6_Support_Vector_Machine

1) 문제 정의

- SVM(Support Vector Machine)이란, **두 클래스를 구분하는 decision boundary를 찾는 알고리즘**이다. 여기서 **Support Vector**란 **마진 경계에 위치한 샘플들**로, Decision Boundary를 결정하는데 영향을 주는 Vector들이다.
- **마진(margin)은 샘플과 decision boundary 간의 거리**이다. 마진 밖에 있는 샘플들은 decision boundary에 결정에 영향을 주지 않으며, 따라서 최적화 과정에서 무시된다. **마진을 최대화하는 것이 SVM의 목표**이다. $|x^1 - x^2| = \frac{2}{|w|}$ (w는 decision boudnary의 법선벡터). $\frac{|w|}{2}$를 최소화하는 것으로 볼 수 있으며, convex form인 $\frac{|w|^2}{2}$로 바꾸어 최소값을 구하는 것이 좋다. $w^* = \min_{w,b} \sum_j \frac{w_j^2}{2}$

2) 해당 문제에 대한 일반적인 접근법

- C라는 파라미터 조정을 통해 학습 데이터에 얼마나 민감하게 반응할 것인지 결정할 수 있다.

  $w^* = \min_{w,b} \sum_j \frac{w_j^2}{2} + C \sum_{i=1}^{n} \xi_i$

- C가 작을수록 소프트 마진(soft margin)이다. 마진을 넓게 유지하기 위해 조금의 에러를 허용하는 방식이다. C가 클수록 하드 마진(hard margin)이다. 데이터가 wrong side에 위치하는 것을 절대 허용하지 않는 방식이다.
- 분류를 잘 하는 수준에서 가장 작은 C값을 구하기 위해서는 Cross-validation을 통해 validation error가 가장 적을 경우의 C값을 구한다. 이렇게 구한 C값으로 최적의 SVM이 결정된다.

3) 일반적인 접근법의 제한사항(from 논문 or 본인생각)

- 데이터 포인트들이 평면 상에서 직선으로 분리된다면 기존의 SVM 방식을 사용하면 된다. 하지만, 실제로 수많은 데이터 포인트들이 정확히 평면 상에서 분리되지 않는 경우도 많을 것이다. **결정 경계 결정 함수가 선형이 아니고 $x_1 = x_2^2 + 5$와 같이 비선형**이면, 새로운 방식을 통해 decision boundary를 해결할 필요가 있다.
- **하드 마진**의 경우 **이상치에 민감한 모델이 된다는 것이 단점**이다. 하드 마진일 경우에도 이상치에 민감하지 않게 두 클래스가 잘 구분되는 decision boundary를 구할 필요성이 있다.

4) 제한사항에 대한 해결방안(from 논문 or 본인생각)

- **커널 트릭(kernel trick)**을 통해 비선형 데이터인 경우와 하드 마진일 경우에도 두 클래스를 잘 분류하는 decision boundary을 구할 수 있다. 커널 트릭이란, **실제로 특성을 추가하지 않으면서 다항식 특성을 많이 추가한 것과 같은 결과를 낼 수 있는 방식**이다. 원래 데이터 공간을 **더 높은 차원의 공간으로 확장하여 비선형 문제를 해결**하는 것이다.
- 커널 트릭을 사용할 경우 **이상치와 결정 경계(decision boundary) 간의 거리가 멀어져서 이상치가 결정경계(decision boundary)에 미치는 영향을 줄이기 때문에 하드 마진에서도 더 강력한 모델을 구축할 수 있는 것이 장점**이다.