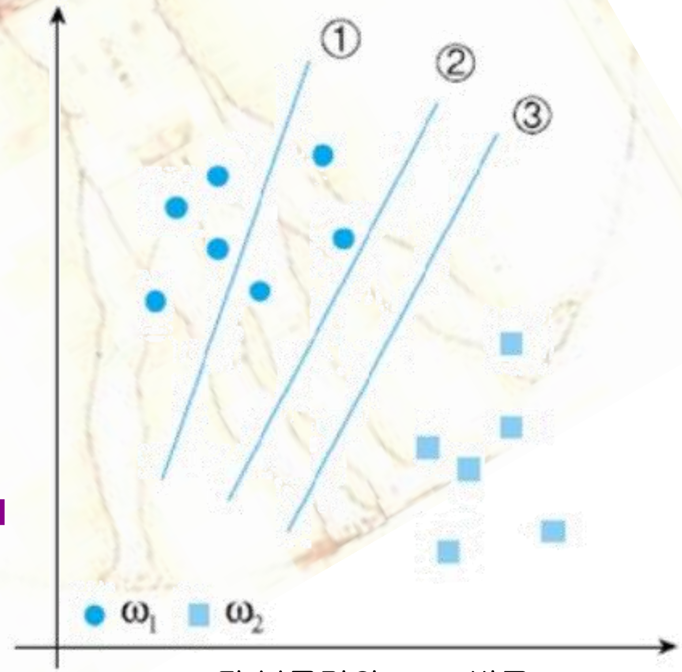
Support Vector Machine(SVM)

Pattern Recognition & Machine Learning Laboratory
Geon-jun Yang July 8, 2019

Introduction

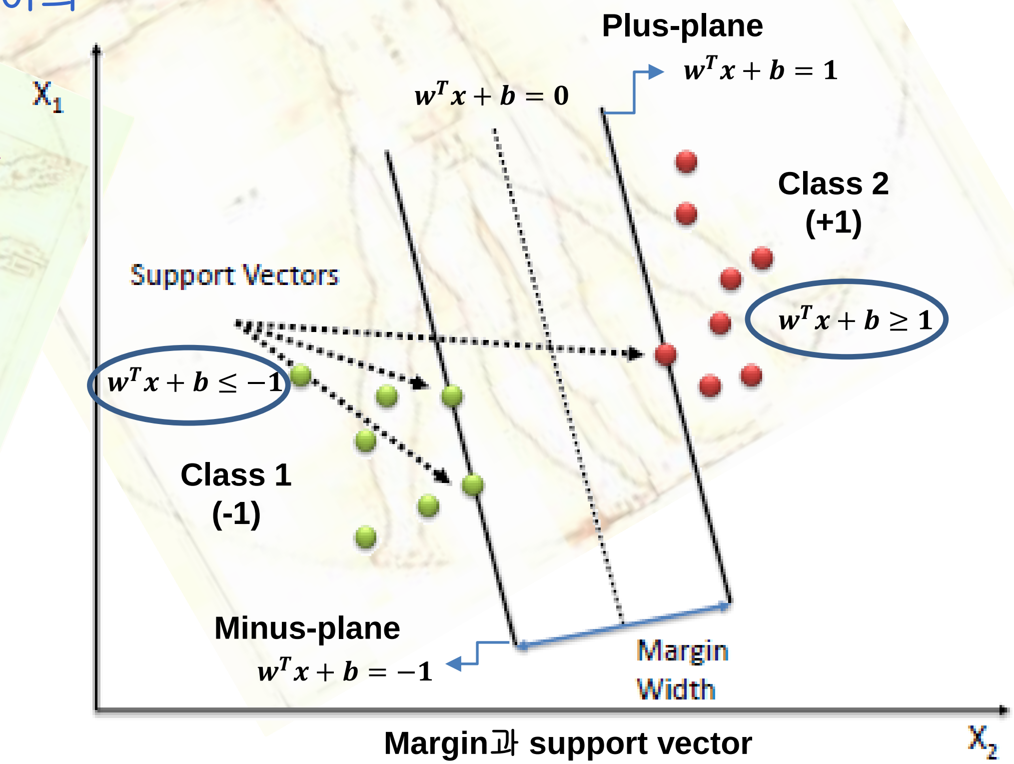
- SVM의 정의 및 특징
  - 결정 경계, 즉 분류를 위한 기준 선을 정의하는 모델
  - 기존 분류기는 '오류율을 최소화'가 목표
  - SVM은 '마진을 최대화 ' 하여 일반화 능력(Generalization ability) 극대화하는 것
  - 분류기의 일반화 능력
    - ②, ③ 모두 분류를 정확히 하였음
    - 신경망은 ①에서 시작하여 ②에 도달하면 멈춤
    - 하지만 SVM 은 마진을 최대화 하여 일반화 능력이 뛰어나고 분류기 품질이 좋음
  - 생각해봐야할 문제
    - 마진의 수학적 표현 방법
    - 두 class를 나누는 hyperplane (초 평면)은 무한히 많음
    - 가장 좋은 hyperplane의 기준
    - 선형으로 나누어지지 않는 경우의 비선형 SVM
  - Hyperplane의 일반식
    - $w^T x + b = 0$
    - w : normal vector of the hyperplane
    - b : bias



타 분류기와 SVM 비교

Linearly Separable Problems

- **Margin, Support Vector의 정의**
  - 각 클래스에서 가장 가까운 관측치 사이의 거리
  - Margin은 w (기울기)로 표현이 가능
  - Support vector란 그림에서 실선위에 존재하는 벡터들
  - y값(라벨 값) +1 또는 -1
  - Minus-plane과 Plus-plane사이의 거리가 마진임
    - 이 마진을 최대화하는 hyperplane을 찾는 것이 목표
  - $x^+ = x^- + \lambda w$
  - λ의 유도 과정
    - $w^T x^+ + b = 1$
    - $w^T(x^- + \lambda w) + b = 1$
    - $w^T x^- + b + \lambda w^T w = 1$
    - $-1 + \lambda w^T w = 1$
    - $\lambda = \frac{2}{w^T w}$



Margin과 support vector

Linearly Separable Problems

- **Vector norm과 마진**
  - $||w||_p = (\sum_i |w_i|^p)^{1/p}$
  - L_2 norm
    - $||w||_2 = (\sum_i |w_i|^2)^{1/2} = \sqrt{w_1{}^2 + w_2{}^2 + \cdots w_n{}^2} = \sqrt{w^T w}$
    - $(w^T = w_1 + w_2 + w_3 + \cdots + w_n)$
  - $Margin = distance(x^+, x^-)$
    - $= ||x^+ - x^-||_2$
    - $= ||(x^- + \lambda w) - x^-||_2$
    - $= ||\lambda w||_2$
    - $= \lambda\sqrt{w^T w}$
    - $= \frac{2}{w^T w} \cdot \sqrt{w^T w} \quad (\lambda = \frac{2}{w^T w})$
    - $= \frac{2}{\sqrt{w^T w}} = \frac{2}{||w||_2}$
    - 즉 margin은 2를 w의 L_2 norm으로 나눈 값
  - Margin의 최대화
    - $max\ Margin = \ max \frac{2}{||w||_2} \leftrightarrow min \frac{1}{2}||w||_2 \leftrightarrow min \frac{1}{2}||w||_2{}^2$

Linearly Separable Problems

- Original problem
  - $min \frac{1}{2}||w||_2{}^2$
  - $subject\ to\ y_i(w^T x_i + b\ ) \geq 1\ ,\ \ i = 1, 2, \cdots, n$

- **Convex optimization problem**
  - Objective function is quadratic (2차) and constraint is linear (1차)
  - Quadratic programming (2차 계획법) →convex optimization→전역최적해 존재
  - Training data가 linearly separable한 경우에만 해가 존재

- **Lagrangian Formulation**
  - Lagrangian multiplier를 이용하여 Original problem을 Lagrangian primal 문제로 변환
    - Lagrangian primal
      - $min\ L(w, b, \alpha) = \frac{1}{2}||w||_2{}^2 \ - \sum_{i=1}^{N} \alpha_i (y_i (w^T x_i + b) - 1)$
      - $subject\ to\ \alpha_i \geq 0, i = 1, 2, \cdots, n$
  - 최소값을 구하려면 미분 값 = 0 에서 최소값을 가짐
    - $\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$
    - $\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \longrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$

Linearly Separable Problems

- $\underbrace{\frac{1}{2}||w||_2^{\,2}}_{①} - \underbrace{\sum_{i=1}^{N} \alpha_i(y_i(w^T x_i + b) - 1)}_{②}$
- ①w에 $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ 을 대입하여 정리
  - $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j {x_i}^T x_j$
- ②에 w값과 $\sum_{i=1}^{n} \alpha_i y_i = 0$를 이용하여 정리
  - $-\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j {x_i}^T x_j + \sum_{i=1}^{n} \alpha_i$
- 최종 정리를 하면 다음과 같음
  - $\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j {x_i}^T x_j$
  - $where \sum_{i=1}^{n} \alpha_i y_i = 0$

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

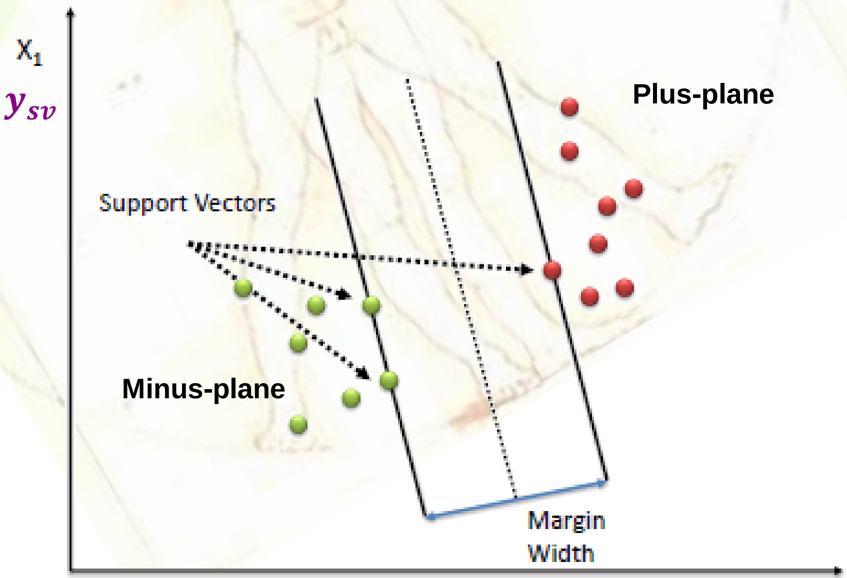
$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \longrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

Lagrangian Primal을 w, b로 미분한 값

- $(\mathrm{w}, \mathrm{b}, \alpha)$가 Lagrangian dual problem의 최적해가 되기 위한 조건 KKT (Karush-Kuhn-Tucker) conditions :
  - Stationarity $\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$, $\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$
  - Primal feasibility $y_i(w^T x_i + b) \geq 1,\ i = 1, 2, \cdots, n$
  - Dual feasibility $\alpha_i \geq 0, i = 1, 2, \cdots, n$
  - Complementary slackness $\boxed{\alpha_i(y_i(w^T x_i + b) - 1) = 0}$

Linearly Separable Problems

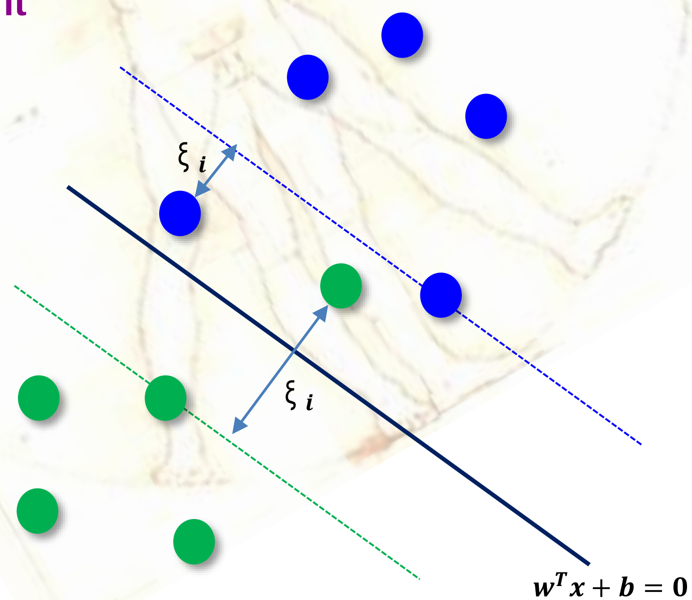
- $\alpha_i\left(y_i\left(w^T x_i + b\right) - 1\right) = 0$
  - $\alpha_i > 0$ and $\left(y_i\left(w^T x_i + b\right) - 1\right) = 0$ ➡ x_i가 plus-plane 또는 minus-plane(마진) 위에 있음 (support vector에 해당)
  - $\alpha_i = 0$ and $\left(y_i\left(w^T x_i + b\right) - 1\right) \neq 0$ ➡ x_i가 plus-plane 또는 minus-plane(마진) 위에 있지 않음, hyperplane 구축에 영향없음
  - Support vector만을 이용하여 optimal hyperplane (decision boundary)을 구할 수 있음
- 또한, 임의의 support vector 하나를 이용하여 b^*를 구할 수 있음
  - $w^{*T} + b^* = y_{sv}$
  - $w^{*T} + b^* = \sum_{j=1}^{n} \alpha_i^* y_i x_i^T x_{sv} + b^* = y_{sv}$
  - $b^* = y_{sv} - \sum_{j=1}^{n} \alpha_i^* y_i x_i^T x_{sv}$
- New data가 hyperplane 밑에 있음
  - $w^{*T} x_{new} + b^* < 0$
  - Class 1로 예측 (minus-plane)
- New data가 hyperplane 위에 있음
  - $w^{*T} x_{new} + b^* > 0$
  - Class 2로 예측 (plus-plane)



New data를 분류 하는 방법

Linearly Non-separable Problems

- **Lagrangian Formulation**
  - **Original Problem**
    - $min\frac{1}{2}||w||_2{}^2 + c\sum_{i=1}^{n}\xi_i$
    - $subject\ to\ y_i\left(w^T x_i + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, 2, \cdots, n$
  - **C는 margin과 training error에 대한 trade-off를 결정하는 parameter**
    - **C↑ : training error를 많이 허용하지 않음 → overfit**
    - **C↓ : training error를 많이 허용 → underfit**
  - **Lagrangian Primal로 변환하여 식을 정리**
    - $max\ \frac{1}{2}||w||_2{}^2 - \sum_{i=1}^{n}\alpha_i\alpha_i y_i y_j x_i{}^T x_j$
    - $subject\ to\ \sum_{i=1}^{n}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C$
  - **KKT condition으로부터 다음과 같은 정보를 얻을 수 있음**
    - $\alpha_i\left(y_i\left(w^T x_i + b\right) - 1 + \xi_i\right) = 0$
    - $\alpha_i = C - \gamma_i,\ \ \gamma_i\xi_i = 0,\ i = 1, 2, \cdots, n$

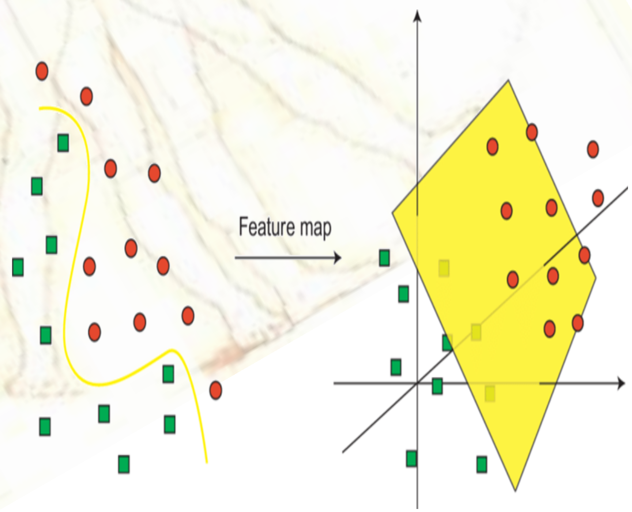


Linearly Non-separable Problems

Nonlinear Classification

- **Mapping Original Space to Kernel Space**
  - **SVM**을 **original space**가 아닌 **feature space**에서 학습
  - 기본적 원리는 원래특징 공간에서는 선형 분리가 불가능하나, 더 높은 공간으로 매핑하여 선형 분리가 가능하게 함.
  - $\Phi : \mathbf{x} \rightarrow z = \Phi(x)$
    - **Ex)** $\Phi: (x_1, x_2) \rightarrow (x_1, x_2, x_1{}^2, x_2{}^2, x_1, x_2)$

      2D 5D
  - 고차원 **feature space**에서는 관측치 분류가 더 쉬울 수 있음
  - **Lagrangian dual formulation**
    - $max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{\Phi(x_i)^T \Phi(x_j)}$
    - $s.t \sum_{i=1}^{n} a_i y_i = 0$
  - 커널 함수의 성질을 이용해 다음과 같이 변경가능
    - $max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i x_j)$
    - 위와 같이 변경이 가능한 이유는 특징벡터가 혼자 등장하지 않고 다른 특징 벡터와의 내적으로 나타나기 때문임



커널 트릭을 활용한 고차원에서의 선형분리

Nonlinear Classification

- 커널 대치 (커널 트릭)
  - 어떤 수식이 벡터 내적을 포함할 때, 그 내적을 커널 함수로 대치하여 계산하는 기법
    - 실제 계산은 저 차원 공간에서 커널 함수의 계산으로 이루어짐
    - 고차원 공간에서 작업하는 효과
    - 실제 계산은 저 차원에서 이루어지지만 분류는 선형분류에서 유리한 고 차원에서 수행
    - 꾀(트릭)를 부려 차원의 저주를 피한 셈
- **SVM**이 사용하는 커널
  - **SVM** 사용시 딱히 기준이 없어서 커널을 결정하는 것이 어려움
  - 사용하는 커널에 따라 **feature space**의 특징이 달라지기 대문에 데이터에 맞는 커널을 사용하는 것이 중요함
    - 다항식 커널 - $K(x, y) = (x \cdot y + 1)^p$
    - **RBF (Radial Basis Function)** 커널 - $K(x, y) = e^{-||x-y||^2/2\sigma^2}$
    - 하이퍼 볼릭 탄젠트 커널 - $K(x, y) = tanh(\alpha x \cdot y + \beta)$
- **SVM**의 특성
  - 사용자가 설정해야 하는 매개변수가 적음 (커널 , C)
  - 다른 모델들에 비해 일반화 능력이 뛰어남
  - 최적의 커널을 자동으로 선택하는 알고리즘은 없음