

<머신러닝 톺아보기 5주차 과제>

박민서

1. SVM은 고차원 데이터셋 분류를 일반화하는데 매우 효과적이고, 그것의 훈련은 quadratic 프로그래밍 문제를 풀 수 있도록 틀이 짜여 있다.
2. 목적: Training data(X, Y)를 통해 w, b 를 찾기
3. 두 개의 계층으로 분류하기

두 class를 나누는 hyperplane은 무한히 많다. 그렇다면 어떤 hyperplane이 가장 좋은 hyperplane인가? 훈련세트에서 margin을 최대화=일반화 오류를 최소화하는 hyperplane이 가장 뛰어난 예측 성과를 내는 좋은 hyperplane이라 할 수 있다.

4. Margin: 각 클래스에서 가장 가까운 관측치 사이의 거리이며, w 로 표현한다.
5. Linearly nonseparable problems: 실제 주어진 데이터들은 linear하게 나뉘지기는 보다는 복잡하게 얽혀 있어 두 개의 클래스로 명확하게 구분하기 어려운 경우가 많다. 이 경우 linear decision boundary를 이용하여 완벽하게 나누는 것은 불가능한데, 이는 어느 정도의 error를 허용하는 방안을 통해 해결 가능하다.
6. Training error를 많이 허용하지 않는 경우 overfit 될 수 있고, training error를 많이 허용하는 경우 underfit 될 수 있기에 C 의 값을 적절히 조절하는 것이 필요하다.
7. Kernel Methods for Nonlinear Classification: 관측치 x 를 더 높은 차원으로 변환시켜 분류하는 방법, 저차원 공간에서는 분류가 어려웠던 관측치들을 고차원 feature space로 가져와 관측치 분류를 용이하게 하며 더욱 효율적으로 계산한다.
8. SVM을 사용하는 경우 kernel을 결정하는 것은 명확한 기준이 없기에 어려운 문제라고 할 수 있다. 사용하는 kernel에 따라 feature space의 특징이 달라지기 때문에 데이터의 특성에 맞는 kernel을 결정하는 것은 중요하다. 일반적으로는 RBF kernel, sigmoid kernel, low degree polynomial kernel 등이 주로 사용된다.