CSE/ISyE 6740
Computational Data Analytics

Support Vector Machine

09/22/2025

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

Georgia Tech

Outline

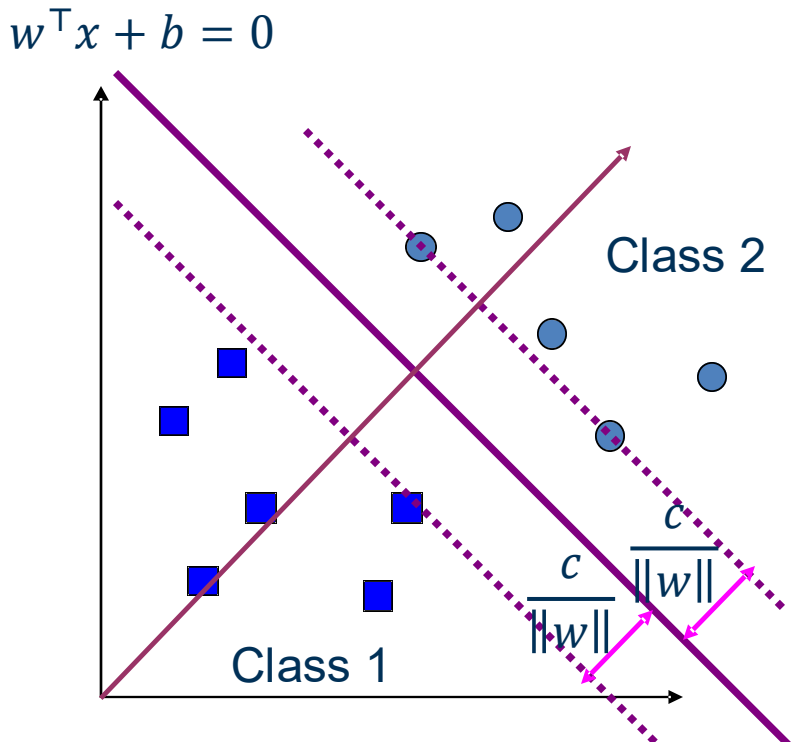
- **Supervised Learning**
  - **Support Vector Machine**
    - From hard-margin SVM to soft-margin SVM
    - Primal and dual problem of soft-margin SVM
    - Hinge loss
    - Support vectors
    - Comparison of SVM and logistic regression

From Hard Margin to Soft Margin SVM

Hard-margin SVM

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

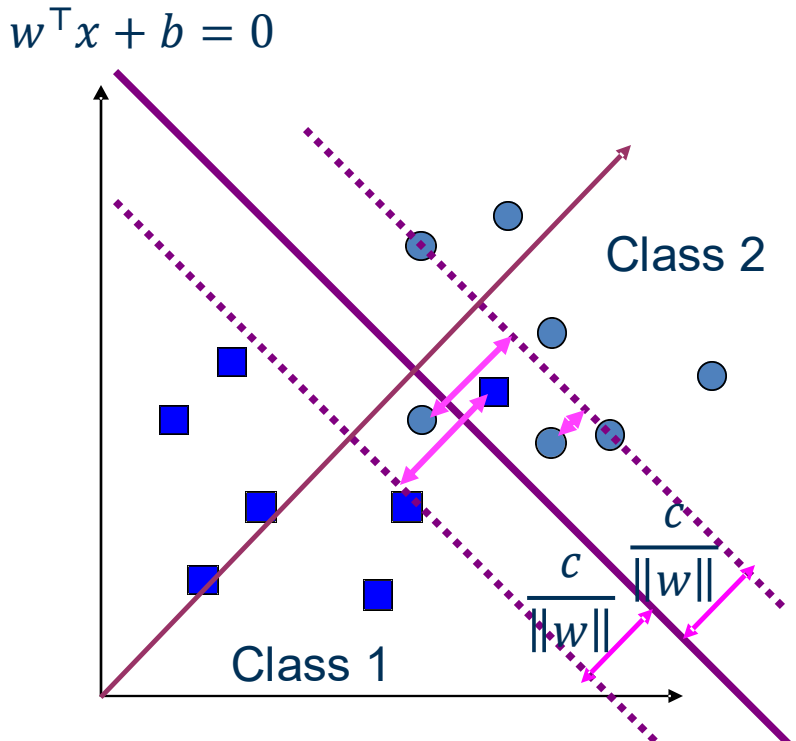
$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$s.t.\ y^i\left(w^\top x^i + b\right) \geq 1, \qquad \forall i$$



Georgia Tech

Soft-margin SVM

- What if the data is **NOT** linearly separable?
- We will allow points to violate the hard-margin constraint



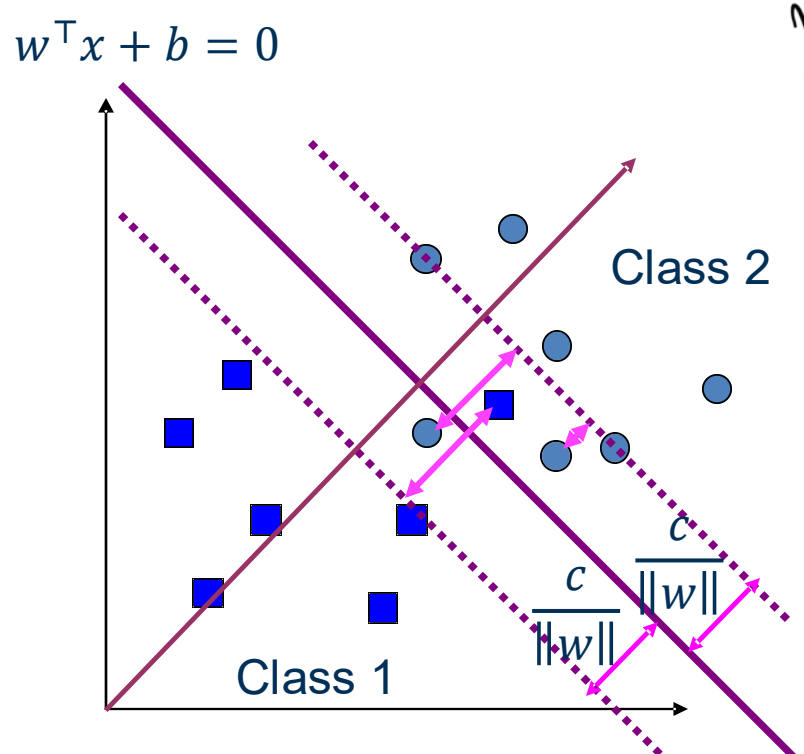
Soft-margin SVM

- Find decision boundary w as far from data point as possible (with some exceptions ξ)

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi^i$$

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

penalty should be positive



Georgia Tech

Soft-margin SVM Primal Problem

- Express the optimization problem in standard form:

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

Soft-Margin SVM Dual Problem

Deriving the Dual Problem

put C as C_i

constant number.

=> there is diff weights on diff penalty.

- Define Lagrangian:

$$L(w, b, \xi, \alpha, \beta)$$

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

- Taking derivative and set to zero:

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

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

$$\frac{\partial L}{\partial \xi^i} = C - \alpha_i - \beta_i = 0$$

Plug Back Relation of w, b, and ξ

$$g(\alpha,\beta) \coloneqq \inf_{w,b,\xi} L(w,b,\xi,\alpha,\beta) = L(w^*,b^*,\xi^*,\alpha,\beta)$$

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

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

- After simplification

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

The Dual Problem

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

$$s.t. C - \alpha_i - \beta_i = 0, \qquad \alpha_i \geq 0, \qquad \beta_i \geq 0 \qquad \forall i$$

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

- The constraint $C - \alpha_i - \beta_i = 0,\ \alpha_i \geq 0,\ \beta_i \geq 0$ can be simplified to $C \geq \alpha_i \geq 0$
- This is a constrained quadratic programming
- Nice and concave, and global maximum can be found

The Dual Problem

soft margin. SVM

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \left(x^{i\top} x^j \right)$$
$$s.t. \; C \geq \alpha_i \geq 0 \quad \forall i$$
$$\sum_{i=1}^{n} \alpha_i y^i = 0$$

- **Intuition**: → no points violate margin.
  - When $C \to \infty$, then this becomes the hard margin SVM.
  - Nice and concave, and global maximum can be found

Hard-margin SVM

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \left(x^{i\top} x^j \right)$$
$$s.t. \; \alpha_i \geq 0 \quad \forall i = 1,2,\dots,m$$
$$\sum_{i=1}^{n} \alpha_i y^i = 0$$

Soft-Margin Kernel SVM

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j K(x^i, x^j)$$
$$s.t. \; C \geq \alpha_i \geq 0 \quad \forall i$$
$$\sum_{i=1}^{n} \alpha_i y^i = 0$$

Hard-margin kernel SVM

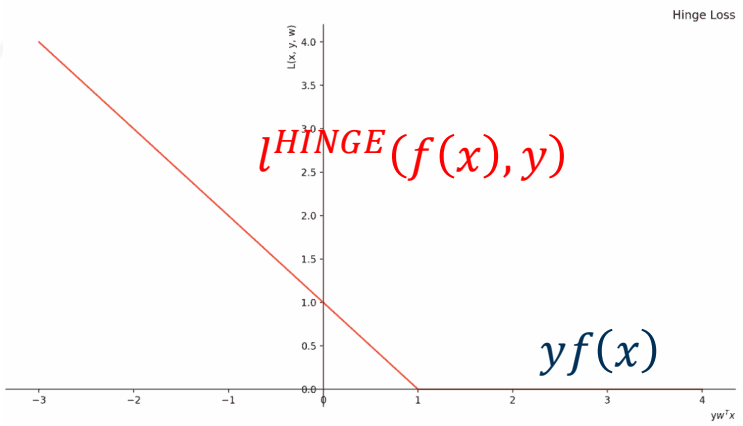
$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j K(x^i, x^j)$$
$$s.t. \; \alpha_i \geq 0 \quad \forall i = 1,2,\dots,m$$
$$\sum_{i=1}^{n} \alpha_i y^i = 0$$

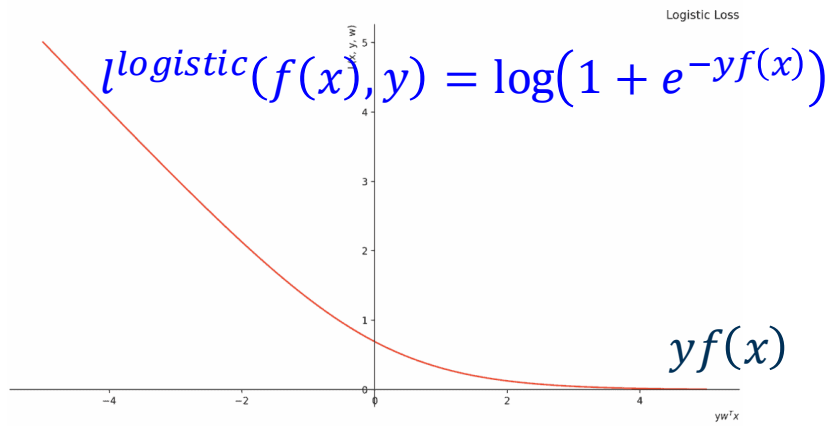
Hinge Loss

Hinge Loss

- **Definition**: Assuming the label $y \in \{-1,1\}$ and the decision rule is $h(x) = sign(f(x))$ with $f(x) = w^{\mathsf{T}}\phi(x) + b$

$$l^{HINGE}(f(x), y) = \begin{cases} 0 & if\ yf(x) \geq 1 \\ 1 - yf(x) & otherwise \end{cases}$$





- **Intuition**: penalize more if incorrectly classified (the left size of the curve)
- **Convenient shorthand**:

$$l^{HINGE}(f(x), y) = \max(0, 1 - yf(x))$$

Equivalence of Hinge Loss to SVMs

- Minimizing the total hinge loss on all the training data

$$\min_{w,b} \sum_{i=1}^{n} max(0, 1 - y^i[w^\top \phi(x^i) + b]) + \frac{\lambda}{2}\|w\|^2$$

Regularization

Hinge loss

- This is analogous to the regularized least square, which balances two terms (the loss and the regularizer). Conventionally, we rewrite the objective function as ($C = \frac{1}{\lambda}$.):

$$\min_{w,b} C \sum_{i=1}^{n} max(0, 1 - y^i[w^\top \phi(x^i) + b]) + \frac{1}{2}\|w\|^2$$

- We further rewrite this into another equivalent form

$$\min_{w,b,\xi} C \sum_{i=1}^{n} \xi^i + \frac{1}{2}\|w\|^2$$

$$s.t. max(0, 1 - y^i[w^\top \phi(x^i) + b]) = \xi^i \quad \forall i$$

Equivalence of Hinge Loss to SVMs

- **Primal formulation**:

$$\min_{w,b,\xi} C\sum_{i=1}^{n}\xi^i + \frac{1}{2}\|w\|^2$$
$$s.t.\ 1 - y^i\left[w^\top\phi\left(x^i\right) + b\right] \le \xi^i \quad \forall i$$
$$0 \le \xi^i$$

Where all ξ^i are called slack variables.

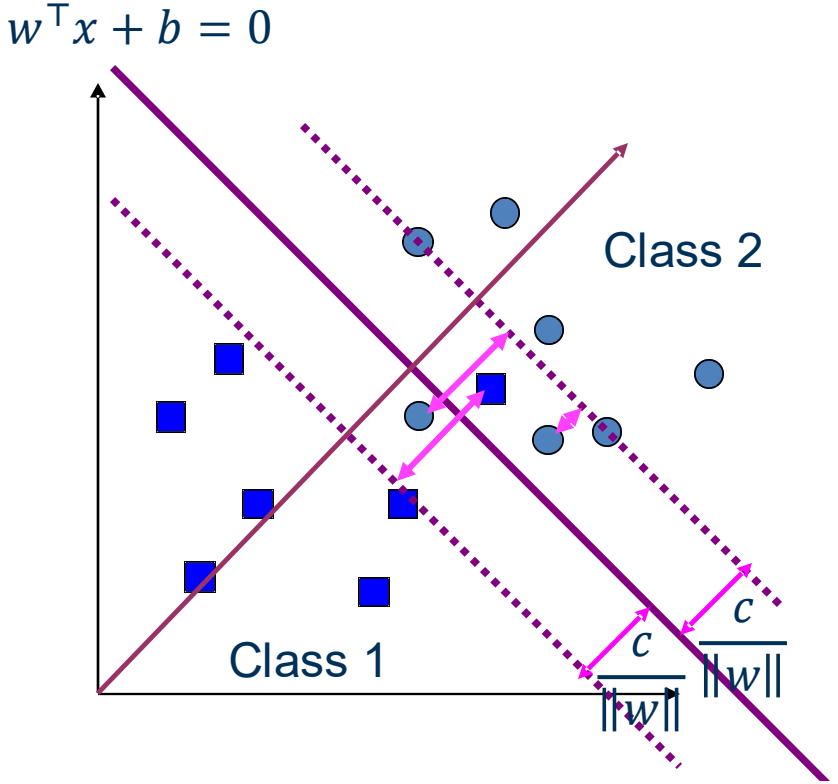
Soft-margin SVM

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^i$$
$$s.t.\ y^i\left(w^\top x^i + b\right) \ge 1 - \xi^i,$$
$$\xi^i \ge 0 \quad \forall i$$

Support Vectors

Complementary Slackness

- What is the geometric meaning of α_i, β_i?



Primal problem

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^i$$
$$s.t.\ 1 - \xi^i - y^i\left(w^\top x^i + b\right) \le 0,$$
$$-\xi^i \le 0 \quad \forall i$$

Dual problem

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y^i y^j\left(x^{i\top}x^j\right)$$
$$s.t.\ C \ge \alpha_i \ge 0 \quad \forall i$$
$$\sum_{i=1}^{n}\alpha_i y^i = 0$$

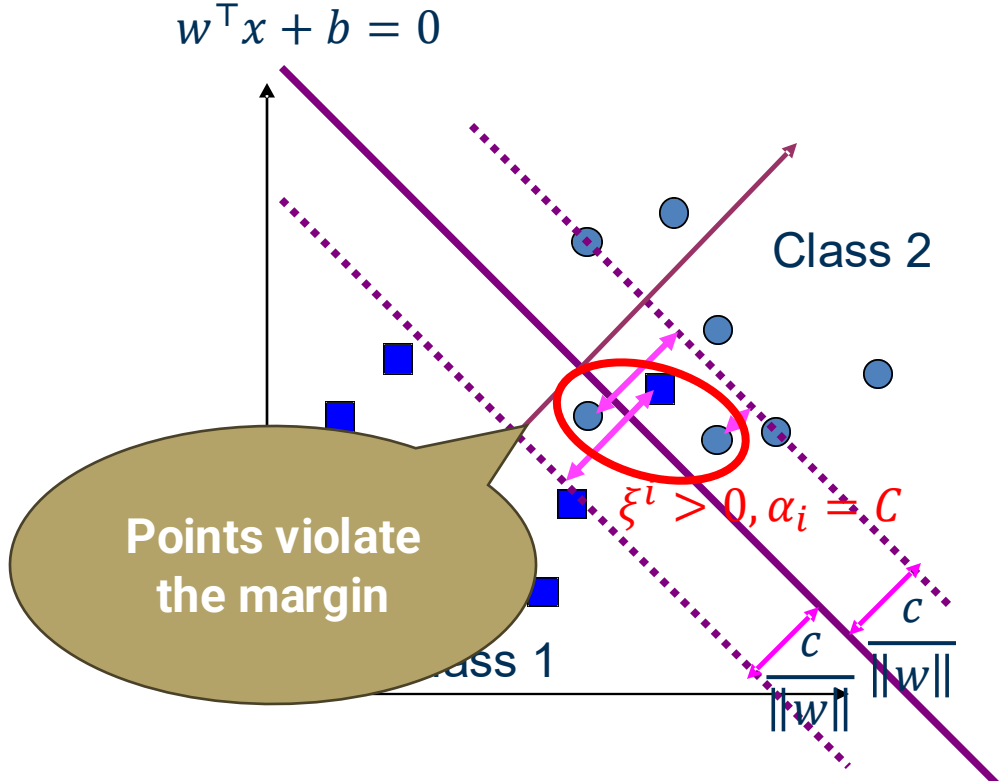
Georgia Tech

Complementary Slackness

- **Non-margin support vectors $\alpha_i = C \neq 0$:**
  - This means $\beta_i = C - \alpha_i = 0$
  - Therefore, $\xi_i > 0$, and $1 - y^i(w^\top x^i + b) = \xi^i > 0$

all the α can be C

$\alpha = C$



Primal problem

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi^i$$
$$s.t.\ 1 - \xi^i - y^i(w^\top x^i + b) \leq 0,$$
$$-\xi^i \leq 0 \quad \forall i$$

Dual problem

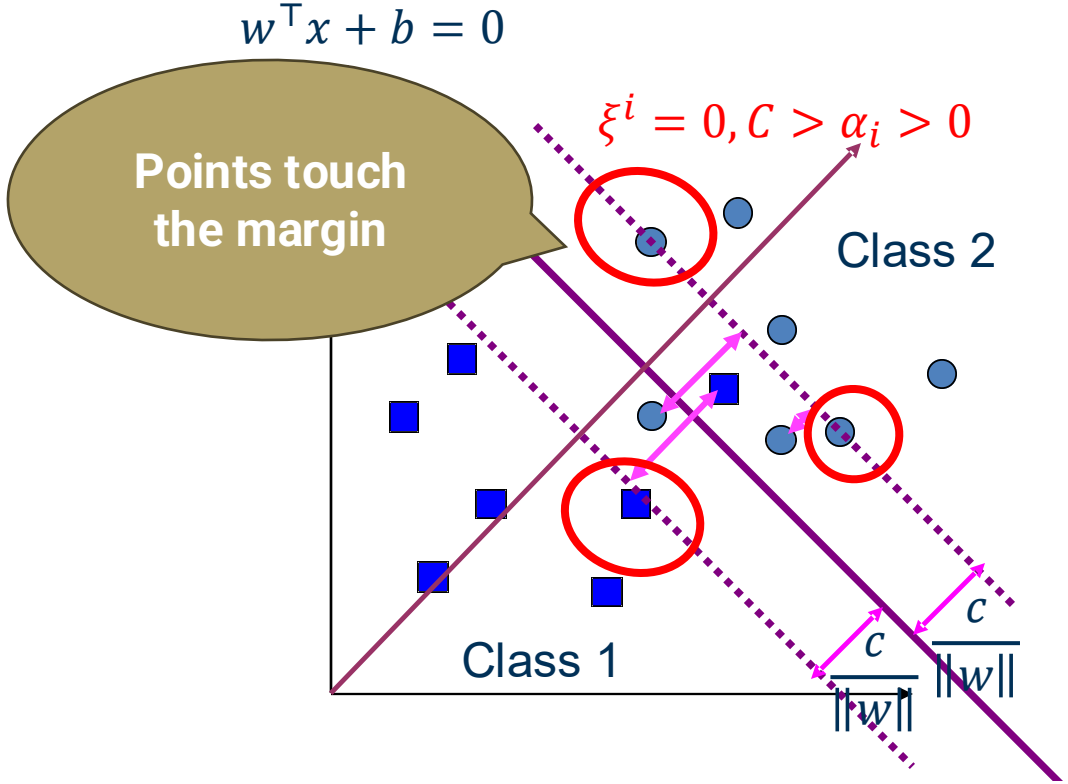
$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j \left(x^{i\top} x^j\right)$$
$$s.t.\ C \geq \alpha_i \geq 0 \quad \forall i$$
$$\sum_{i=1}^{n} \alpha_i y^i = 0$$

Georgia Tech

Complementary Slackness

- **Non-margin support vectors** $\alpha_i = C \neq 0$:
  - This means $\beta_i = C - \alpha_i = 0$
  - Therefore, $\xi_i > 0$, and $1 - y^i\left(w^\top x^i + b\right) = \xi^i > 0$

- **Margin support vectors** $0 < \alpha_i < C$:
  - This means $\beta_i = C - \alpha_i > 0$
  - Therefore, $\xi_i = 0$, and $1 - y^i\left(w^\top x^i + b\right) = \xi^i = 0$



Primal problem

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

Dual problem

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y^i y^j\left(x^{i\top}x^j\right)$$
$$s.t.\ C \geq \alpha_i \geq 0 \quad \forall i$$
$$\sum_{i=1}^{n}\alpha_i y^i = 0$$

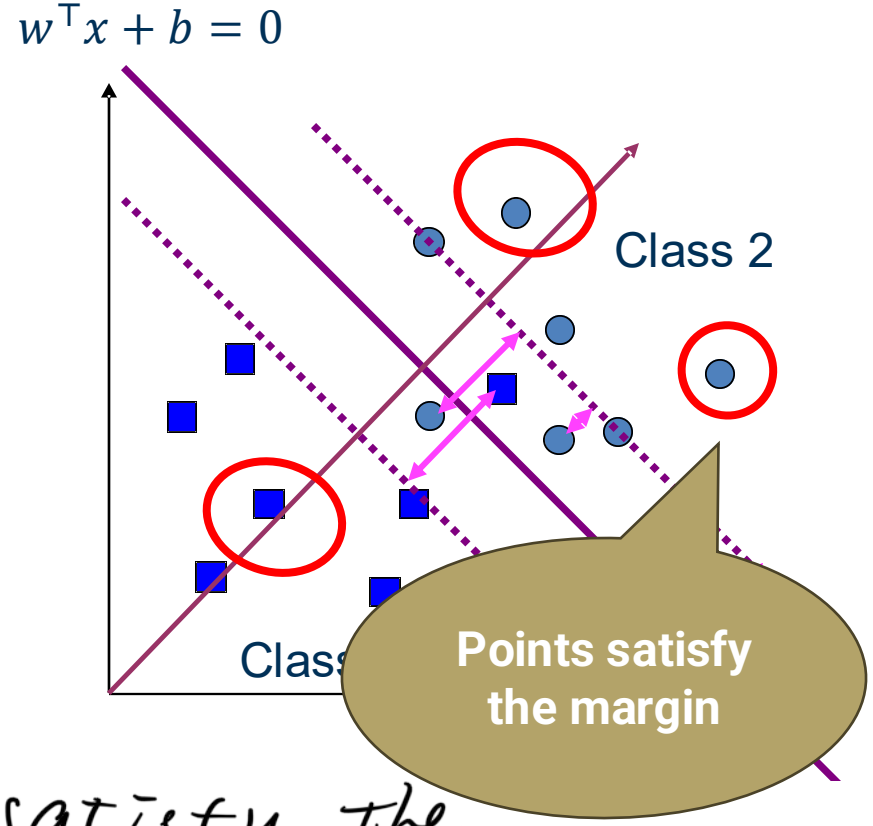
Georgia Tech

Complementary Slackness

- **Non-margin support vectors** $\alpha_i = C \neq 0$:
  - This means $\beta_i = C - \alpha_i = 0$
  - Therefore, $\xi_i > 0$, and $1 - y^i(w^\top x^i + b) = \xi^i > 0$

- **Margin support vectors** $0 < \alpha_i < C$:
  - This means $\beta_i = C - \alpha_i > 0$
  - Therefore, $\xi_i = 0$, and $1 - y^i(w^\top x^i + b) = \xi^i = 0$

- **Non-support vectors** $0 = \alpha_i$: → perfectly satisfy the margin
  - This means $\beta_i = C - \alpha_i = C > 0$
  - Therefore, $\xi_i = 0$, i.e., $1 - y^i(w^\top x^i + b) < \xi^i = 0$



Georgia Tech

Inference of Support Vector Machines

- The inference is the same as hard-margin SVM. We make decisions by comparing each new example z with only the support vectors:

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

Primal

- Primal solution w^*, b^*
- Inference: $sign(w^\top z + b)$
- Computation cost: $O(d)$
- Kernel formulation: no

biggest diff

Dual

- Dual solution α with n' support vectors (nonzero α_i)
- Inference: $sign\left(\left(\sum_{i \in support\ vectors} \alpha_i y^i \left(x^{i\top} z\right)\right) + b\right)$
  - Need to pre-compute the offset $b \in \mathbb{R}$ with O(1) cost
- Computation cost: $O(dn')$
- Kernel formulation: yes

Can get better mapping

Comparison of SVM and Logistic Regression

SVMs v.s. Logistic Regression

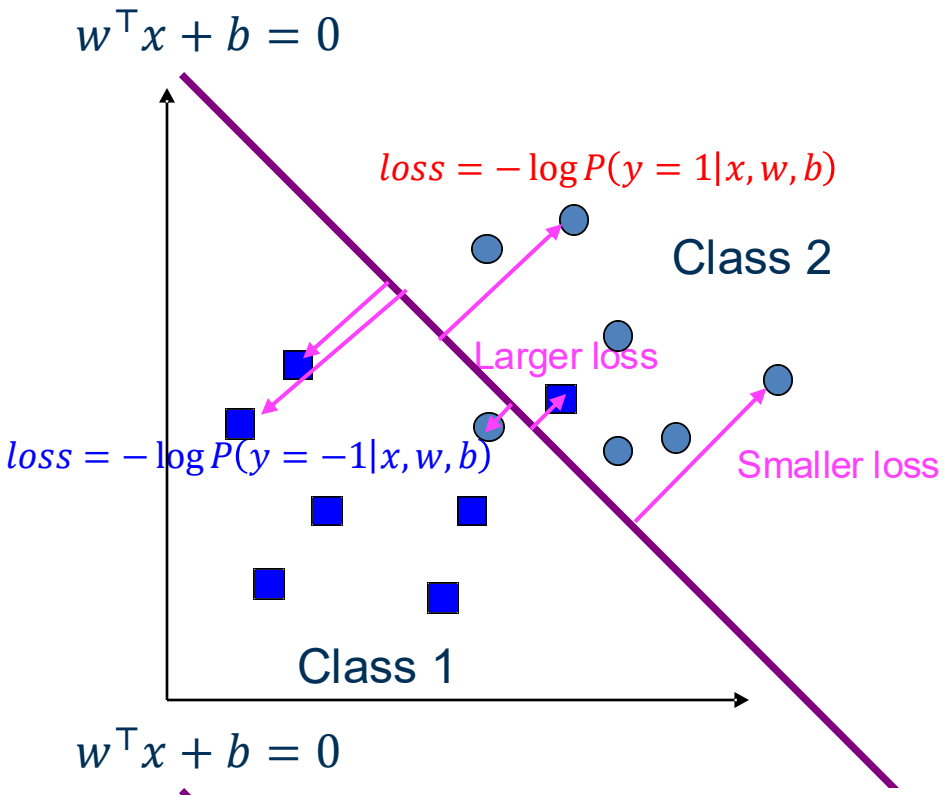
- **Logistic regression** focuses on maximizing the probability of the data. The farther the data lies from the separating hyperplane (on the correct side), the happier LR is.

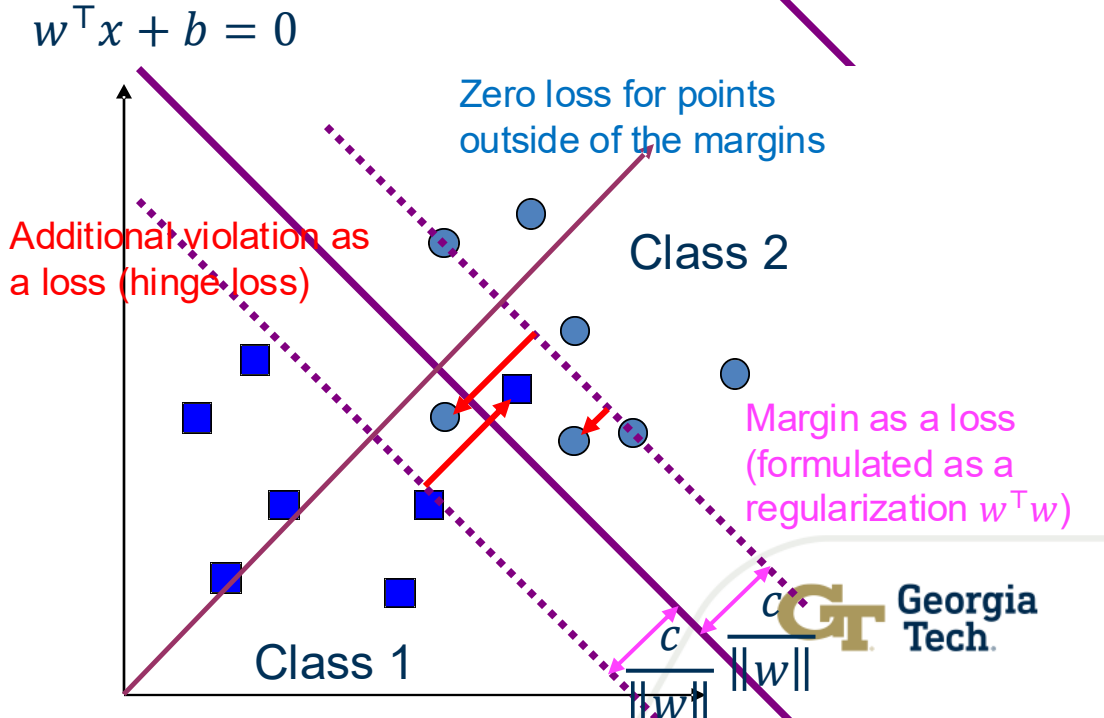
All points have loss.

- **SVM** tries to find the separating hyperplane that maximizes the distance of the closest points to the margin (the support vectors). If a point is not a support vector, it doesn't really matter.

inside margins : loss ✓

outside margins : no loss





Georgia Tech

SVMs v.s. Logistic Regression

	SVMs	Logistic regression
Loss function	Hinge loss	Log-loss

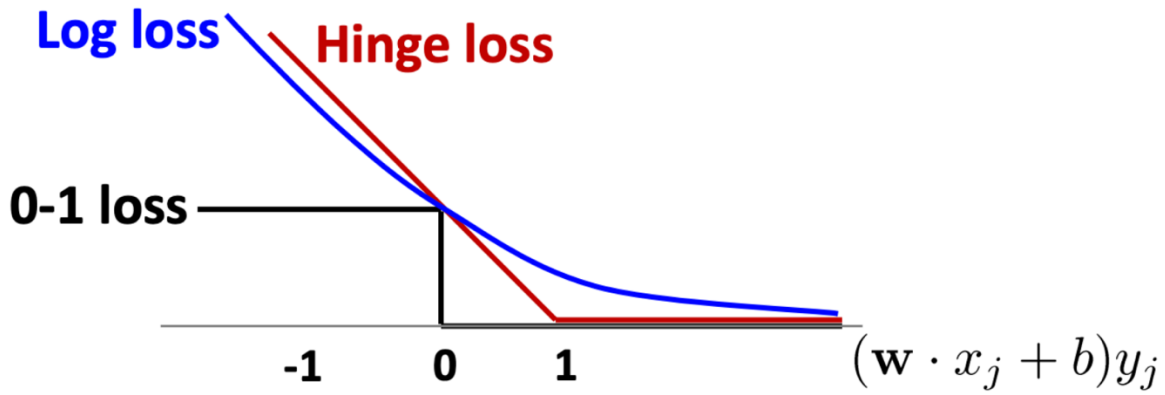
SVMs v.s. Logistic Regression

- **SVM**: **Hinge loss**

$$loss\left(f\left(x^i\right), y^i\right) = \left(1 - \left(w^\top x^i + b\right) y^i\right)_+$$

- **Logistic regression**: **log loss** (negative log likelihood)

$$loss\left(f\left(x^i\right), y^i\right) = -\log P\left(y^i | x^i, w, b\right) = \log\left(1 + e^{-\left(w^\top x^i + b\right) y_i}\right)$$



SVMs v.s. Logistic Regression

	SVMs	Logistic regression
Loss function	Hinge loss	Log-loss
High dimensional features	Yes	Yes
Dual solution sparse	Often yes!	Almost always no!
Semantics of output	Margin	Probabilities

SVMs v.s. Logistic Regression: Complexity

- n: number of training examples
- n': number of support vectors
- d: number of features
- E: number of stochastic gradient descent (SGD) epochs

- **Logistic regression**
  - Train time complexity: $O(ndE)$ (SGD)
  - Test time complexity: $O(d)$

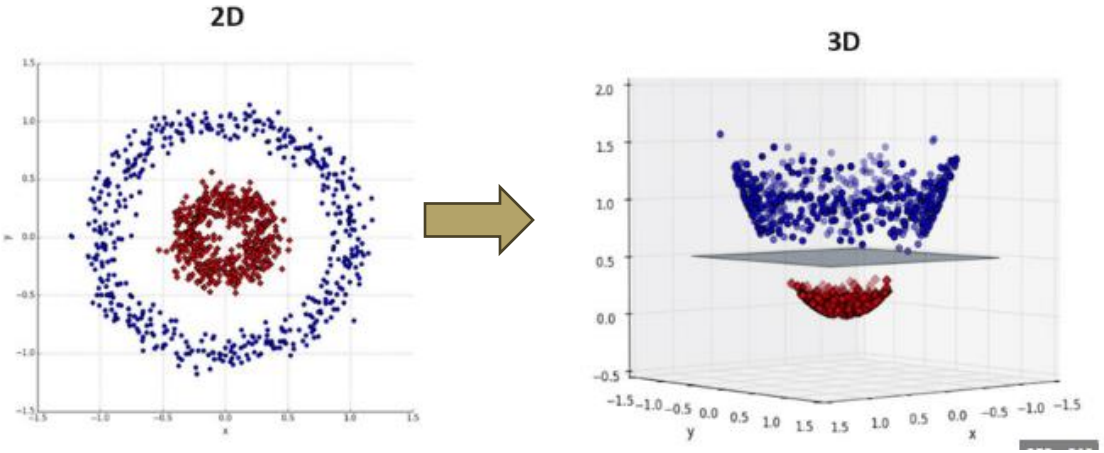
- **SVM**
  - Train time complexity: $O(n^2d)$ (QP) or $O(ndE)$ (SGD)
  - Test time complexity: $O(d)$ (linear SVM) or $O(dn')$ (kernel SVM)

Which One to Use?

- **Logistic regression**
  - LR gives calibrated probabilities that can be interpreted as confidence in a decision, while SVM does not have a direct probabilistic interpretation.
  - LR can be (straightforwardly) used within Bayesian models.
  - LR gives us an unconstrained, smooth objective.

- **SVMs**
  - SVMs don’t penalize examples for which the correct decision is made with sufficient confidence. This may be good for generalization.
  - SVMs have a nice dual form, giving sparse solutions.
  - SVM results are strongly dependent on a suitable choice for the softening parameter C.

Next Lecture: Kernel Methods



- What is the kernel function K?
- What exactly does the kernel method do?
- Why can the kernel method help extend to nonlinearity?

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

$$s.t. \frac{C}{n} \geq \alpha_i \geq 0 \quad \forall i$$

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