**Machine Learning – 12week** 과제

20145165 정균모

빅데이터

SVM(서포트 벡터 머신)

머신 러닝의 분야 중 하나로 매우 강력하고 선형이나 비선형 분류, 회귀, 이상치 탐색에도 사용할 수 있는 다목적 모델이다. 주로 패턴 인식과 자료 분석을 위해 사용하고 있다. SVM은 N차원에 공간을 (N-1)차원으로 나눌 수 있는 초 평면을 찾는 것이다. 한 데이터 집합이 주어졌을 때 데이터들 사이에 최적의 경계를 찾고자 한다. 이는 마진(클래스들 사이의 간격, 각 클래스의 말단에 위치한 데이터들 사이의 거리)을 최대화할 수 있는 경계를 찾는 것이기도 한다

SVM은 선형과 비선형으로 나누어진다. 우선 선형 SVM부터 살펴보자면 하드 마진과 소프트 마진으로 볼 수 있다. 하드 마진은 두 클래스를 분류할 수 있는 최대 마진의 초 평면을 찾는 방법인데 단, 모든 훈련 데이터는 마진의 밖에 위치하게 선형으로 구분해야 된다.(하나의 오차도 허용하면 안된다는 것) 하지만 모든 데이터를 선형으로 오차없이 나눌 수 있는 결정 경계를 찾는 것은 사실상 어렵다. 그래서 이를 해결하고자 나온 개념이 소프트마진이다. 소프트마진은 하드 마진이 가진 한계를 개선하고자 나온 개념으로써 완벽하게 분류하는 초 평면을 찾는 것이 아니라 어느 정도의 오차를 허용하는 방식이다. 소프트마진에서 오차를 허용하고 이를 고려하기 위해 Slack variable을 사용한다. Slack variable은 해당 결정 경계로부터 잘못 분류된 데이터의 거리를 측정하기 위해 사용한다.

비선형 SVM은 선형SVM이 불가능한 상태의 데이터 집합을 2차원으로 변환하여 선형SVM이 가능한 형태로 만든 것 즉 선형 분리가 불가능한 입력 공간을 선형 분리가 가능한 고차원 특성공간으로 보내 선형 분리를 진행하고 다시 기존의 입력공안으로 변환하는 것이다 입력 공간을 특성공간으로 변환하기 위해서 Mapping function을 사용한다. 하지만 여기서 문제가 발생하는데 고차원의 특성공간으로 변환하고 목적 함수에 대한 문제를 푸는 것이 간단한 차원에서는 가능하나 그 차수가 커질수록 계산량의 증가가 어마 하다는 것이다. 이를 다시 해결하고자 나오는 것이 커널 트릭이라는 개념이다. 커널 트릭은 실제로는 데이터의 특성을 확장하지 않지만 마치 확장한 것처럼 만들어 계산하는 방식인데 특성 공간의 차원이나 매핑 관계를 명시적으로 알 필요 없이, 확장된 특성 공간의 두 벡터의 내적만을 계산하면 된다.

마지막으로 SVM의 장단점을 정리하자면

장점:

1. 과 적합을 피할 수 있다.
2. 분류 성능이 좋다.
3. 저 차원, 고차원 공간의 적은 데이터에 대해서 일반화 능력이 우수하다.
4. 잡음에 강하다.
5. 데이터 특성이 적어도 좋은 성능을 나타낸다.

단점:

1. 커널 함수 선택이 명확하지 않다.
2. 파라미터 조절을 적절히 수행하여 최적의 모델을 찾을 수 있다.
3. 계산량 부담이 있음
4. 데이터 특성의 스케일링에 민감하다.