CSE/ISyE 6740
Computational Data Analysis

Support Vector Machine

09/17/2025

Kai Wang, Assistant Professor in Computational Science and Engineering
kwang692@gatech.edu

Georgia Tech

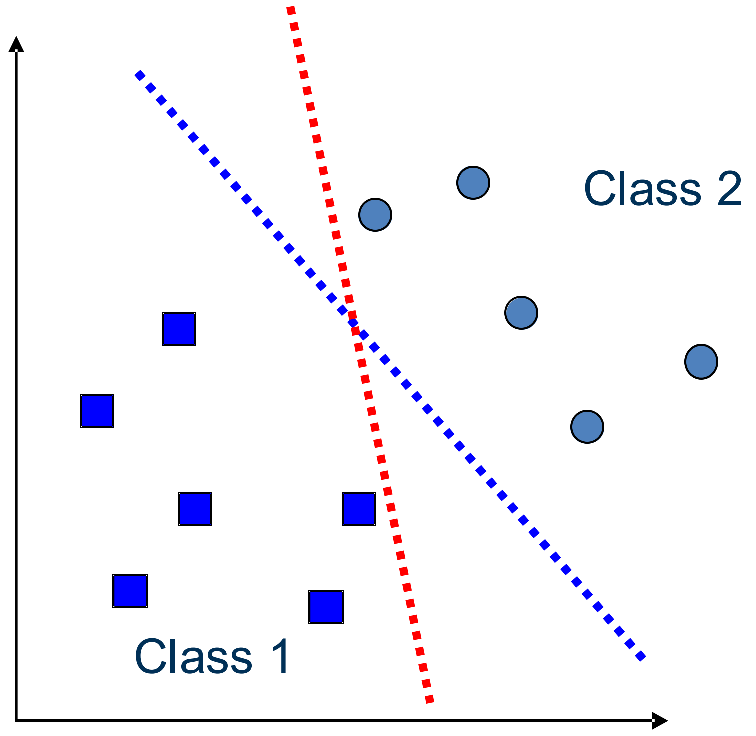
Outline

- **Supervised Learning**
  - Support Vector Machine (SVM)
    - Decision boundary
    - Maximum margin problem
    - Lagrangian duality
    - Dual problem of SVM
    - Inference and support vectors
    - Kernel SVM

Support Vector Machines

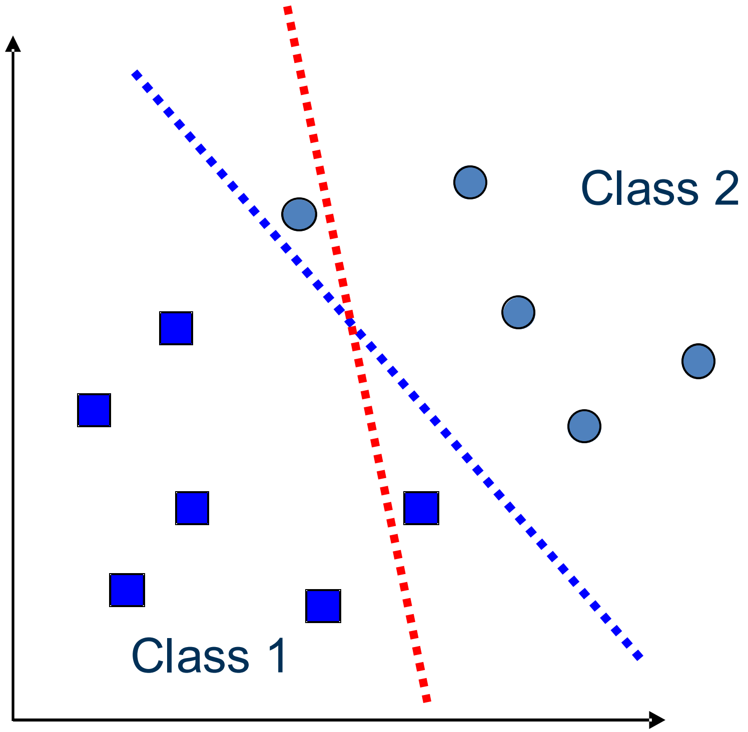
Which Decision Boundary is Better?

- Suppose the training samples are linearly separable
- We can find a decision boundary which gives zero training error
- But there are many such decision boundaries
- Which one is better?



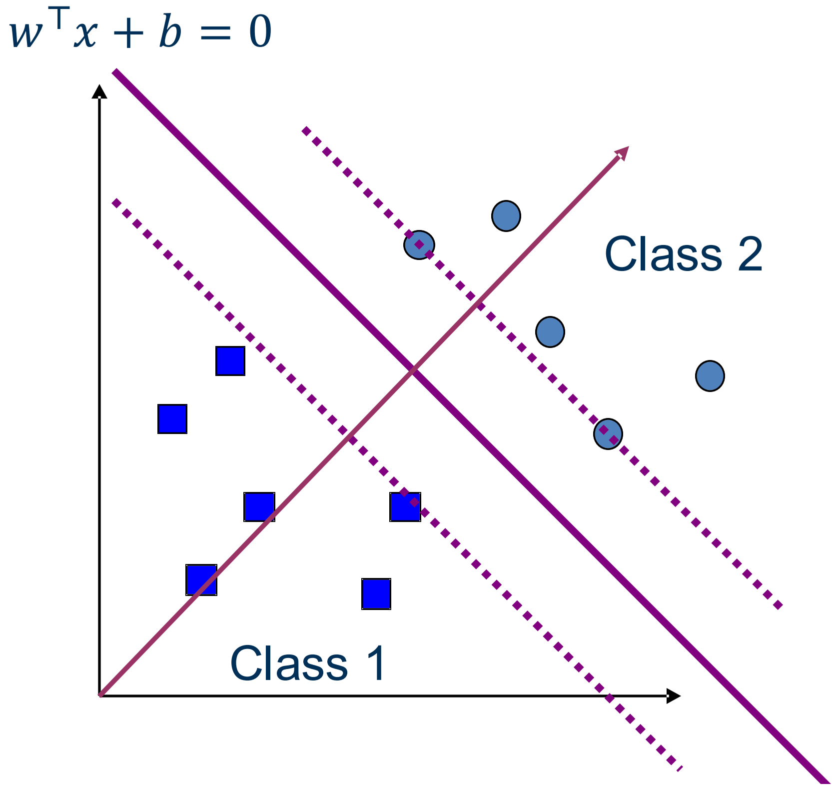
Which Decision Boundary is Better?

- Suppose we perturb the data, which boundary is more robust to error?



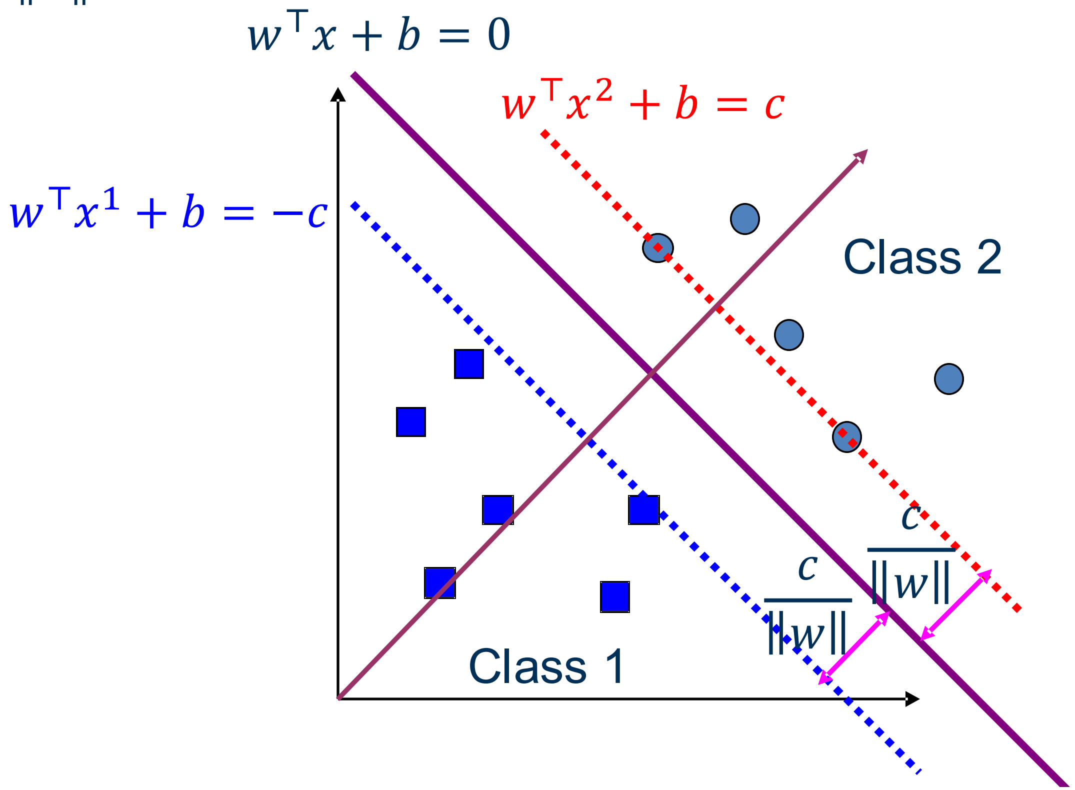
Geometric Interpretation of a Classifier

- Parameterizing decision boundary as: $w^\top x + b = 0$
  - w denotes a vector orthogonal to the decision boundary
  - b is a scalar offset term
- Dash lines are parallel to the decision boundary, and they just hit the data points



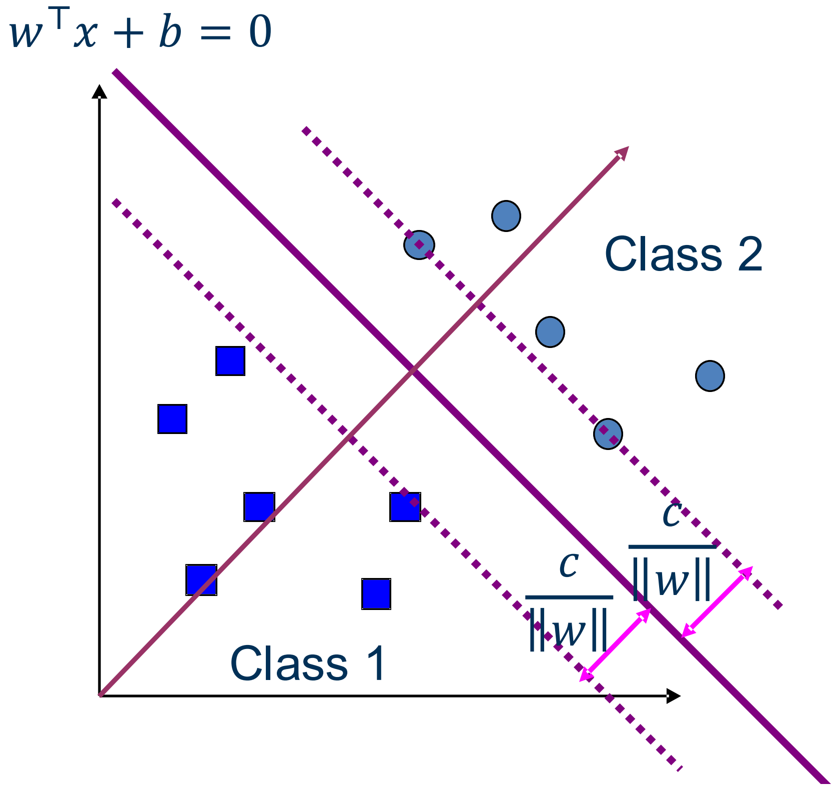
Formal Definition of Margin

- Pick two data points x^1 and x^2 which are on each dash line respectively
  - $w^\top x^2 + b = c$
  - $w^\top x^1 + b = -c$
- The unnormalized margin is $\tilde{\gamma} = w^\top(x^2 - x^1) = 2c$
- The margin is $\gamma = \frac{2c}{\|w\|}$



Formal Definition of Constraints on Data Points

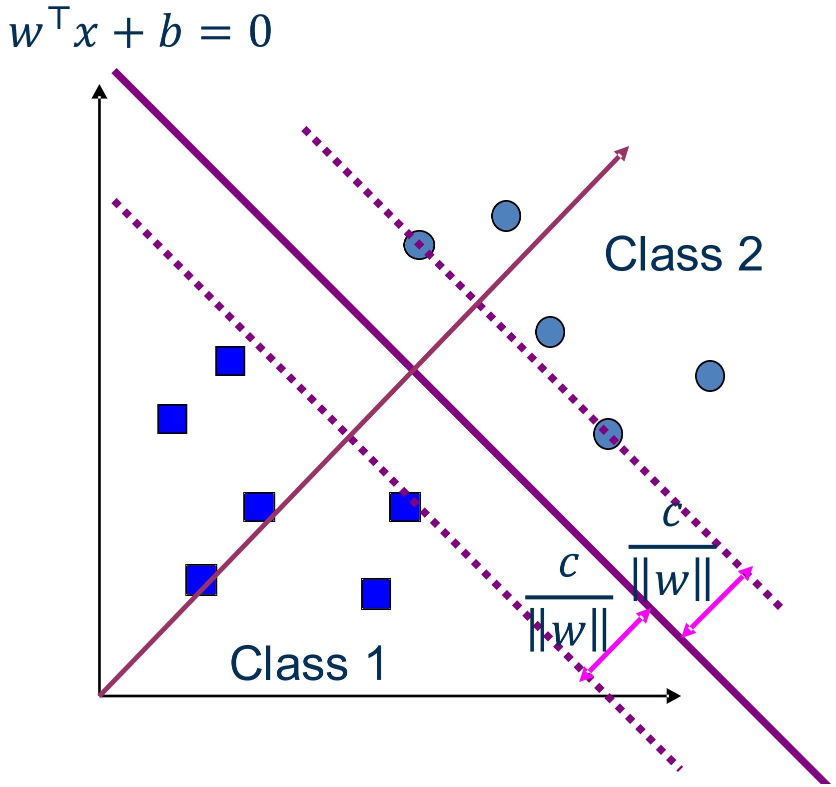
- Constraints on data points
  - For all x in class 2, $y = 1$ and $w^{\top}x + b \geq c$
  - For all x in class 1, $y = -1$ and $w^{\top}x + b \leq -c$
- Or more compactly, $(w^{\top}x + b)y \geq \mathrm{c}$



Maximum Margin Classifier

- Find decision boundary w as far from data point as possible

$$\max_{w,b} \quad \gamma = \frac{2c}{\|w\|}$$
$$s.t.\, y^i(w^\top x^i + b) \geq c, \qquad \forall i$$



Equivalent Form

$$\max_{w,b} \frac{2c}{\|w\|}$$
$$s.t.\, y^i\left(w^\top x^i + b\right) \geq c, \qquad \forall i$$

- Note that the magnitude of c merely scales w and b, and does not change the relative goodness of different classifiers

- Set $c = 1$ and drop the constant to get a cleaner problem

$$\max_{w,b} \frac{1}{\|w\|}$$
$$s.t.\, y^i\left(w^\top x^i + b\right) \geq 1, \qquad \forall i$$

Support Vector Machine Optimization Form

- A constrained convex quadratic programming problem (standard form)

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$
$$s.t.\, 1 - y^i\left(w^\top x^i + b\right) \leq 0, \qquad \forall i$$

- After optimization, the margin is given by $\frac{2}{\|w\|}$

Lagrangian Duality

Optimization and Lagrangian Duality

- The **primal** problem

$$
\begin{aligned}
&\min_{w} f(w) \\
s.t.\ & g_i(w) \le 0, \qquad i = 1,2,\ldots,k \\
& h_i(w) = 0, \qquad i = 1,2,\ldots,l
\end{aligned}
$$

- The Lagrangian function

$$L(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

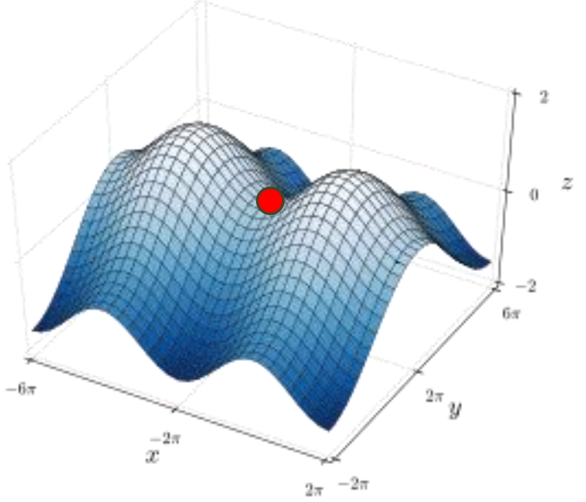
$\alpha_i \ge 0$ and β_i are called the Lagrangian multipliers

The KKT Conditions

- The Lagrangian function

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

- If there exists some **saddle point** of L, then the saddle point satisfies the following "Karush-Kuhn-Tucker" (KKT) conditions:



1. Stationarity: $\frac{\partial L}{\partial w} = 0$
2. Primal feasibility: $g_i(w) \leq 0, \;\; h_i(w) = 0$
3. Dual feasibility: $\alpha_i \geq 0$
4. Complementary slackness: $\alpha_i g_i(w) = 0$

Properties of Lagrangian

- For any feasible dual variables (Lagrangian multipliers) $\alpha_i \geq 0$ and β_i

- We have:

$$\min_{\substack{w \text{ feasible:} \\ g_i(w) \leq 0,\ h_i(w)=0\ \forall i}} f(w) \geq \inf_w L(w, \alpha, \beta)$$

Why?

- This is great! Can we find the best lower bound by optimizing over $\alpha_i \geq 0$ and β_i?

$$\min_{\substack{w \text{ feasible:} \\ g_i(w) \leq 0,\ h_i(w)=0\ \forall i}} f(w) \geq \max_{\substack{\alpha,\beta \text{ feasible} \\ \alpha_i \geq 0\ \forall i}} \inf_w L(w, \alpha, \beta)$$

Why?

Lagrangian Dual Problem

- **Dual problem**: maximizing the lower bound

$$\max_{\substack{\alpha,\beta \text{ feasible} \\ \alpha_i \geq 0\ \forall i}} \inf_w L(w, \alpha, \beta)$$

  - Dual objective: $g(\alpha, \beta) := \inf_w L(w, \alpha, \beta) = \inf_w f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$
  - Dual variables: α, β
  - Dual constraints: $\alpha \geq 0$

- **Dual problem (in a different form)**:

$$\max_{\alpha,\beta} g(\alpha, \beta)$$
$$s.t.\ \alpha \geq 0$$

Strong and Weak Duality

- **Primal problem**:

$$p^* = \min_w f(w)$$
$$s.t.\ g_i(w) \le 0, \qquad i = 1,2,\dots,k$$
$$h_i(w) = 0, \qquad i = 1,2,\dots,l$$

- **Dual problem**: maximizing the lower bound

$$d^* = \max_{\substack{\alpha,\beta \text{ feasible} \\ \alpha_i \ge 0\ \forall i}} g(\alpha,\beta), \qquad where\ g(\alpha,\beta) = \inf_w L(w,\alpha,\beta)$$

- **Duality gap**: $p^* - d^* \ge 0$
  - **Weak duality**: we know that $p^* \ge d^*$ always holds
  - **Strong duality**: $p^* - d^* = 0$

SVM Dual Problem

Dual Problem of Support Vector Machines

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

$$s.t. \, 1 - y^i\left(w^\top x^i + b\right) \leq 0, \qquad \forall i$$

- The Lagrangian function

$$L(w, b, \alpha) = \frac{1}{2} w^\top w + \sum_{i=1}^{n} \alpha_i \left(1 - y^i\left(w^\top x^i + b\right)\right)$$

Deriving the Dual Problem

- $L(w, b, \alpha) = \frac{1}{2} w^\top w + \sum_{i=1}^{n} \alpha_i \left(1 - y^i (w^\top x^i + b)\right)$

- **Dual objective**: $\mathrm{g}(\alpha) \coloneqq \inf_{w,b} L(w, b, \alpha)$

- Taking derivative and set to zero to find optimal w and b

$$\frac{\partial L}{\partial w} = w^* - \sum_{i=1}^{n} \alpha_i y^i x^i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y^i = 0$$

Plug Back the Relation of w and b

$$g(\alpha) := L(w^*, b^*, \alpha) = \frac{1}{2} {w^*}^\top w^* + \sum_{i=1}^{n} \alpha_i \left(1 - y^i \left({w^*}^\top x^i + b\right)\right)$$

$$= \frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i y^i x^i\right)^\top \left(\sum_{j=1}^{n} \alpha_j y^j x^j\right) + \sum_{i=1}^{n} \alpha_i \left(1 - y^i \left(\left(\sum_{j=1}^{n} \alpha_j y^j x^j\right)^\top x^i + b\right)\right)$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \left({x^i}^\top x^j\right) - b \underbrace{\sum_{i=1}^{n} \alpha_i y^i}_{0}$$

The Dual Problem of SVM

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \left({x^i}^{\top} x^j \right)$$

$$s.t.\, \alpha_i \geq 0 \quad \forall i = 1,2, \dots, m$$

$$\underbrace{\sum_{i=1}^{n} \alpha_i y^i = 0}_{\text{Remember to put constraints back!}}$$

- This is a constrained quadratic program
- Nice and **concave**, and global maximum can be found
- How to use the dual solution α to make prediction/classification?

Why?

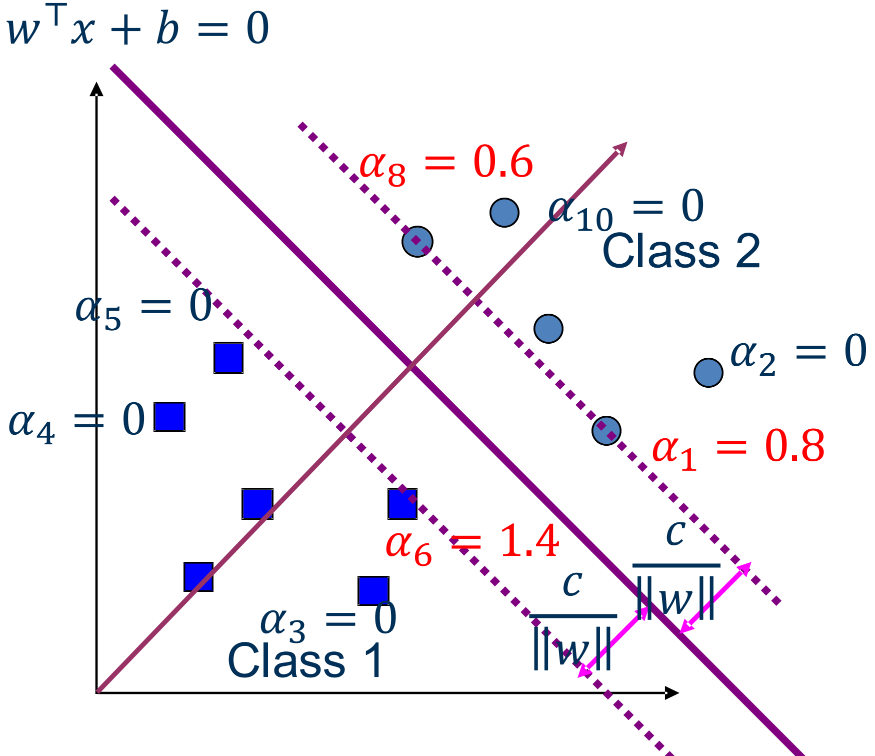
Inference and Support Vectors

Support Vectors

- Note that the KKT condition $\alpha_i\, g_i(w) = 0$ (complementary slackness)

$$\alpha_i \left(1 - y^i\left(w^\top x^i + b\right)\right) = 0$$

  - For data points with $\left(1 - y^i\left(w^\top x^i + b\right)\right) < 0$, then $\alpha_i = 0$
  - For data points with $\left(1 - y^i\left(w^\top x^i + b\right)\right) = 0$, then $\alpha_i \geq 0$



Call the training data points whose α_i's are nonzero the **support vectors** (SV)

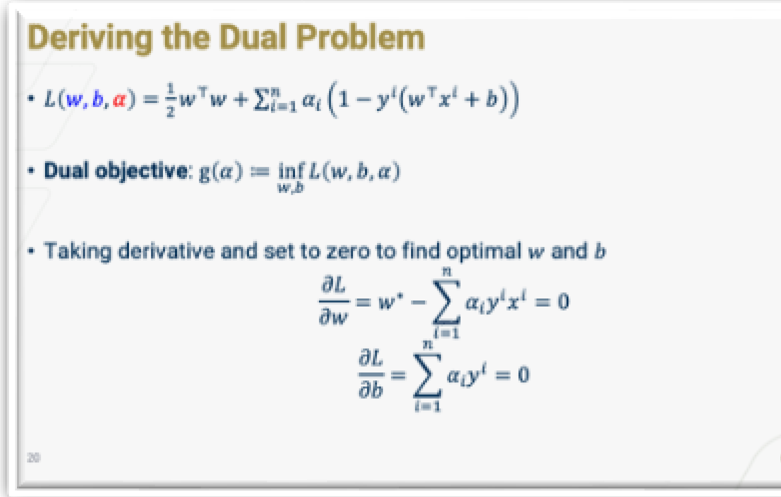
- Those points that support the margins

Classification by Support Vectors

- Recall that we have:

$$w = \sum_{i=1}^{n} \alpha_i y^i x^i$$

Slide 20:



Deriving the Dual Problem

- $L(w,b,\alpha) = \frac{1}{2} w^\top w + \sum_{i=1}^{n} \alpha_i \left(1 - y^i (w^\top x^i + b)\right)$
- **Dual objective**: $g(\alpha) := \inf_{w,b} L(w,b,\alpha)$
- Taking derivative and set to zero to find optimal w and b

$$\frac{\partial L}{\partial w} = w^* - \sum_{i=1}^{n} \alpha_i y^i x^i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y^i = 0$$

20

- Pick any data point with $\alpha_i > 0$ to compute b by:

$$1 - y^i\left(w^\top x^i + b\right) = 0$$

- For a new test point z
  - Compute

$$w^\top z + b = \left(\sum_{i=1}^{n} \alpha_i y^i \left({x^i}^\top z\right)\right) + b = \left(\sum_{i \in support\ vectors} \alpha_i y^i \left({x^i}^\top z\right)\right) + b$$

  - Classify z as class 1 if the result is negative (or $w^\top x + b \leq -c$)
  - Classify z as class 2 if the result is positive (or $w^\top x + b \geq c$)

Interpretation of Support Vector Machines

- The optimal solution $w = \sum_{i=1}^{n} \alpha_i y^i x^i$ is a linear combination of a small number of data points. This is a sparse and memory friendly representation.

- To use support vector machines, we need to specify only the inner products (or kernel) between the examples ${x^i}^\top x^j$

- We make decisions by comparing each new example z with only the support vectors:

$$y^* = sign\left(\left(\sum_{i \in support\ vectors} \alpha_i y^i \left({x^i}^\top z\right)\right) + b\right)$$

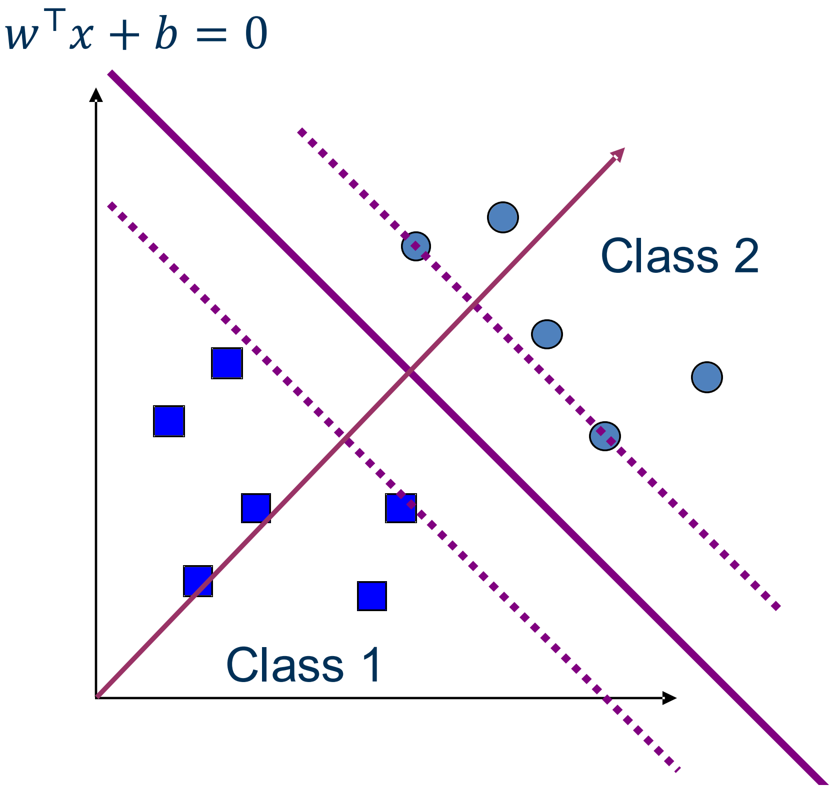
Kernel (nonlinear) SVM

Support Vector Machine (Dual Problem)

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \left({x^i}^\top x^j \right)$$

$$s.t.\, \alpha_i \geq 0 \quad \forall i = 1,2, \dots, m$$

$$\sum_{i=1}^{n} \alpha_i y^i = 0$$



Kernel SVM

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \, K(x^i, x^j)$$

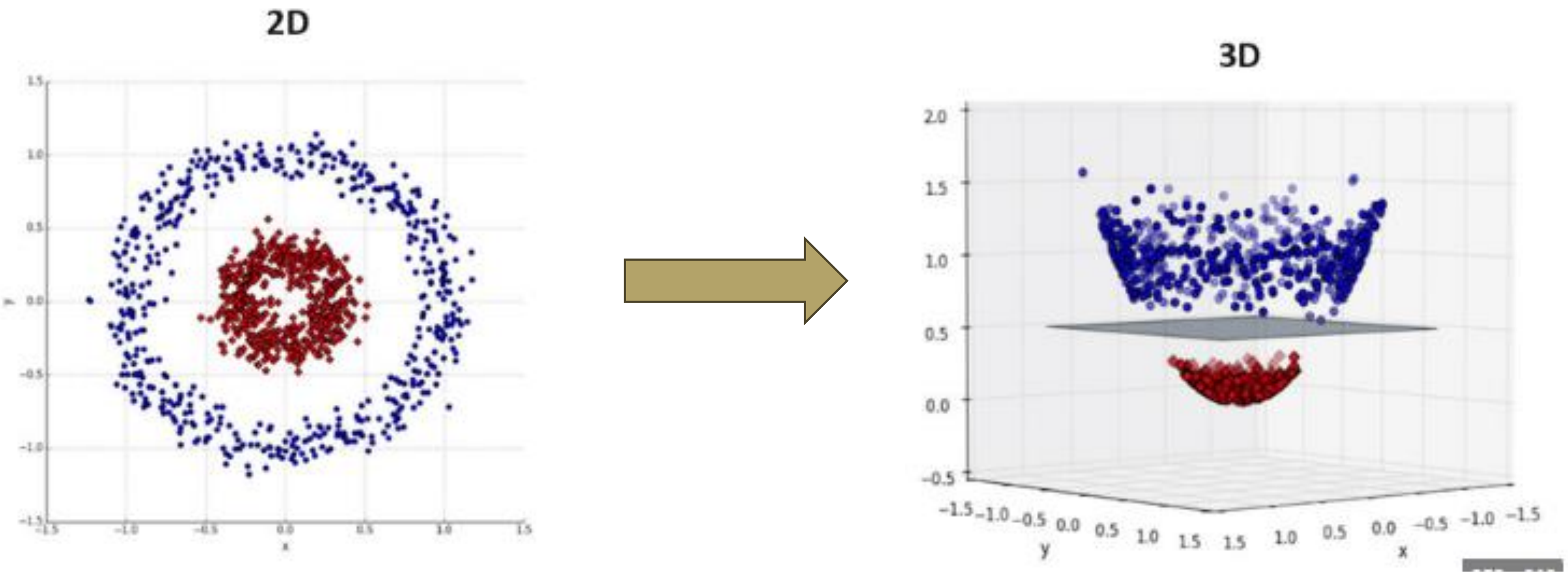
$$s.t.\, \alpha_i \geq 0 \quad \forall i = 1,2, \dots, m$$

$$\sum_{i=1}^{n} \alpha_i y^i = 0$$

- Kernel $K(x, x')$ measures the similarity of any pair $x, x' \in X$
  - **General kernel**: $K: X \times X \rightarrow \mathbb{R}$
  - **Inner product-like kernel**: $K(x, x') = \langle \phi(x), \phi(x') \rangle_V$, where $\phi: X \rightarrow V$

  - Map the data points x^i into a higher-dimensional space $\phi(x^i)$ where the separation becomes easier.

Kernel SVM



- What kernel to use in SVM?
  - Kernel $K(x, x')$ measures the similarity between two points $x, x' \in X$
  - Pick your similarity measure!

Image by Suraj Yadav

Advantage of SVM (Primal v.s. Dual)

- **Short answer**: **kernel**
- **Long answer: keeerrneeel**

- **Primal**: no good kernel extension
- **Dual**: naturally extend to different kernels

- **Primal**: learns w, b and classifies z by computing $w^\top z + b$
- **Dual**: learns α and classifies z by computing $\sum_{i \in support\ vectors} \alpha_i y^i \left({x^i}^\top z \right) + b$

Outline

- **Supervised Learning**
    - Support Vector Machine (SVM)
        - Decision boundary
        - Maximum margin problem
        - Lagrangian duality
        - Dual problem of SVM
        - Inference and support vectors
        - Kernel SVM

    - Soft-margin SVM (Sep 22^{nd})