18-660: Numerical Methods for Engineering Design and Optimization

Xin Li
Department of ECE
Carnegie Mellon University
Pittsburgh, PA 15213

Electrical & Computer ENGINEERING

Overview

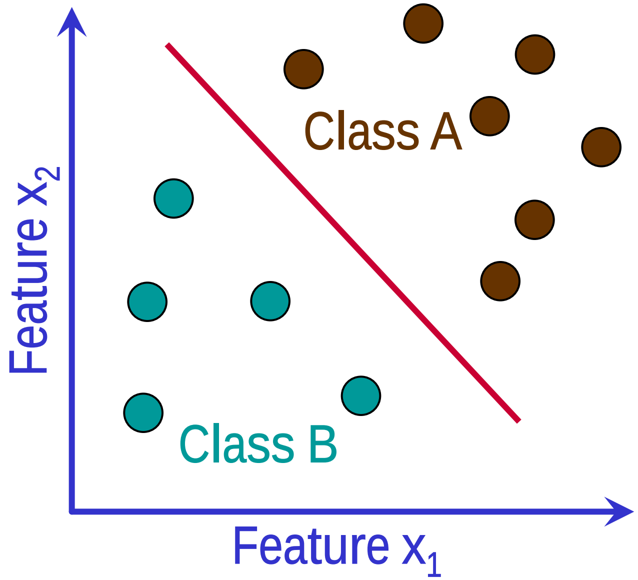
- Classification
  - Support vector machine
  - Regularization

Classification

- Predict categorical output (i.e., two or multiple classes) from input attributes (i.e., features)

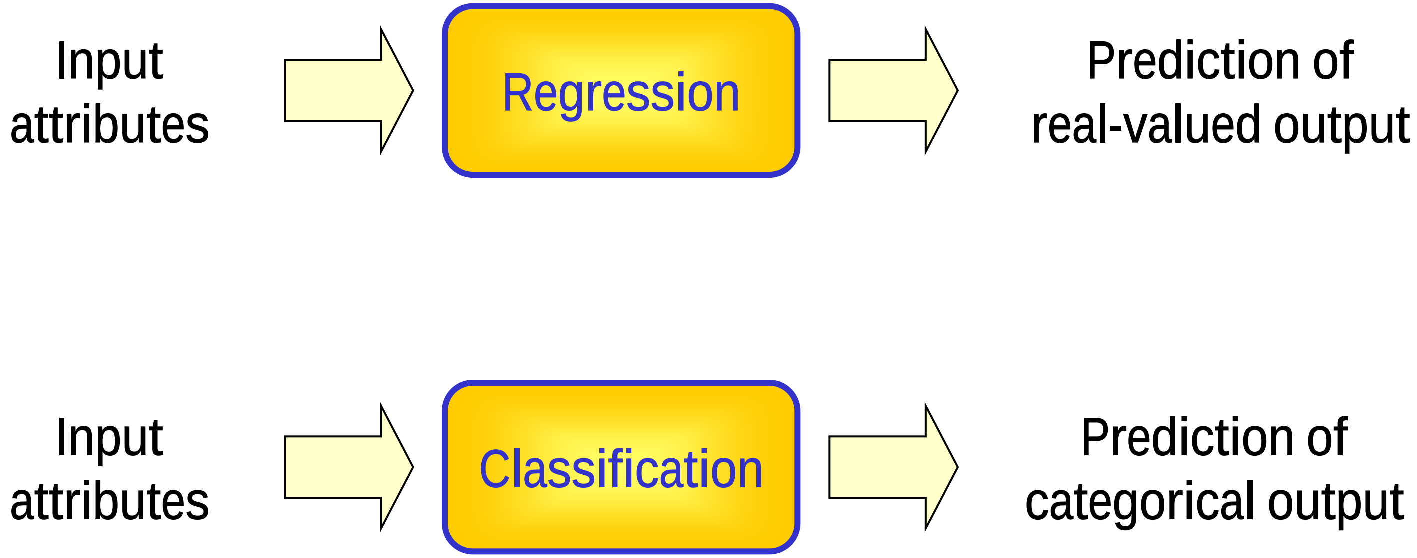
- Example: two-class classification

$$f(X) = W^T X + C \begin{cases} \geq 0 & (Class\ A) \\ < 0 & (Class\ B) \end{cases}$$



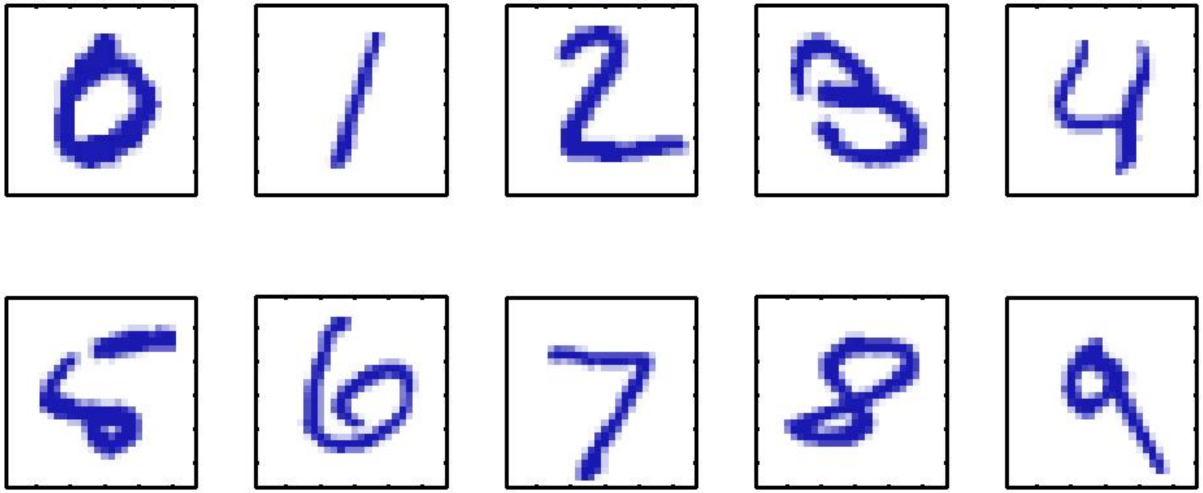
Classification

- Classification vs. regression



Classification Examples

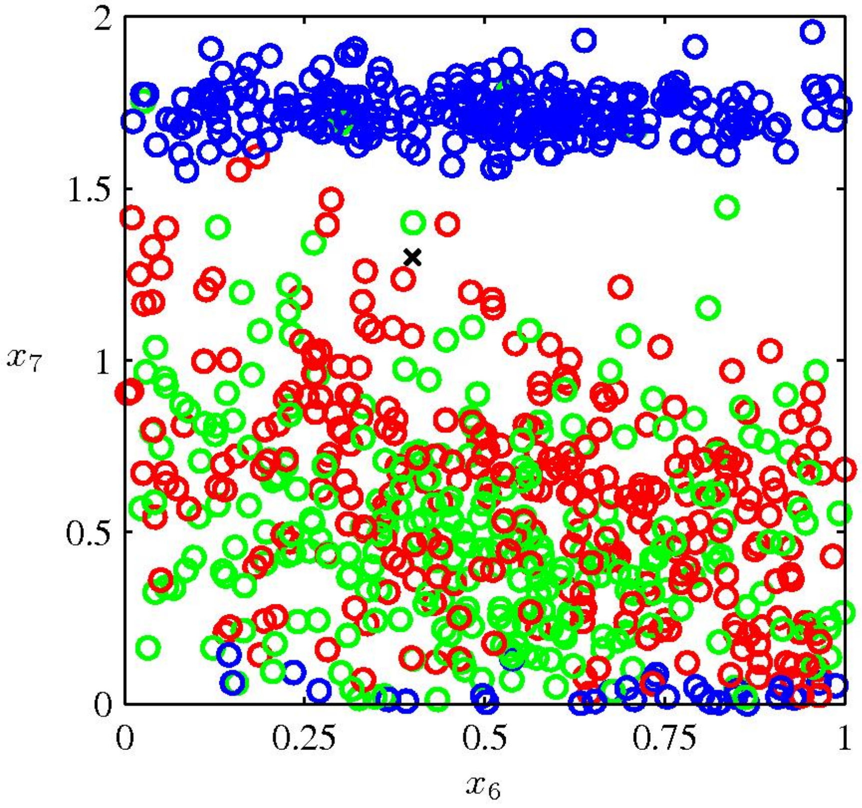
- Identify hand-written digits from US zip codes



Bishop, Pattern recognition and machine learning, 2007

Classification Examples

- Identify geometrical structure from oil flow data



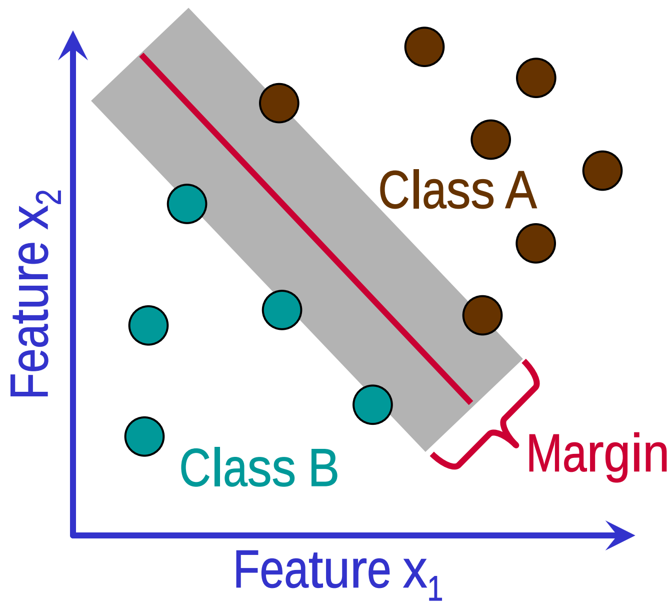
Blue: geometrical structure 1
Green: geometrical structure 2
Red: geometrical structure 3

Bishop, Pattern recognition and machine learning, 2007

Support Vector Machine (SVM)

- Support vector machine (SVM) is a popular algorithm used for many classification problems
  - Key idea: maximize classification margin (immune to noise)

- Two-class linear support vector machine

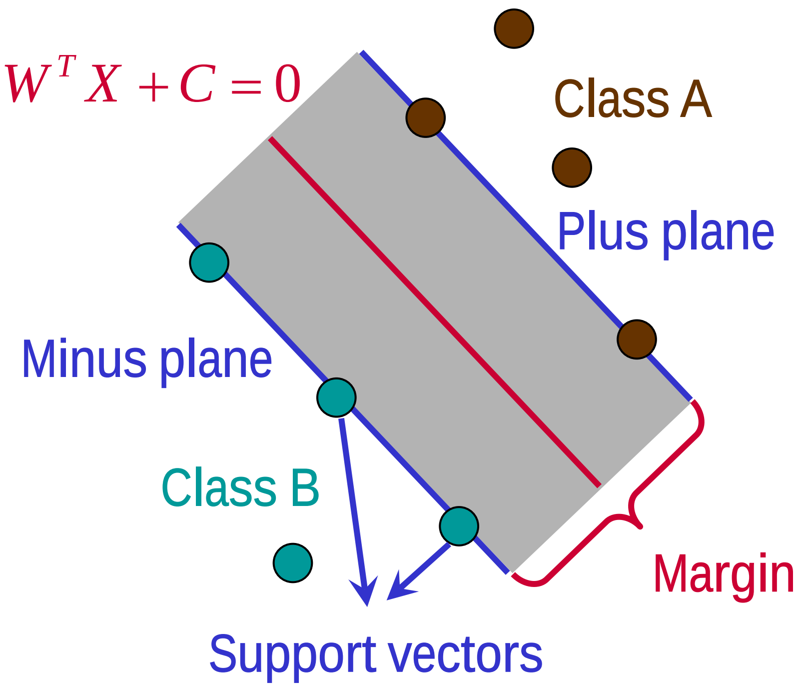


$$f(X) = W^T X + C \quad \begin{cases} \geq 0 & (Class\ A) \\ < 0 & (Class\ B) \end{cases}$$

Determine W and C with maximum margin

Margin Calculation

- To maximize margin, we must first represent margin as a function of W and C



$$f(X) = W^T X + C \quad \begin{cases} \geq 0 & (Class\ A) \\ < 0 & (Class\ B) \end{cases}$$

Plus plane $W^T X + C = 1$

Minus plane $W^T X + C = -1$

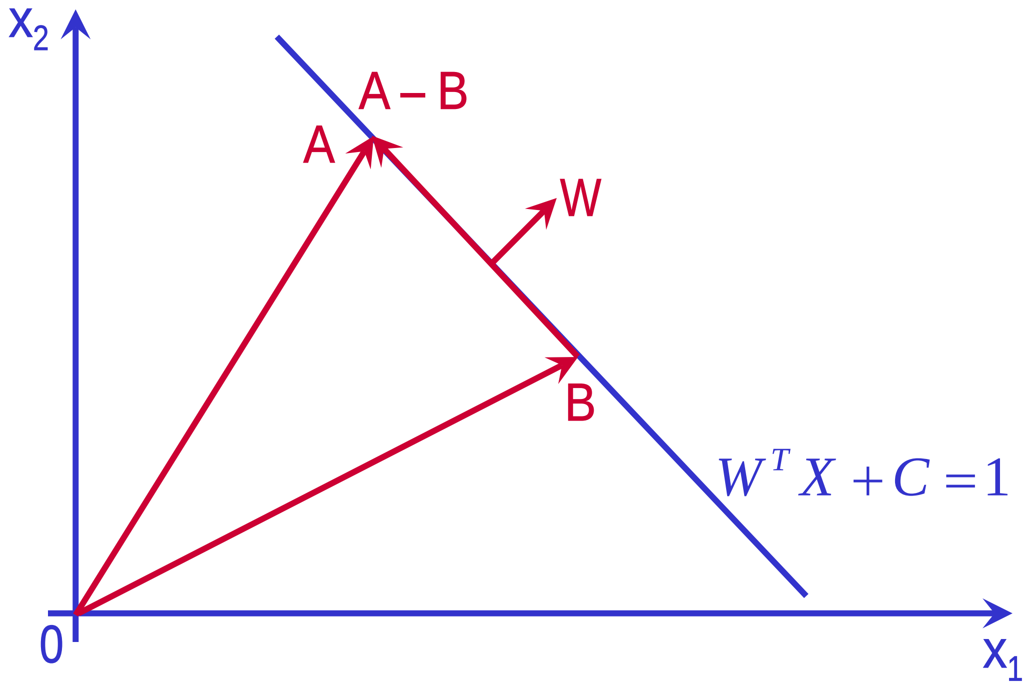
(Right-hand side can be normalized to ±1)

Margin Calculation

- W is perpendicular to plus/minus planes

Plus plane $W^T X + C = 1$

Minus plane $W^T X + C = -1$



$$W^T A + C = 1$$

$$W^T B + C = 1$$

$$\Downarrow$$

$$W^T \cdot (A - B) = 0$$

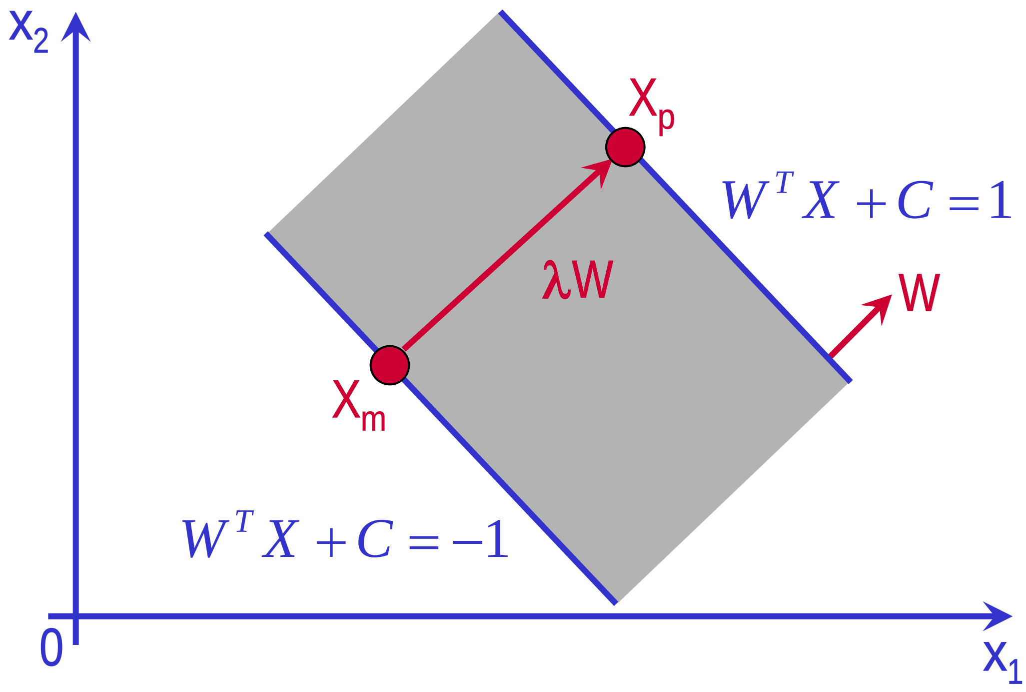
W is perpendicular to (A – B)

Margin Calculation

- Margin equals to the distance between X_m and X_p

$$X_p = X_m + \lambda W \quad \Rightarrow \quad Margin = \left\| X_p - X_m \right\|_2 = \left\| \lambda W \right\|_2$$

Find λ to determine margin



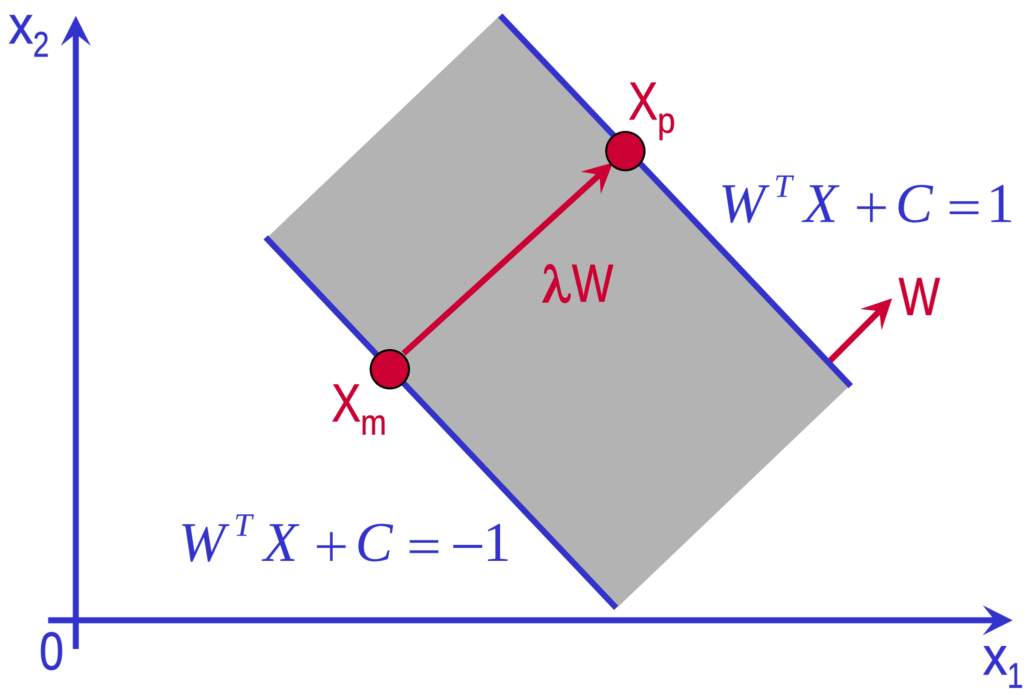
Margin Calculation

$$X_p = X_m + \lambda W$$

$$W^T X_p + C = 1$$

$$W^T X_m + C = -1$$

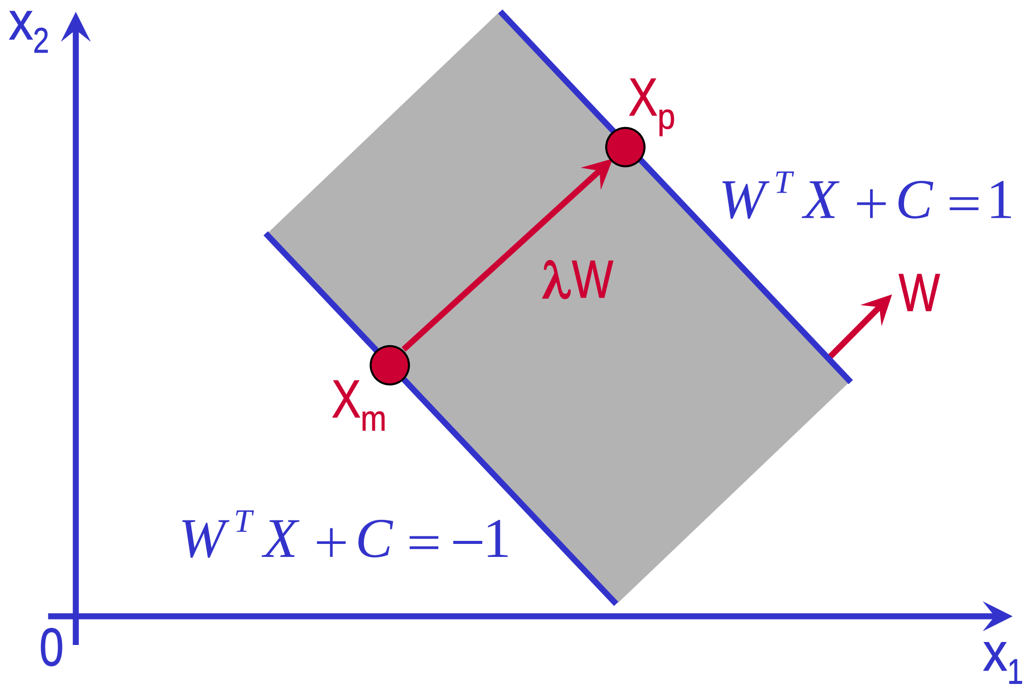
$$\Rightarrow \quad W^T \cdot \left(X_p - X_m \right) = \lambda W^T W = 2$$



Margin Calculation

$$\lambda W^T W = 2 \Rightarrow \lambda = \frac{2}{W^T W} \Rightarrow Margin = \|\lambda W\|_2 = \lambda \cdot \sqrt{W^T W} = \frac{2}{\sqrt{W^T W}}$$

Maximizing margin implies minimizing $||W||_2$



Mathematical Formulation

- Start from a set of training samples

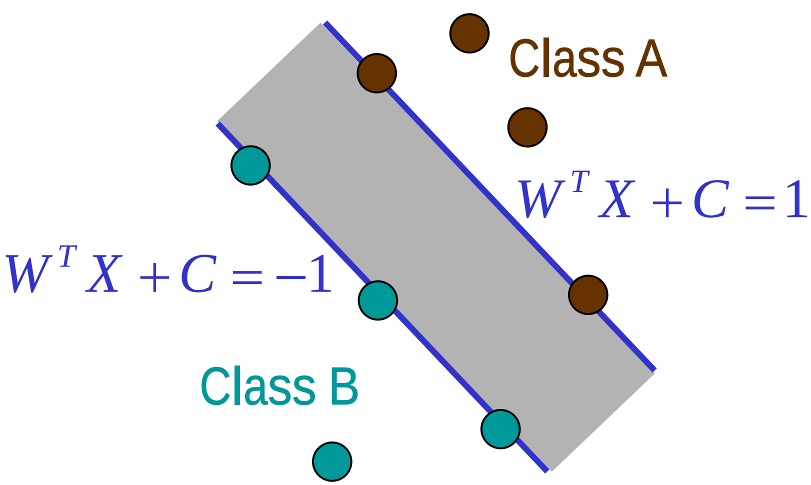
$$(X_i, y_i) \quad (i = 1, 2, \cdots, N)$$

X_i: input feature of i-th sampling point

y_i: output label of i-th sampling point

Class A $\rightarrow$ $y_i = 1$

Class B $\rightarrow$ $y_i = -1$



Class A:

$$W^T X_i + C \geq 1 \quad y_i = 1$$

$$y_i \cdot \left(W^T X_i + C\right) \geq 1$$

Class B:

$$W^T X_i + C \leq -1 \quad y_i = -1$$

$$y_i \cdot \left(W^T X_i + C\right) \geq 1$$

Mathematical Formulation

- Formulate a convex optimization problem

$$\max_{W,C} \frac{2}{\sqrt{W^T W}}$$ → Maximize margin

$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1$$ → All data samples are in the right class

$$\left(i = 1, 2, \cdots, N\right)$$

⇓

$$\min_{W,C} W^T W$$ → Convex quadratic function

$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1$$ → Linear constraints

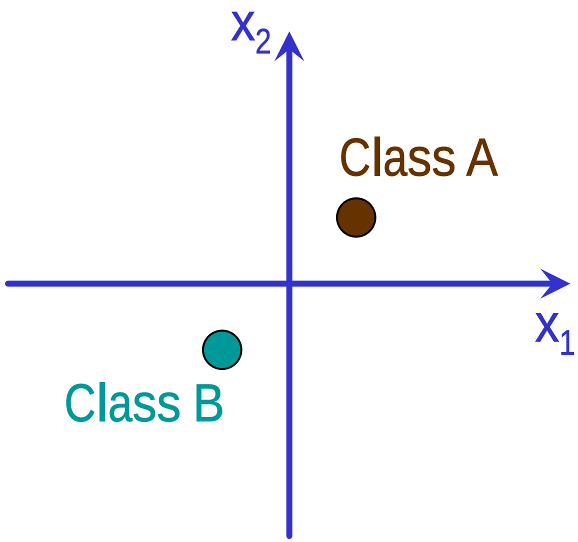
$$\left(i = 1, 2, \cdots, N\right)$$

(Convex optimization)

A Simple SVM Example

- Two training samples
  - Class A: $x_1 = 1$, $x_2 = 1$ and $y = 1$
  - Class B: $x_1 = -1$, $x_2 = -1$ and $y = -1$

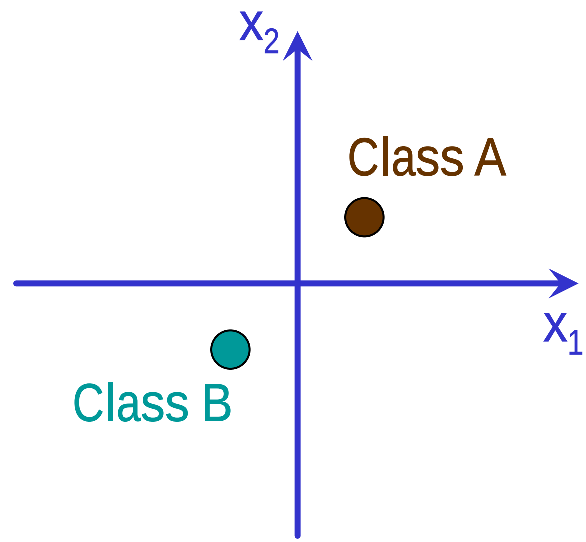
$$f(X) = w_1 x_1 + w_2 x_2 + C \quad \begin{cases} \geq 0 & (Class\ A) \\ < 0 & (Class\ B) \end{cases}$$



Solve w_1, w_2 and C to determine classifier

A Simple SVM Example

- Two training samples
  - Class A: $x_1 = 1$, $x_2 = 1$ and $y = 1$
  - Class B: $x_1 = -1$, $x_2 = -1$ and $y = -1$



$$\min_{W,C} \quad W^T W$$
$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1$$
$$\left(i = 1, 2, \cdots, N\right)$$

$$\min_{W,C} \quad w_1^2 + w_2^2$$
$$\text{S.T.} \quad 1 \cdot \left(w_1 + w_2 + C\right) \geq 1$$
$$-1 \cdot \left(-w_1 - w_2 + C\right) \geq 1$$

A Simple SVM Example

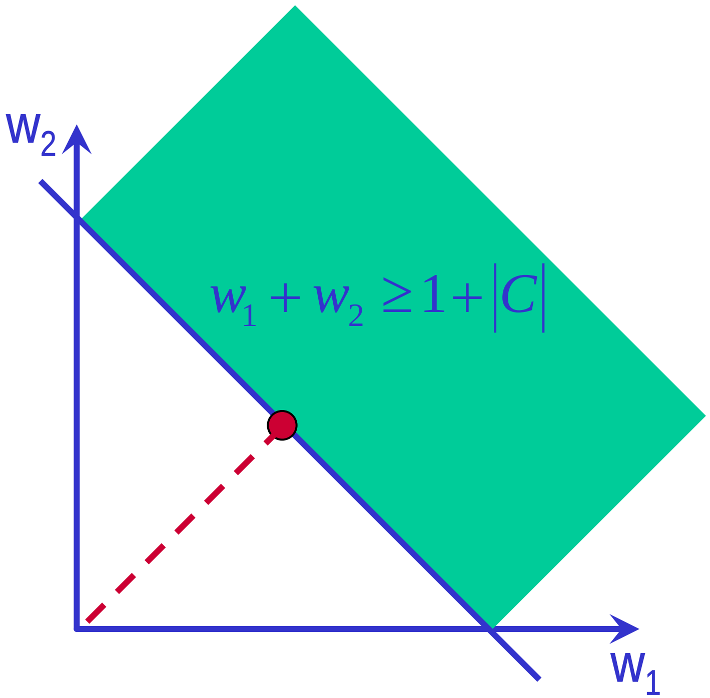
$$\min_{W,C} \quad w_1^2 + w_2^2$$
$$\text{S.T.} \quad 1\cdot(w_1 + w_2 + C) \geq 1$$
$$-1\cdot(-w_1 - w_2 + C) \geq 1$$

⇓

$$\min_{W,C} \quad w_1^2 + w_2^2$$
$$\text{S.T.} \quad w_1 + w_2 \geq 1 - C$$
$$w_1 + w_2 \geq 1 + C$$

⇓

$$\min_{W,C} \quad w_1^2 + w_2^2$$
$$\text{S.T.} \quad w_1 + w_2 \geq 1 + |C|$$



$$w_1 = w_2 = 0.5$$
$$C = 0$$

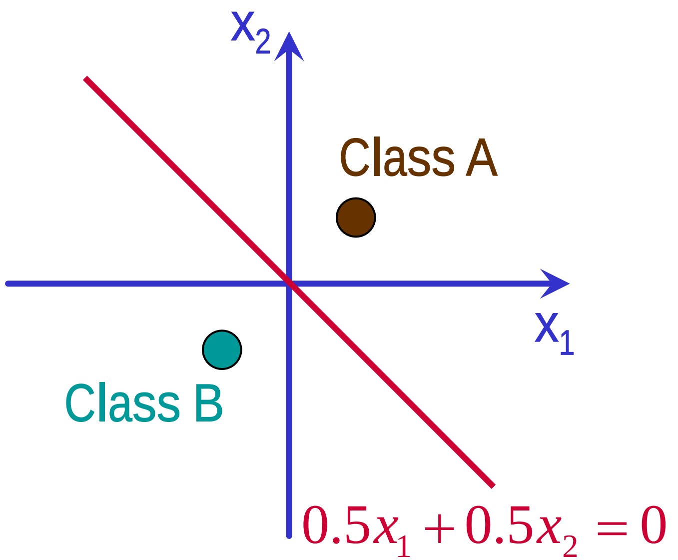
A Simple SVM Example

- Two training samples
  - Class A: $x_1 = 1$, $x_2 = 1$ and $y = 1$
  - Class B: $x_1 = -1$, $x_2 = -1$ and $y = -1$

$$w_1 = w_2 = 0.5$$

$$C = 0$$

$$f(X) = 0.5x_1 + 0.5x_2 \begin{cases} \geq 0 & (Class\ A) \\ < 0 & (Class\ B) \end{cases}$$



Support Vector Machine with Noise

- In practice, training samples may contain noise or are not linearly separable

$$\min_{W,C} \quad W^T W$$

$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1$$

$$\left(i = 1,2,\cdots,N\right)$$

(No feasible solution)

Parameter determined by cross validation

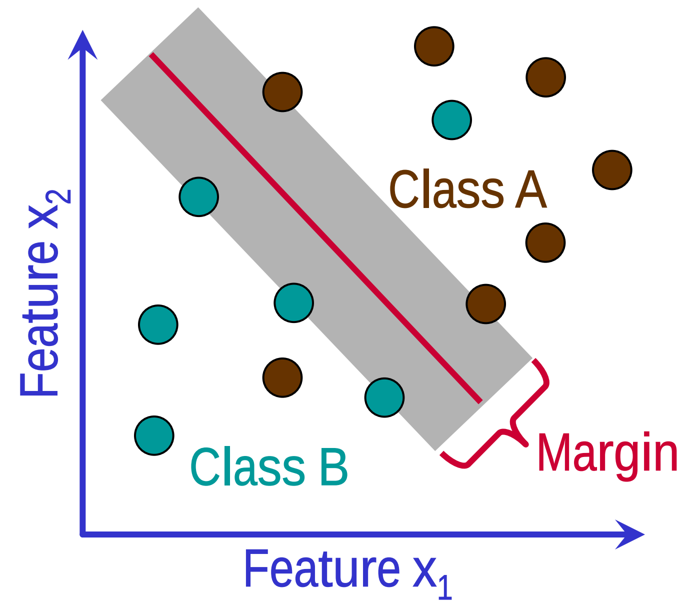
$$\min_{W,C,\xi} \quad \sum \xi_i + \lambda \cdot W^T W$$

$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$\left(i = 1,2,\cdots,N\right)$$

Error of i-th training sample



Support Vector Machine with Noise

- Can be solved by convex programming
  - Cost : sum of two convex functions
  - Constraints: linear and hence convex

Linear (convex) — Quadratic (convex)

$$\min_{W,C,\xi} \sum \xi_i + \lambda \cdot W^T W \longrightarrow \text{Convex}$$

$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C \right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$\left(i = 1, 2, \cdots, N \right)$$

$\Big\} \rightarrow$ Linear

(Convex optimization)

Regularization

- Regression vs. classification

$$\min_{\alpha} \quad \|A \cdot \alpha - B\|_2^2 + \lambda \cdot \|\alpha\|_2^2$$

Regression

Regularization

$$\min_{W,C,\xi} \quad \sum \xi_i + \lambda \cdot W^T W$$

$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \ge 1 - \xi_i$$

$$\xi_i \ge 0$$

$$\left(i = 1, 2, \cdots, N\right)$$

Support vector machine

Other regularization forms can also be used for support vector machine

Regularization

- L_1-norm regularization is used to find a sparse solution of W

L_1-norm regularization

$$\min_{W,C,\xi} \sum \xi_i + \lambda \cdot W^T W$$
$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$
$$\left(i = 1,2,\cdots,N\right)$$

$\Rightarrow$

$$\min_{W,C,\xi} \sum \xi_i + \lambda \cdot \|W\|_1$$
$$\text{S.T.} \quad y_i \cdot \left(W^T X_i + C\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$
$$\left(i = 1,2,\cdots,N\right)$$

Important for feature selection

Regularization

- Feature selection

$$f(X) = W^T X + C \quad \begin{cases} \geq 0 & (Class\ A) \\ < 0 & (Class\ B) \end{cases}$$

$$\underbrace{\begin{bmatrix} 0 & 0 & \times & 0 & \times \end{bmatrix}}_{W^T} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}}_{X}$$

Important features (x_3, x_5)

Summary

- Classification
  - Support vector machine
  - Regularization