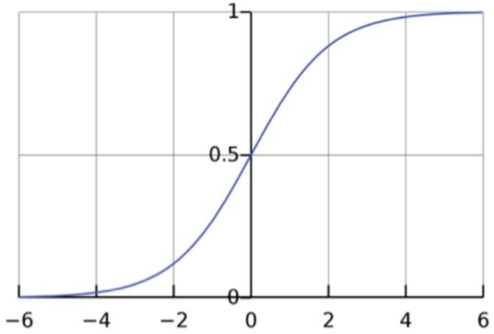
Outline

- Support vector machines
  - Intuition: margins
  - Problem definition
  - Functional and geometric margins
  - The optimal margin classifier
  - Regularization and the non-separable case

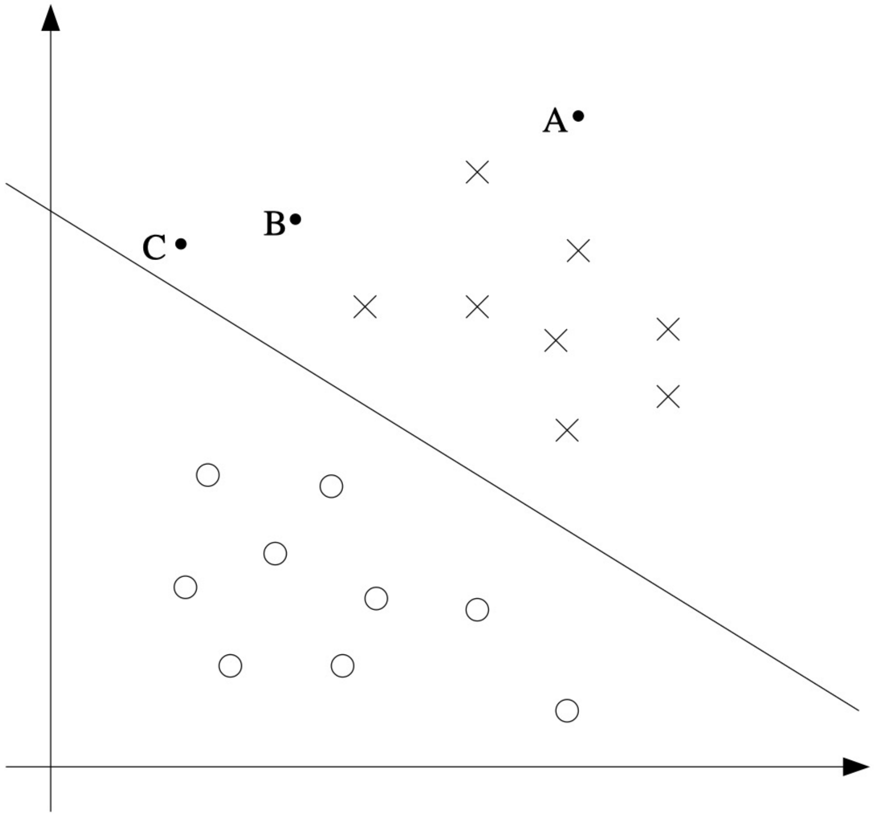
Intuition: margins

The confidence of predictions

- Recall in the logistic regression
  - Predict the probability $p(y = 1|x; \theta)$ using $h_\theta(x) = g(\theta^\top x)$
  - Predict the label $y = 1$ if $h_\theta(x) > 0.5$
  - Predict the label $y = 0$ otherwise
- Consider different examples
  - For x with $\theta^\top x \gg 0$, being confident to predict $y = 1$
  - For x with $\theta^\top x \approx 0.0005$, being NOT confident to predict $y = 1$



Illustration



- Confidence of the prediction: A>B>C

The confidence of predictions

- We have a good model if the θ satisfies
  - When $y = 1$, $\theta^\top x \gg 0$
  - When $y = 0$, $\theta^\top x \ll 0$

- This reflects a very confident (and correct) set of classifications

- Our objective: introduce the functional margins (confidence) to evaluate the performance

New formulation of classification

Formulation

- To better evaluate the sigh of the label
  - Label $y \in \{-1,1\}$
- Linear classifier (based on parameter w, b)

$$h_{w,b}(x) = g(w^T x + b)$$

  - b plays the role of previous θ_0, w plays the role of previous $[\theta_1, \theta_2, \dots, \theta_d]$
- Activation function
  - $g(z) = 1$ if $z \geq 0$
  - $g(z) = 0$ otherwise
- Difference from logistic regression: do not predict the probability

Functional and geometric margins

Functional margin

- Define the functional margin w.r.t. training example i

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

- Intuition: to make the margin larger
  - When $y^i = 1$, hope $w^\top x^i + b$ to be a large positive number
  - When $y^i = -1$, hope $w^\top x^i + b$ to be a large negative number
  - If $\hat{\gamma}^i > 0$: prediction is correct

  - A large functional margin represents a confident and a correct prediction.

Functional margin

- Given the training set $S = \{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$
- Define the functional margin w.r.t. training set

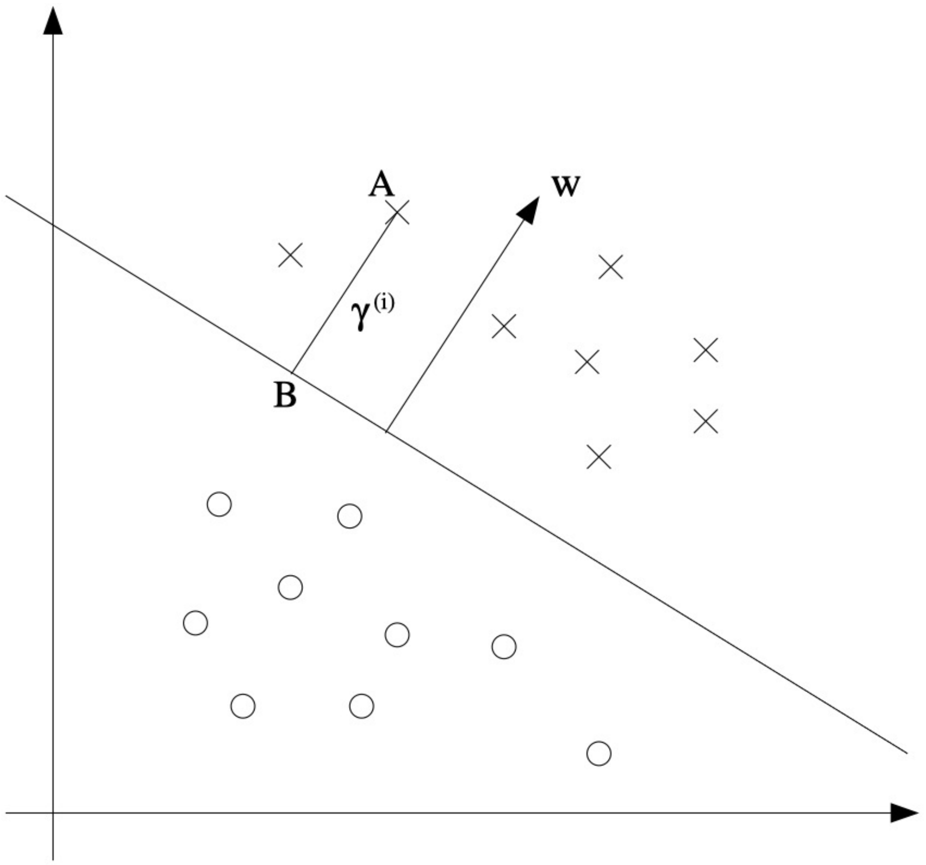
$$\hat{\gamma} = \min_{i=1,\ldots,n} \hat{\gamma}^{(i)}$$

Limitation

- If we replace w, b with $2w, 2b$
  - The prediction $g(w^\top x^i + b)$ does not change (since the sigh does not change
  - But the function margin changes $\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$

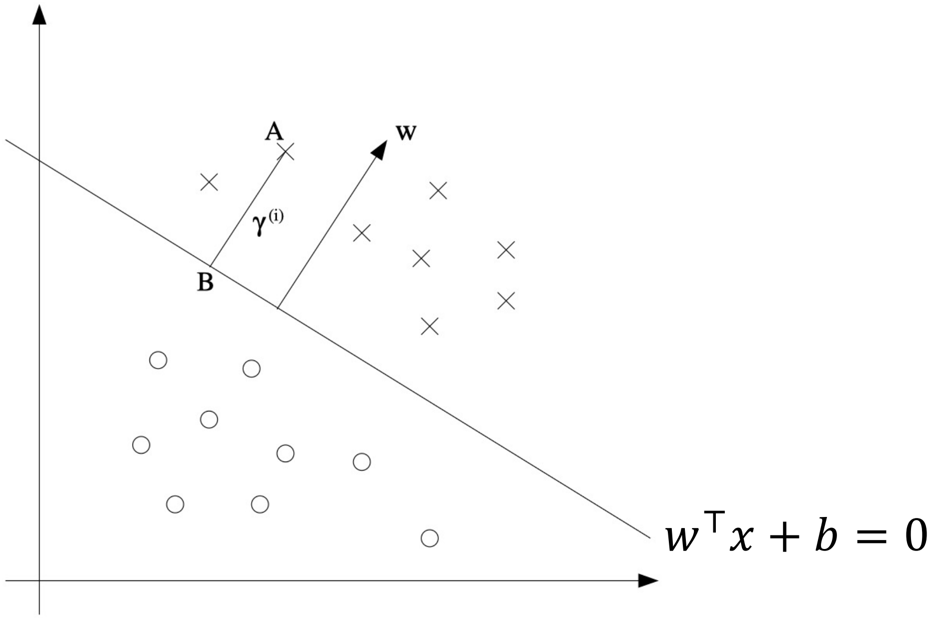
- From this view, optimizing the functional margin changes anything meaningful

Improvement: geometric margins



$w^\top x + b = 0$

Improvement: geometric margins



- How to compute the function margin?

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||}$$

Geometric margins: formal definition

- For any training example (x^i, y^i)

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

- If $||w|| = 1$, the function margin equals to geometric margin

- Finally, given training set $S = \{(x^1, y^1), (x^2, y^2), \dots, (x^n, y^n)\}$

$$\gamma = \min_{i=1,\dots,n} \gamma^{(i)}$$

The optimal margin classifier

The optimization objective

- Given a training set that is linearly separable
- How to achieve the maximum geometric margin

$$
\begin{aligned}
\max_{\gamma,w,b} \quad & \gamma \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, n \\
& ||w|| = 1.
\end{aligned}
$$

- Using optimization algorithm to solve it
- But …
  - $\|w\| = 1$ is a non-convex constraint, no standard optimization algorithm

Transforming the problem

- New form

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

- $||w||$ is a non-convex

Keep going

- Recall that scaling constraint on w and b without changing anything on prediction but influences the margin
- We can scale w and b to ensure $\hat{\gamma} = 1$
- Then maximizing $\frac{\hat{\gamma}}{||w||}$ equivalents to minimizing $\frac{1}{2}||w||^2$
- New problem

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$

Quadradic convex objective

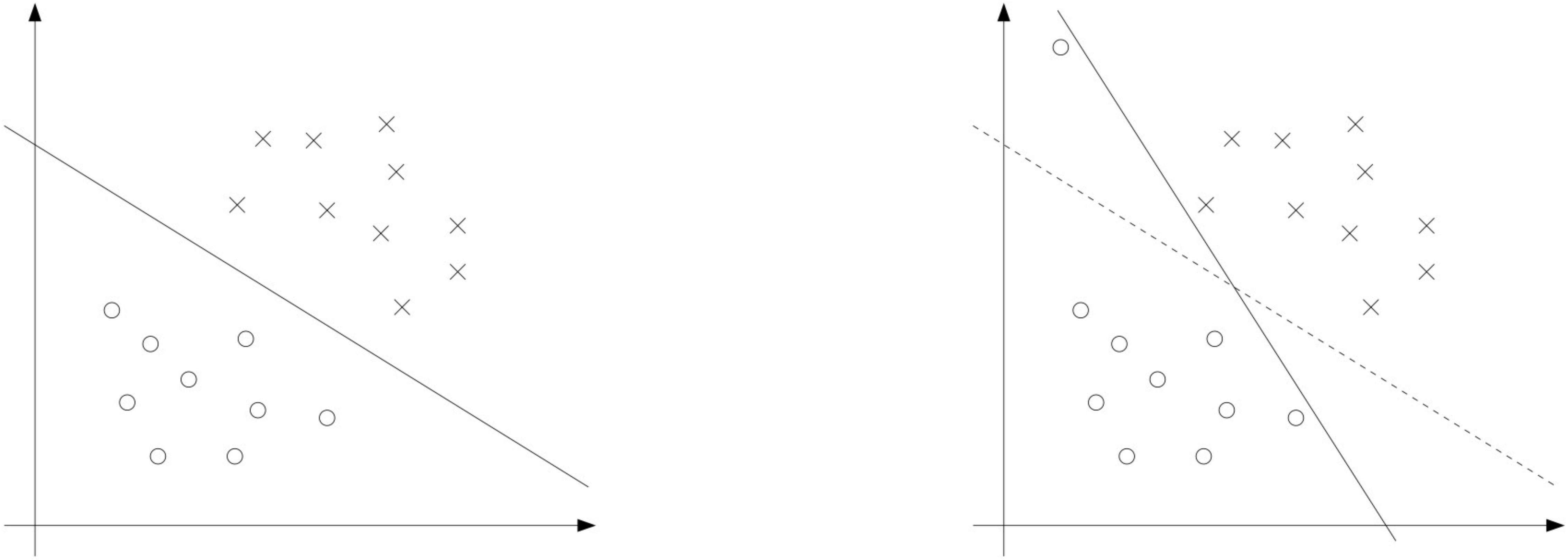
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n$$

Linear constraint

- The dual form and extension using kernel tricks are omitted

Regularization and the non-separable case

What happens if the data is non-separable



Solution

- To make the algorithm work for non-linearly separable datasets as well as be less sensitive to outliers

$$
\begin{aligned}
\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\
& \xi_i \geq 0, \quad i = 1, \ldots, n.
\end{aligned}
$$

Summary

- Support vector machines
  - Intuition: margins
  - Problem definition
  - Functional and geometric margins
  - The optimal margin classifier
  - Regularization and the non-separable case